Universidade Estadual de Campinas
Instituto de Computação

Edgar Kenji Tanaka

Multilingual abstractive summarization of podcasts with Longformers

Sumarização abstrativa multilíngue de podcasts utilizando Longformers

CAMPINAS
2022

Edgar Kenji Tanaka

Multilingual abstractive summarization of podcasts with Longformers

Sumarização abstrativa multilíngue de podcasts utilizando Longformers

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Jacques Wainer
Co-supervisor/Coorientadora: Profa. Dra. Ann Clifton

Este exemplar corresponde à versão final da Dissertação defendida por Edgar Kenji Tanaka e orientada pelo Prof. Dr. Jacques Wainer.

CAMPINAS
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

T153m Tanaka, Edgar Kenji, 1984-
Multilingual abstractive summarization of podcasts with Longformers / Edgar Kenji Tanaka. – Campinas, SP : [s.n.], 2022.

Orientador: Jacques Wainer.
Coorientador: Ann Clifton.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Processamento de linguagem natural (Computação). 2. Sumarização automática. 3. Podcasts. 4. Aprendizado de máquina. 5. Aprendizado profundo. I. Wainer, Jacques, 1958-. II. Clifton, Ann. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Sumarização abstrativa multilíngue de podcasts utilizando Longformers
Palavras-chave em inglês:
Natural language processing (Computer science)
Automatic summarization
Podcasts
Machine learning
Deep learning
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Jacques Wainer [Orientador]
Julio Cesar dos Reis
Thiago Alexandre Salgueiro Pardo
Data de defesa: 06-12-2022
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-1664-2118
- Currículo Lattes do autor: http://lattes.cnpq.br/3061526797142751

Universidade Estadual de Campinas
Instituto de Computação

Edgar Kenji Tanaka

Multilingual abstractive summarization of podcasts with Longformers

Sumarização abstrativa multilíngue de podcasts utilizando Longformers

Comissão Examinadora:

- Prof. Dr. Jacques Wainer
  IC/Unicamp
- Prof. Dr. Julio Cesar dos Reis
  IC/Unicamp
- Prof. Dr. Thiago Alexandre Salgueiro Pardo
  ICMC/USP

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 06 de dezembro de 2022

Acknowledgements

First, I would like to thank my family who have always supported and encouraged my academic history. I want to thank my parents who have always set an example on the importance of learning and studying, no matter how old or young you are.

I would like to thank my industrial advisor Ann Clifton who has been an incredible mentor and colleague. I am very grateful to Ann for everything I have learned about Machine Learning, Natural Language Processing and applied research. She is a role model as a scientist and an inspiration as a leader. Thanks for all the questions, advice and discussions who helped me grow as a professional.

I would also like to thank Rosie Jones who sponsored my Master thesis project at LiLT and who contributed to the design, production and release of the Portuguese Podcast dataset.

I want to thank my academic advisor Prof. Jacques Wainer for all of the guidance, feedback and discussions throughout this research project. My collaborations with Prof. Jacques have taught me how to reason as a researcher in a pragmatic manner without being careless.

I would also like to thank Md. Iftekhar Tanveer for all the suggestions and feedback which contributed to the multilingual aspect of this research.

Finally, I would like to thank my partner Lucione for all the love, support and patience during this past year.

Resumo

Podcasts se estabeleceram como uma importante fonte de conteúdo em áudio nos dias de hoje. Conforme o número de podcasts aumenta, fica cada vez mais evidente a necessidade de boas descrições que ajudem usuários a decidir se vão ou não escutar a um determinado episódio. No entanto, as descrições fornecidas pelos criadores de podcasts geralmente carecem de informações importantes sobre o episódio. Além disso, estas descrições são frequentemente usadas para propaganda de produtos ou divulgação de redes sociais. Como alternativa a essas descrições fornecidas pelos criadores, a tarefa de sumarização automática de podcasts foi proposta na conferência TREC 2020. Muitos pesquisadores propuseram diferentes modelos baseados em deep learning para resolver esse problema. No entanto, todos modelos propostos estavam restritos a apenas podcasts em inglês. À medida que o consumo de podcasts aumenta globalmente, é fundamental explorar modelos capazes de ingerir e gerar texto em vários idiomas.

Nesta dissertação de mestrado, investigamos a aplicação de modelos multilíngues baseados em transformadores para gerar automaticamente resumos abstrativos a partir de transcrições de podcasts. Experimentamos e contrastamos modelos com um mecanismo de full self-attention e um mecanismo de Longformer self-attention. Além disso, estudamos o impacto do ajuste fino desses modelos de forma monolíngue e bilíngue. Por fim, exploramos o fenômeno de cross lingual transfer learning no contexto de sumarização de podcasts multilíngue. O escopo de nossa pesquisa se limita ao inglês e português, mas a metodologia proposta aqui pode ser generalizada para qualquer outro conjunto de idiomas.

Abstract

Podcasts are now established as an important source of audio content today. As the number of podcast shows increases, so has the need for high-quality descriptions which assist consumers to decide whether to listen to an episode or not. However, descriptions provided by podcast creators often lack important information about the episode. Not only that, they are often used for self-promotion instead of describing the actual content. As an alternative to the creator provided descriptions, the task of automatic podcast summarization was proposed in the TREC conference 2020. Many researchers proposed different deep learning based models to solve this problem but they were all restricted to podcasts in English. As podcast consumption continues to rise globally, it is critical to explore models capable of ingesting and generating text in multiple languages.

In this Master thesis, we investigate the application of transformer-based multilingual models to automatically generate abstractive summaries from podcast transcripts. We experiment and contrast models with a full self-attention mechanism and a Longformer attention mechanism. In addition, we study the impact of finetuning these models monolingually and bilingually. Lastly, we explore cross lingual transfer learning in this domain of multilingual podcast summarization. We scope our research to English and Portuguese but the methodology proposed can be generalized to any other set of languages.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Terminology

Let us first define the terms commonly used in this Master thesis:

- Podcast: audio format distributed via internet and characterized by talk instead of music. This term will be used to refer to podcast content in general (e.g. podcast summarization, podcast dataset).
- Creator (or Podcast Creator): the person or group of people who produce and release podcast content.
- Episode (or Podcast Episode): a playable unit with podcast content. Episodes are usually released on a regular cadence.
- Show (or Podcast Show): a collection of Episodes produced by the same Creator.

1.2 Motivation and Problem Statement

Podcasts have become a very popular medium among audio listeners in the last decade. According to a survey by Edison Research[46], the share of people who are familiar with podcast in the United States has grown from 46% to 79% in the past 10 years. During the same period, the number of monthly podcast listeners has more than tripled in United States. In addition to podcast consumption growth, podcast creation has also increased in the past years. It is now possible to find podcasts in almost any imaginable genre including true crime stories, news, sports, education or religion. According to [1], the number of podcast shows in the Apple Podcasts platform has increased from 550,000 to 2,440,383 between 2018 and 2022.

With this massive scale of podcast content available, browsing and picking what shows and episodes to listen to has become increasingly harder for audio consumers. Given that podcast episodes are long forms of audio, typically ranging from 30 minutes to 3 hours, scanning parts of an episode in order to make this decision is a time consuming and tedious task. Although recommendations systems can certainly assist users in this content selection process, contextual information about an episode is key before making a

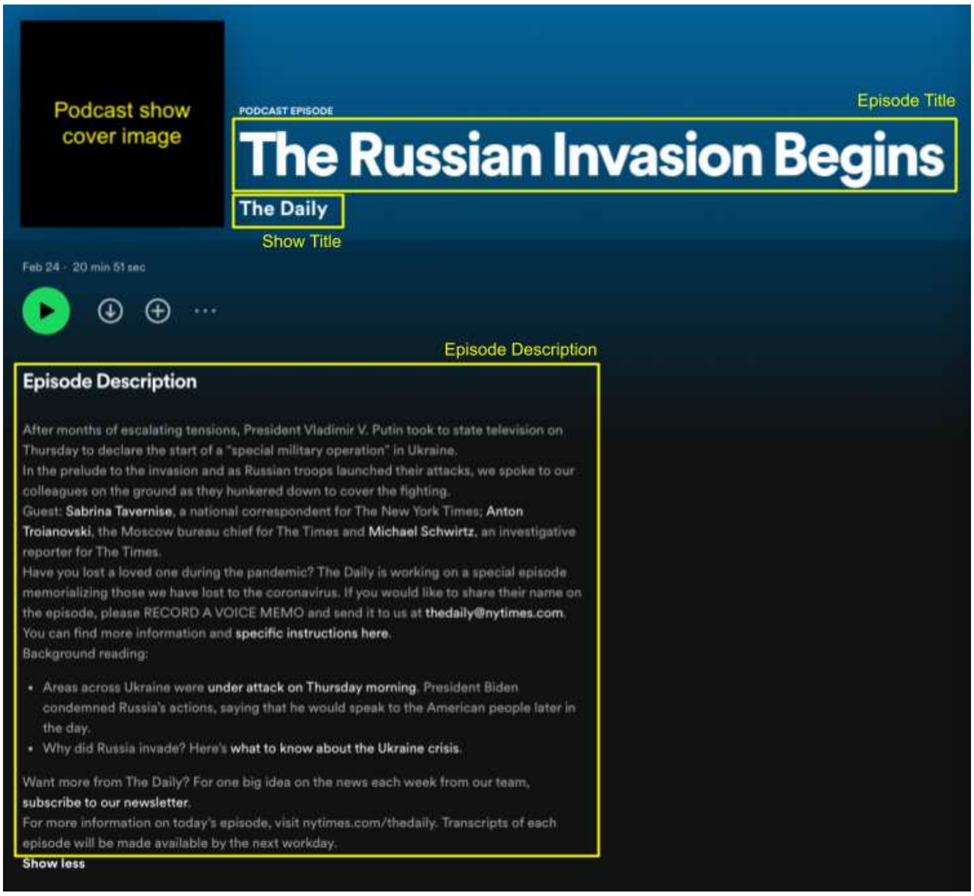


Figure 1.1: Example of a podcast episode. We highlight the episode title, episode description and the show title. We can also see the episode duration (20 min 51 sec) and the publication date (Feb 24). This is a screenshot of the Spotify application taken on July of 2022.

final decision. A few examples of contextual information include: who are the participants of this episode, what are the main topics discussed in the episode, what is the duration, what is the format of this episode (e.g. interview, monologue).

Podcast episodes usually come with metadata such as title, description and duration (see Figure 1.1). Episode descriptions providing contextual information can be of great help to users when deciding whether to listen or not to an episode. However, some of the episode descriptions are simply uninformative or filled with ads and social media links. Given that writing useful and informative episode descriptions is a time consuming task, we believe that using state-of-the-art summarization methods to fully or semi-automate this process would have a positive impact not only on podcast creators but also on podcast listeners.

In [25], human evaluators scored summaries generated automatically from transcriptions of episodes on the scale EGFB (Excellent, Good, Fair, Bad). On the same set of episodes, the evaluators also scored the episode descriptions written by the podcast creators using the same scale. In total, 179 episodes were used in this study. The results show that automatically generated summaries were perceived as having higher quality than the episode descriptions. While 71 of the episode descriptions were scored as Good

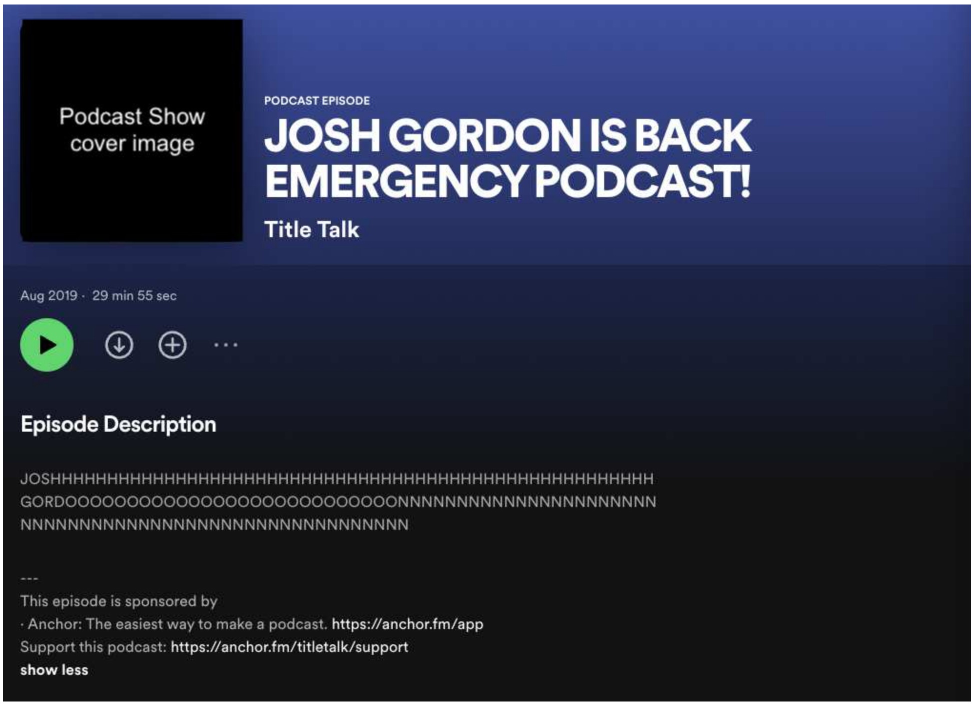


Figure 1.2: Example of a podcast episode with uninformative description.

or Excellent, 105 of the auto-generated summaries received the same scores. These figures suggest that episode descriptions are not always written with the intention to act as summaries[25] and, at the same time, demonstrate promising results in the application of deep learning models to summarize podcasts. In Figure 1.2, we can see an example of episode description which does not provide any useful information to the user.

For the past decades, researchers have studied the problem of summarization in different media: text, audio and video. In this research project, we are particularly interested in summarization of audio content as podcasts are an audio-only form of media. There are three approaches when performing audio summarization: conditioning the summary using only audio features, extracting the text inside the audio signal and conditioning the summarization process using textual methods and an hybrid approach combining the first two [21]. Podcasts largely contain spoken-word content so when considering an audio-based summarization method, we would need to consider models which operate on the linguistic level (i.e. identifying phonemes and words) such as HuBERT [24]. There are two challenges when taking this approach: 1) the great majority of the summarization datasets available are in text format and not audio and 2) processing audio is computationally much more expensive than processing text. In light of recent success of many transformer-based models in the NLP domain and availability of podcast transcripts, we are taking the second approach in this Master thesis. In a sense, we are transforming the problem of audio summarization into a problem of text summarization.

Text summarization has long been researched in the Natural Language Processing (NLP) community. However, most works have focused on summarizing news articles while the summarization of podcasts still remains fairly unexplored. Summarizing podcasts is a challenging task due to a number of reasons. Firstly, there are many podcast formats such as interviews, debates, monologues. Secondly, podcast transcripts are noisy as they are ill punctuated and the audio often contains fillers and overlapping speakers. Thirdly, these transcripts are very long and state-of-the-art models can ingest only a limited number of tokens. And lastly, podcast content is produced in many different languages including some with scarce data resources for training.

In 2020, the Text Retrieval Conference (TREC)[25] debuted the Podcast track and invited researchers to explore the problem summarizing podcast episodes. Several proposals have been submitted but they were restricted only to the English language.

For robust summarization on diverse datasets, including spoken language and long-form text and low-resource scenarios, it is crucial to expand our methods outside of English alone and explore multilingual models. A multilingual model is a single model capable of ingesting and generating text in multiple languages. This method circumvents the need of training a monolingual model for every single language. Moreover, multilingual models have demonstrated better performance than monolingual models, especially for low-resource languages. [13]

1.3 Objectives

In this research project, we investigate the application of transformer-based multilingual models to produce abstractive summaries from podcast transcripts. We experiment with a number of different baselines as well as multiple combinations of two models and finetuning variations.

Our first set of experiments will use MBART[52] and finetune it to the summarization task using podcasts data. We experiment and compare finetuning it monolingually and bilingually with podcasts in English and Portuguese.

On our second set of experiments, we create a modified version of MBART[52] by replacing the original attention mechanism module with Longformer's[5] attention mechanism. We call this new version LongMBART. The main advantage of LongMBART is the capacity to ingest 8 times more text than the original MBART. We compare the summaries produce by these two models using ROUGE scores as our metric. Additionally, we compare the effects of finetuning LongMBART monolingually and bilingually.

Furthermore, we experiment finetuning the LongMBART model in two rounds: first using the XL-SUM dataset and later with podcasts data. We compare this finetuning scheme with the LongMBART finetuned only with podcasts data.

Lastly, we evaluate if cross-lingual transfer learning occurs when finetuning MBART and LongMBART in one language and testing in another.

Previous works have explored multilingual summarization, summarization of long documents and the summarization of noisy documents. However, the intersection of all of those still remains unexplored. Table 1.1 compares previous works in the summarization domain and the contribution provided by this research project.

Previous Work	Summarization	Long Text	Multilingual
BART	Yes	No	No
LED (Longformer Encoder-Decoder)	Yes	Yes	No
XL-SUM	Yes	No	Yes
My research	Yes	Yes	Yes

Table 1.1: Comparison of previous works and this research project.

1.4 Contributions

We summarize the main contributions of our work as follows:

1. We have published a dataset consisting of 123,054 podcast episodes in Portuguese from 16,131 shows and encompassing more than 76,000 hours of speech audio.

2. We have proposed new methods to the problem of multilingual podcast summarization. We have evaluated results with humans and intrinsic metrics.

3. We have concluded that there is a high correlation between the summary quality perceived by humans and intrinsic metrics such as ROUGE.

4. We have studied cross-lingual transfer learning when applying multilingual transformer-based models to summarize podcast episodes.

5. We have studied the impact of using Longformer's attention mechanism when summarizing podcast episodes multilingually. We have also contrasted these results with the full attention mechanism.

1.5 Related Publications

- TREC 2021 Podcasts Track Overview; Jussi Karlgren, Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Sravana Reddy, Edgar Tanaka, Md Iftekhar Tanveer.

- TREC 2021 Conference Paper "Multilingual Podcast Summarization using Longformers", Edgar Tanaka, Ann Clifton, Md Iftekhar Tanveer.

- Arxiv preprint "Cem Mil Podcasts: A Spoken Portuguese Document Corpus", Edgar Tanaka, Ann Clifton, Joana Correia, Sharmistha Jat, Rosie Jones, Jussi Karlgren, and Winstead Zhu.

1.6 Outline

We organize this Master thesis as follows:

- Chapter 2: We review the literature on text summarization and deep learning architectures which served as foundations for the state-of-the-art summarization models. We discuss the Longformer architecture, a multilingual summarization model and the most recent publications in podcast summarization.
- Chapter 3: We detail the methods and datasets used in this research project. We discuss the different models and different training strategies used as well as how we evaluate the results. We also explain the hypothesis motivating each of the experiments planned.
- Chapter 4: We expose the results of our research from two perspectives: human evaluation and intrinsic evaluation. We also discuss the correlation between these two evaluation methods.
- Chapter 5: We present our conclusions of this work and propose opportunities for future research on multilingual podcast summarization.

1.7 Summary

In this chapter, we have first defined important terms commonly used in this Master's thesis. We have also defined and motivated the problem of summarizing podcast episodes using multilingual models. To support the motivation, we considered the following: 1) automatically generated podcast summaries were perceived as having higher quality than the episode descriptions provided by podcast creators in a previous human evaluation, 2) A multilingual model circumvents the need for training a monolingual model for every single language and 3) multilingual models have demonstrated better performance than monolingual models, especially for low-resource languages.

We then compare this Master thesis with previous related workers, calling out the fact that there is no previous work in the intersection of the following domains: summarization, long text and multilingual.

Lastly, we list the five contributions of our work and an outline of this Master's thesis.

Chapter 2

Literature Review

2.1 Automatic Text Summarization

Automatic Text Summarization (ATS) is one of the most challenging tasks in Natural Language Processing (NLP) and it has been studied since the 1950s. According to [17], ATS research began with Luhn's work[35] which proposed a method to automatically generate abstracts of magazine articles and technical papers. Since then, ATS has branched into many sub-domains as depicted in Figure 2.1. We will briefly define some of the classifications of ATS systems proposed in [17].

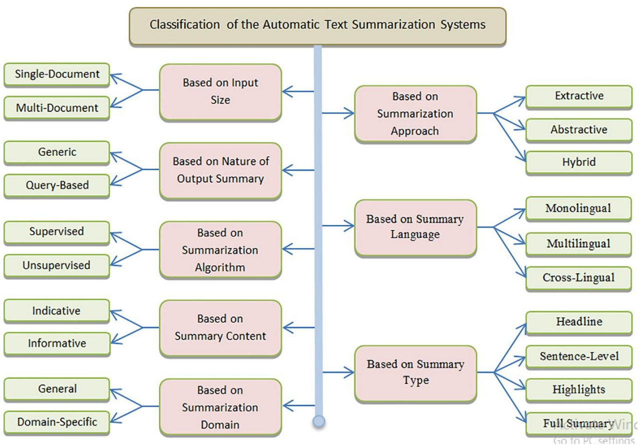


Figure 2.1: Classification of Automatic Text Summarization systems. Extracted from [17]

Classification based on the Input Size: Single-document or Multi-document. Single-document ATS systems will generate one summary based on one input document. Multi-document ATS systems will generate one summary based on a set of input documents.

Classification based on Nature of Output Summary: Generic or Query-Based. Generic ATS systems aim to provide a general sense of one or more input documents by distilling the most important information. Query-Based ATS systems generates a summary with the most important information according to an input query. A query may specify a topic, a user or keywords.

Classification based on Summarization Algorithm: Supervised or Unsupervised. The definition is the same as of any Machine Learning system. Supervised methods require annotated training data whereas Unsupervised methods do not. Creating annotated data is expensive, specially for the summarization task. However, a number of datasets containing pairs of article-summary - most of them collected from major news sources - have been published recently .

Classification based on summarization domain: General or Domain-Specific. General ATS systems are designed to summarize documents belonging to any domain. In other words, they are domain-agnostic. On the other hand, Domain-Specific ATS systems are designed to summarize documents from a specific domain such as news articles [36][20], scientific papers [11][34], legal documents [3][15] and transcripts [43][6]. This Master thesis focuses on the domain of transcripts, more specifically podcast transcripts.

Classification based on summarization approach: Extractive, Abstractive or Hybrid. The extractive text summarization approach splits a document into sentences and ranks them according to a relevance score. The output summary is composed of the most relevant sentences. The abstractive summarization approach will first create a representation of the input document(s) and then generate a summary based on this representation. Unlike extractive summaries, abstractive summaries are written in novel sentences, i.e. they may differ from the sentences in the original document(s). The hybrid text summarization approach combines both the extractive and abstract approaches.

Classification based on summary language: Monolingual, Multilingual or Cross-Lingual. A monolingual ATS system is designed to ingest documents in one language and produce summaries in that same language. A multilingual ATS system is designed to ingest documents in multiple languages and produce summaries in the same language as the source document. Cross-Lingual ATS systems can ingest documents in one language and produce summaries in a different language. Most of the literature in ATS focuses on monolingual methods for the English language.

Based on this classification, we propose a **supervised multilingual abstractive single-document method for generic domain-specific (podcast transcripts) summaries**. We do not define the classifications in regards to Summary Type and Summary Content [17] as they are out of scope for this work.

Extractive summaries are well suited for cases where the input documents professionally written such as news articles or research papers. However, the input documents we have used are transcripts automatically produced by Automated Speech Recognition (ASR) from the audio of podcasts. Although ASRs systems have greatly improved in recent years, they still produce errors such as incorrect casing, mispunctuations, incorrect sentence boundaries, misspelled words and names. Fortunately, state-of-the-art abstractive summarization models like BART[30] are trained to reconstruct corrupted text and have demonstrated the ability to generate relatively fluent written summaries even with these ASR errors [9]. Furthermore, both in TREC 2020 [25] and TREC 2021 [28], abstractive systems have scored higher based on human evaluation for the Podcast Summarization track. For all these reasons, we have opted for the abstractive approach instead of the extractive one.

When it comes to monolingual and multilingual summarization, most summarization works have been restricted to monolingual systems in the English language. The lack of datasets in multiple languages (other than English) is one of the main factors limiting research in the field. In other words, there is a need for more diversity language-wise in datasets available to the NLP research community [17]. In [17], 16 out of 19 standard datasets for text summarization are in English. In another survey targeting neural abstractive summarization models[50], we also see a higher concentration of articles only focused on English, more specifically using the CNN/DailyMail dataset [38]. Despite recent efforts to expand the availability of multilingual summarization datasets [48][22], corpora for low resource languages are still not available or composed of a very limited number of documents. With that in mind, we decided to publish the dataset containing 123,054 podcast episodes in Portuguese (pt-BR and pt-PT) used in this Master thesis project as one of our contributions.

2.2 Sequence-to-sequence learning

Back in 2014, Deep Neural Networks (DNN) were already demonstrating impressive performance on difficult tasks such as speech recognition and visual object recognition. However, successful applications of DNNs were still restricted to problems where both its inputs and targets were encoded with vectors of fixed dimensionality. This was a significant limitation given that many relevant problems were expressed with input/output sequences of variable length, i.e. the sequence length could not be defined a-priori. For example, text translation and text summarization are examples of problems where the size of the inputs and output are not fixed[49]. The source (input) document may be composed of a few tokens or dozens of sentences. Likewise, the target (output) document's length will vary and cannot be pre-defined.

The scenario changed when the authors of [49] proposed a sequence to sequence learning method using neural networks. In that paper, they propose an architecture with two main components: an encoder and a decoder. The encoder is a Long Short-Term Memory (LSTM) [23] which reads the input sequence, one timestep at a time, and outputs a large fixed-dimensional vector representation. The decoder is a second LSTM which ingests the encoded representation of the input sequence and then generates the target sequence of variable length. A special character <EOS> is always used to indicate the end of a sequence. Figure 2.2 illustrates the encoding and decoding process.

The authors evaluated this architecture on an English to French translation task from the WMT-14 dataset and results showed that a large deep LSTM with a limited vocabulary could outperform a standard Statistical Machine Translation (SMT) based system whose vocabulary is unlimited. Moreover, the LSTM was able to correctly translate very long sentences even though similar models in the past had reported poor performance in these scenarios[49].

The fact that sequence-to-sequence learning with very little optimization outperformed a mature SMT system, suggested that the same approach would likely do well on other challenging sequence-to-sequence problems[49].

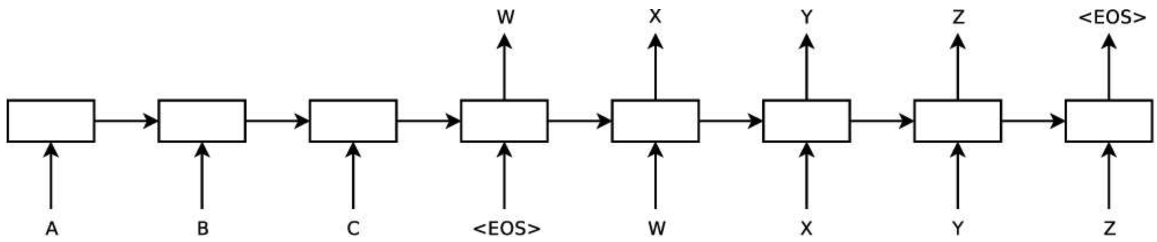


Figure 2.2: Sequence-to-sequence learning. The model reads an input sequence "ABC" and produces the output sequence "WXYZ". <EOS> indicates end-of-sequence. The model stops making predictions after outputting the end-of-sentence token.
Extrated from [49]

2.3 Transformers

Sequence-to-sequence models were established as state-of-the-art for sequence modeling and transduction problems such as language modeling and machine translation. However, those models still relied on recurrent models which meant that computation time was still factored by the length of the input and output sequences. That is because recurrent models compute each element of the input in a sequential manner, i.e. the computation of hidden state h_t is a function of the previous hidden state h_{t-1} and the input for position t. For example, in text data, one token is processed at each timestamp. As a consequence, it is not possible to leverage parallelization within training examples which is a critical limitation for long sequences [53].

In 2017, [53] proposed the Transformer architecture which dispensed the use of recurrence and allowed for significantly more parallelization with the use of positional encodings and an attention mechanism. In other words, a Transformer model could process an entire sequence in one step whereas sequence-to-sequence models required t steps.

2.3.1 Architecture

The transformer architecture uses an encoder-decoder structure. The encoder is responsible for mapping an input sequence of symbol representations $(x_1, ..., x_n)$ to a sequence of continuous representations z = $(z_1, ..., z_n)$. The decoder then uses z as input to generate an output sequence $(y_1, ..., y_n)$ one element at each step. When decoding, the symbol generated in step $t - 1$ is used as an additional input to generate the symbol in step t. Figure 2.3 illustrates the transformer architecture in more details.

In [53], the encoder is composed of a stack of 6 identical layers. Each layer is made of two sub-layers: a multi-head attention and a fully connected feed-forward network. The output of each sub-layer goes through layer normalization which can be written as $LayerNorm(x + Sublayer(x))$, where $Sublayer(x)$ is either the multi-head attention or the feed-forward network.

The decoder is also composed of 6 identical layers, each with three sub-layers. The first sub-layer is a masked multi-head attention which prevents positions from attending to subsequent positions. This is to ensure that predictions for position i can only rely on the outputs seen at positions less than i. In other words, we do not want the model

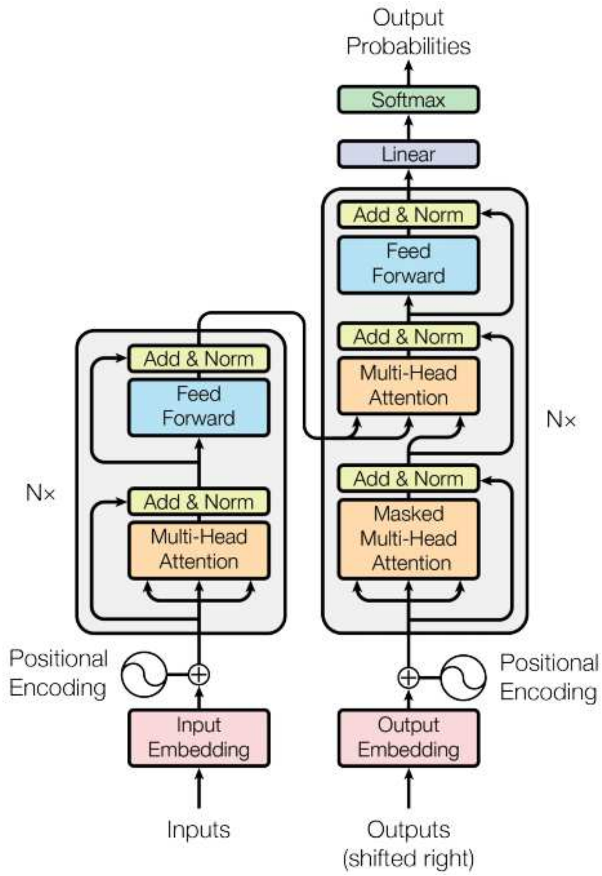


Figure 2.3: The Transformer - model architecture.
Extracted from [53]

to "cheat" by seeing future tokens that it ought to predict. The second sub-layer is a multi-head attention over the output of the encoder stack. The third sub-layer is a fully connected feed-forward network. As in the encoder stack, the same layer normalization is applied on top of each sub-layer output.

The last step after the 6 decoders contains a linear layer followed by a softmax layers. The linear layer, which is a fully connected neural network, contains N cells where N is the size of the vocabulary for the model. The output of this linear layer is called the logits vector. The softmax layer then converts the logits vector into a vector of the same with probabilities for each word. The word with highest probability is then outputted.

The Inputs and Outputs in Figure 2.3 go through the same process: 1) embeddings are generated for every word in the sentences and 2) a positional encoding vector is summed to the word embedding vector. To generate the word embeddings, the Transformer randomly initializes the weight matrix and refines these weights during training.

2.3.2 Attention mechanism

In this section, we will go into more details regarding the Transformer's attention mechanism. But first, it is important to understand the intuition behind the idea of attention. Looking at Figure 2.4, we have an example of the word "it" attending to other words in the sentence "The animal didn't cross the street because it was too tired". We can see that the self-attention mechanism strongly associates "it" to the word "animal". This contextual information is important because the meaning of a word changes depending on its surroundings and this is needed to perform a correct translation. As another example, we could think about the word "apple" which depending on the sentence and context, it could mean the fruit "apple" or it could mean the company "Apple". This example illustrates how import the context is when representing a word as a word embedding. The French translation of the word "apple" could either be "pomme" or just "Apple".

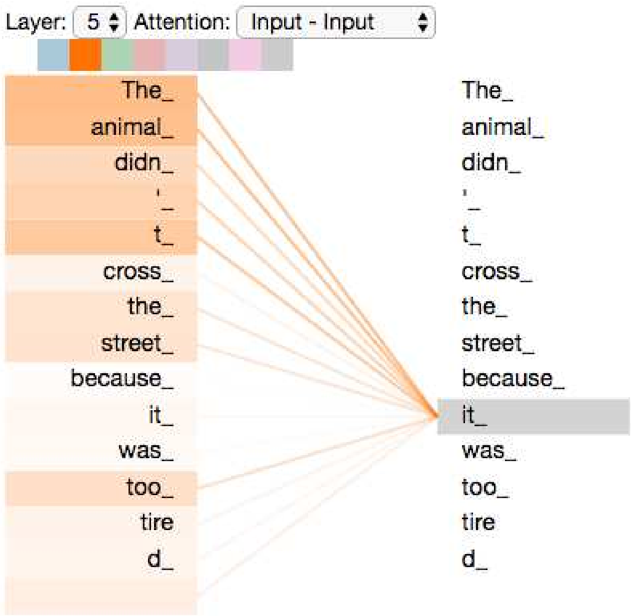


Figure 2.4: Self-attention mechanism illustrated.
Source: https://jalammar.github.io/illustrated-transformer/

This word-to-word attention is represented as a function of Q (query), K (key) and V (value). See Function 2.1 and Figure 2.5 (left).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

We can break down Function 2.1 into the following steps:

1. We start with Q, K and V matrices. These values for Q, K and V will change depending on the attention sub-layer as seen in Figure 2.3. For the first encoder layer, Q, K and V are equal to a matrix containing a stack of embeddings of every word in the input sentence.

2. Compute the dot product of Q and K (same as QK^T). This will produce a self-attention score for every pair of words in the sentence. Higher scores indicate stronger association or stronger attention from one word to the other.

3. Divide the self-attention scores by a scalar value $\sqrt{d_k}$, where d_k is the dimension of queries and keys. The authors of [53] use $d_k = 64$. The reason for this scaling is

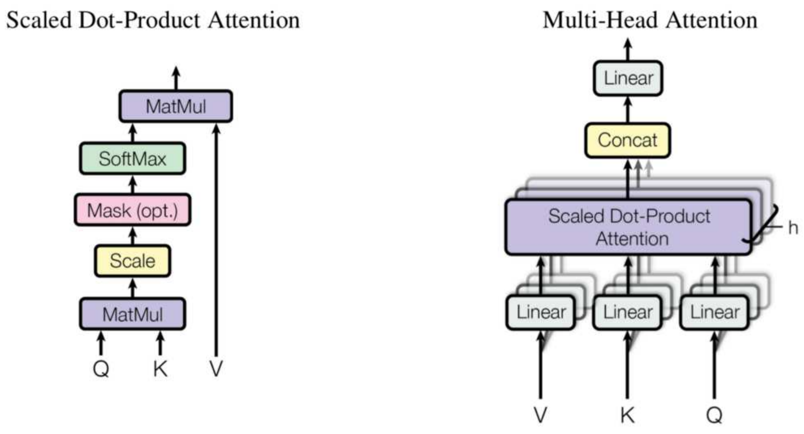


Figure 2.5: (left) Scaled Dot-Produce Attention. (right) Multi-Head Attention consisting of several attention layers running in parallel.
Extrated from [53]

that for larger values of d_k, the dot product of Q and K can grow large at the point of pushing the softmax function into regions where it has extremely small gradients.

4. Apply a softmax function which will map our scaled self-attention scores to the range of 0 and 1. These will serve as weights in the next step. The intuition here is that for each word, very small weights will drown-out other irrelevant words and while keeping the values of important words almost intact [2].

5. Multiply the weights matrix ($softmax(\frac{QK^T}{\sqrt{d_k}})$) with V. The end result will be a matrix where every word (each row) will be expressed as a sum of the weighted value vector of every word in the sentence. We can think of each row in this final matrix as a word embedding which captures the attention to every other word.

We have just explained the attention mechanism applied directly to the query, key and value matrices. We could call this a one-head attention mechanism. In [53], the authors refined this mechanism by adding a projection layer with $h = 8$ sets of matrices (W^Q, W^K and W^V) and therefore creating 8 separate attention heads. See Figure 2.5 (right). These projection matrices were initialized randomly and trained to produce different linear projections of Q, K and V. These projections would then be fed into the attention mechanism as already explained.

In other words, we are replicating the attention mechanism $h = 8$ times in different projections of (W^Q, W^K and W^V). We call each of these replicas a "head". Function 2.2 defines the attention for each head i, where *Attention* is the same as in Function 2.1.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.2}$$

The output of every head is then concatenated and projected with a matrix W^O as in Function 2.3.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.3}$$

All weight matrices W^O, W_i^Q, W_i^K, W_i^V are randomly initialized and adjusted during training.

To better understand the different steps in the multi-head attention mechanism, we can look at the dimensions of each variable and function in Table 2.1. Consider n the number of words in the sentence, d_{model} the size of the embeddings, d_k and d_v the dimension of the projections of K and V respectively, and h the number of attention heads.

	Dimension
Q	n x d_{model}
K	n x d_{model}
V	n x d_{model}
QK^T	n x n
$Attention(Q, K, V)$	n x d_{model}
W_i^Q	d_{model} x d_k
W_i^K	d_{model} x d_k
W_i^V	d_{model} x d_v
QW_i^Q	n x d_k
KW_i^K	n x d_k
VW_i^V	n x d_v
$Attention(QW_i^Q, KW_i^K, VW_i^V)$	n x d_v
$Concat(head_1, ..., head_h)$	n x hd_v
W^O	hd_v x d_{model}
$Multihead(Q, K, V)$	n x d_{model}

Table 2.1: Dimensions of each step of multi-head attention

2.3.3 Positional encoding

When dealing with sequential data, in particular with text, the position of a word in a sentence can completely change its meaning. Consider the sentences "The suspect killed the policeman" and "The policeman killed the suspect". Changing the position of the words "suspect" and "policeman" completely changes the semantics of each sentence. Recurrent neural networks naturally encode positional information by processing each word in a sequential manner. However, this is not the case for Transformers which ingest all words of a sentence in a single step. Positional encoding is used to inject information about the relative or absolute position of each token in the ingested sequence. For transformers, sine and cosine functions of different frequencies are used. See Equation 2.4. These positional encoding vectors have the same dimension d_{model} and are summed to the input embeddings to accomplish this.

$$
\begin{aligned}
PE_{(pos,2i)} &= sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \\
PE_{(pos,2i+1)} &= cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)
\end{aligned}
\tag{2.4}
$$

where pos is the position and i is the dimension.

2.4 BERT: Bidirectional Transformers

Following the publication on transformers [53], researchers at Google AI Language published the paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[16]. BERT stood out from existing language models in two ways. First, it used a masked language modeling task for pre-training, i.e. a percentage of the words in the sequence was masked using a special token [MASK] and the model was trained to predict this masked token. Second, BERT was a bidirectional language model, i.e. the model was able to "see" the context on the right and on the left of every token during training. Language models until then were either trained in only one direction (usually left-to-right) or were composed of a concatenation of two model independently trained (left-to-right and right-to-left). At the time of its publication, BERT achieved new state-of-the-art results on 11 natural language processing tasks and served as a base for numerous other models in the future[16].

The architecture of BERT is a multi-layer bidirectional Transformer encoder and its implementation is almost identical to the original Transformer [53]. Given that BERT is just an encoder without a decoder, it cannot perform sequence-to-sequence tasks such as text translation or text summarization. BERT is well suited for tasks where the output is of a fixed size such as classification (e.g. emotion detection), question and answering, named entity recognition.

To accommodate a diverse range of downstream tasks, BERT represents both a single sentence or a pair of sentences as one token sequence. This is accomplished by the use of a special separator token [SEP] and sentence embeddings. See Figure 2.6. This flexible representation is important given that depending on the task, the input may be one or two sentences. For example, an Emotion Detection task requires only a single sentence but a Question Answering (QA) task requires a pair of sentences <question, alternative of answer>. As illustrated in Figure 2.6, BERT can represent its input by summing token embeddings, sentence embeddings and transformer positional embeddings. The purpose of having the sentence embeddings is similar to the purpose of having transformer positional embeddings but on a sentence level and not on a token level.

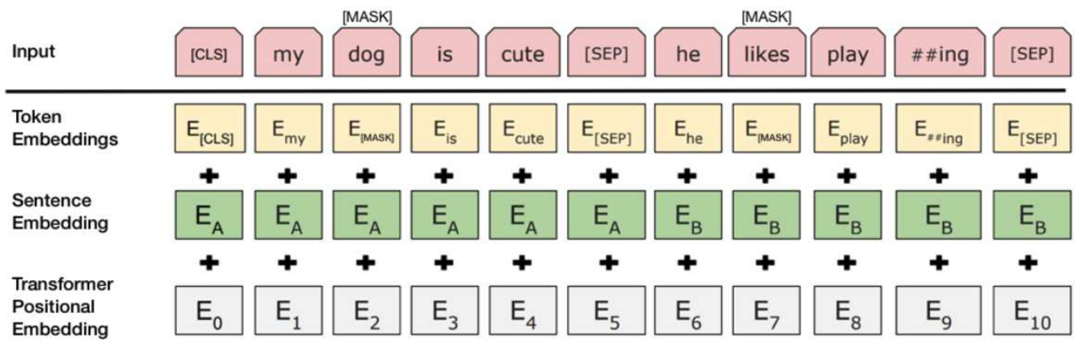


Figure 2.6: BERT embeddings.
Extracted from [16]

BERT is pre-trained jointly on two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, the authors randomly masked some of the tokens from the input and the model is trained to predict the original vocabulary id of the masked word based only on its context. In NSP, pairs of sentences are extracted from a monolingual corpus and the model is trained to predict if the second sentence follows the first sentence (labeled as $IsNext$) or not (labeled as $NotNext$). While the first task (MLM) focuses on learning the probability of a sequence of words occurring in a sentence; the second task (NSP) trains the model to understand sentence relationships, an important ability for downstreams tasks such as Question Answering (QA) ad Natural Language Inference (NLI). Lastly, it is important to note that both tasks are trained in a self-supervised fashion, i.e. the training data was produced from a corpus of unlabeled data.

2.5 BART and mBART

In 2019, Facebook AI published BART [30] which became the state-of-the-art model for abstractive summarization at the time. In its architecture, BART contains a bidirectional encoder and an auto-regressive decoder enabling the model for text generation tasks. BART is pre-trained to reconstruct artificially corrupted documents and optimizes for reconstruction loss, i.e. the cross-entropy between the decoder's output and the original document. [30] has explored both previously proposed as well as novel noising transformations techniques:

1. Token Masking: Random tokens are sampled and replaced with a [MASK] token. The model has to replace [MASK] with the most suitable token. This technique was proposed by BERT [16].

2. Token Deletion: Random tokens are deleted from the input. In contrast to token masking, the model has to decide which positions are missing inputs.

3. Text Infilling: A number of text spans are sampled, with span lengths drawn from a Poisson distribution ($\lambda = 3$). Each span is replaced with a single [MASK] token. 0-length spans correspond to the insertion of [MASK] tokens. Text infilling teaches the model to predict how many tokens are missing from a span.

4. Sentence Permutation: A document is split into sentences based on full stops. These sentences are then randomly shuffled.

5. Document Rotation: A token is chosen at random from a uniform distribution. The document is then rotated so that it begins with that token. This task trains the model to identify the start of the document.

These transformations are illustrated in Figure 2.7.

The transformations applied to the original document are arbitrary, including changing its length. After experimenting with a number of noising approaches, the best performance was found by both randomly shuffling the order of the original sentences and using the infilling technique, where arbitrary length spans of text (including zero length) are

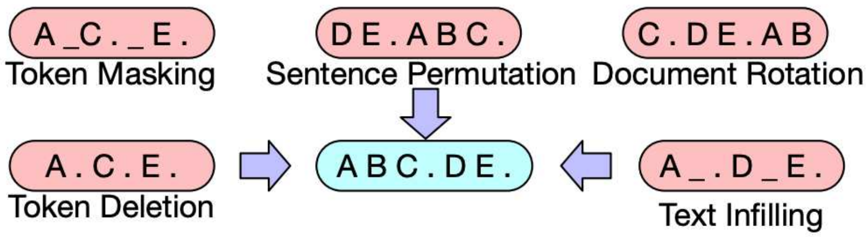


Figure 2.7: Transformations for noising the input during BART's training.
Source: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [30]

replaced with a single mask token. Hence, the model learns how to reason about overall sentence length and how to apply longer range transformations to the input.

Once pre-trained, BART was then fine-tuned in different downstream tasks including question answering, dialogue response and text summarization. For the summarization task, the authors have presented results on two well known datasets for English corpus: CNN/DailyMail [38] and XSum [39]. Both of them contained pairs of a news article and respective summary. However, the summaries in CNN/DailyMail tend to resemble source sentences whereas summaries in XSum are highly abstractive. BART outperformed previous works in both datasets using ROUGE [31] scores as a performance metric. On a qualitative analysis, the summaries generated by the model were fluent, grammatical English, highly abstractive (i.e. with few phrases copied from the original text) and generally factually accurate (i.e. statements in the summary are supported by the original text).

In 2020, Facebook AI expanded the work of BART to the multilingual domain. The authors of [32] published mBART, a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages. mBART was trained to reconstruct the original input text from artificially noised text. Two noising functions proposed in the BART work were used: (1) remove spans of text with arbitrary length and replace them with a <MASK> token and (2) permute the order of sentences within each instance. A standard sequence-to-sequence Transformer architecture [53] with 12 layers of encoder and 12 layers of decoder was used, totalling roughly 680M parameters. During pre-training, the decoder input is the original text with one position offset. A language id symbol <LID> is used as the initial token to start predicting a sentence.

Although mBART was fine-tuned and tested on a machine translation task, interesting insights about how multilingual models behave can be leveraged for our research. Firstly, mBART improved translation performance even with fine-tuning for languages that were not part of the pre-training corpora, suggesting that the pre-training has language universal aspects. Secondly, pre-training on more languages tended to improve performance when the target language monolingual data was limited. On the other hand, when monolingual data was plentiful, pre-training on more languages slight hurt performance which could be explained by the fact that additional languages could reduce the capacity available for each test language.

Two versions of mBART have been released: one trained on 25 languages [32] and another trained on 50 languages [52]. The latter includes the English and Portuguese languages.

2.6 XL-SUM

The lack of datasets for low and mid-resource languages - such as Japanese or Bengali - is one of the main reasons why most of the recent works on abstractive text summarization has focused primarily on high-resource languages like English. That is why the authors of [22] decided to explore abstract text summarization on a large-scale multilingual setting. Their work has made two important contributions. First, they collected and published a dataset comprising 1 million professionally annotated article-summary pairs from BBC in 44 languages ranging from low to high-resource, for many of which no public dataset is currently available. The dataset named XL-Sum is available on HuggingFace [1]. Second, they have trained a single model capable of summarizing articles on those 44 languages by finetuning the mT5 [57] with the XL-SUM dataset. This model is also available on HuggingFace[2].

Table 2.2 displays all languages and number of samples per languages contained in the XL-Sum dataset.

Language	**#Samples**	**Language**	**#Samples**	**Language**	**#Samples**
Amharic	5,461	Korean	4,281	Somali	5,636
Arabic	40,327	Kyrgyz	2,315	Spanish	44,413
Azerbaijani	7,332	Marathi	11,164	Swahili	10,005
Bengali	8,226	Nepali	5,286	Tamil	17,846
Burmese	5,002	Oromo	5,738	Telugu	11,308
Chinese	39,810	Pashto	15,274	Thai	6,928
English	301,444	Persian	25,783	Tigrinya	4,827
French	9,100	Pidgina	9,715	Turkish	29,510
Gujarati	9,665	Portuguese	23,521	Ukrainian	57,952
Hausa	6,313	Punjabi	8,678	Urdu	40,714
Hindi	51,715	Russian	52,712	Uzbek	4,944
Igbo	4,559	Scottish Gaelic	1,101	Vietnamese	23,468
Indonesian	44,170	Serbian (Cyrillic)	7,317	Welsh	11,596
Japanese	7,585	Serbian (Latin)	7,263	Yoruba	6,316
Kirundi	5,558	Sinhala	3,414	**Total**	**1,005,292**

Table 2.2: Languages covered by the XL-Sum dataset, and the number of samples for each language. Here, a sample denotes an article-summary pair.

For the English language, the finetuned mT5 model achieved a ROUGE-2 [31] of 15.18 while the state-of-the-art $PEGASUS_{\text{BASE}}$ model[3] [59] achieved a ROUGE-2 score 16.58

[1] https://huggingface.co/datasets/csebuetnlp/xlsum

[2] https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum

[3] We did not evaluate this model because it is a monolingual English-only summarization model. We are interested in benchmarking multilingual models capable of generating text at least in Portuguese and English.

on the XSUM English dataset, which is similar to XL-Sum. This comparison demonstrates that even though mT5 was finetuned to 44 languages, it is still producing competitive results against the monolingual $PEGASUS_{\text{BASE}}$ model.

Another interesting finding from this work was evidence that training a model multilingually results in positive transfer. The authors have trained 5 monolingual models on low-resource languages (Amharic, Azerbaijani, Bengali, Japanese, Swahili) using a compute-efficient setup and then compared ROUGE scores against the multilingual model. Results have shown that a multilingual model outperforms any of these monolingual models usually by a margin of 2 points in ROUGE-2. This phenomenon known as "positive transfer" [12] manifests when learning to perform a certain task in one language ends up improving learning to perform the same task in another similar language.

2.7 Longformer

Although Transformers [53] had estabilished state-of-the-art results in many NLP tasks, they still suffered from an important limitation. Due to their full self-attention mechanism, computational and memory requirements scaled quadratically with the sequence length which then prevented Transformers from processing very long sequences [5]. For example, BERT [16] was limited to input sequences no longer than 512 tokens. If the sequence was longer than this limit, a typical solution was to simply truncate the sequence and therefore lose part of the data during training and inference. Although this input length limitation may not be critical for news articles, the data loss caused by the truncate operation becomes significant when dealing with long documents such as scientific articles and podcast transcripts.

To address this problem, [5] proposed the Longformer, a Transformer-based model with a modified attention mechanism which scales linearly with sequence length, therefore capable of processing much longer sequences. This modified attention mechanism is a drop-in replacement of the standard self-attention mechanism and allows Longformers to process sequences with up to 16,000 tokens. The attention pattern employed is of a fixed-size window attention surrounding each token. Three variations of this pattern are used:

1. Sliding Window: Given a fixed window size w, each token attends to $\frac{1}{2}w$ tokens on each side. (see Figure 2.8b)

2. Dilated Sliding Window: Window has gaps of size dilation d. (see Figure 2.8c)

3. Global + Sliding Window: Global attention is used on few pre-selected input locations. This attention operation is symmetric: token with a global attention attends to all tokens across the sequence, and all tokens in the sequence attend to it. (see Figure 2.8d)

In [5], the authors have concluded that both attention types - windowed local and global - are important. Through ablation studies and controlled experiments, they observed that the local attention is primarily used to build contextual representations, while

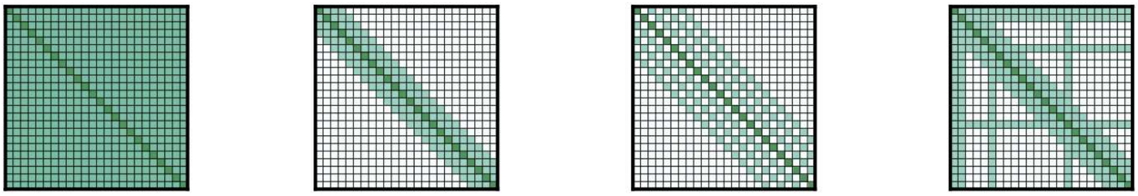


Figure 2.8: Comparing (a) the full self-attention pattern and (b)(c)(d) the configuration of attention patterns in Longformer.
Source: Longformer paper [5]

the global attention allows Longformer to build full sequence representations for prediction.

Although the original Longformer model proposed was an encoder-only transformer, an encoder-decoder variant named Longformer-Encoder-Decoder (LED)[4] was later released [5]. LED parameters were initialized with BART's [30] weights, and it follows BART's exact same architecture in terms of number of layers and hidden sizes. The authors evaluated LED on the summarization task using the arXiv summarization dataset [10] which focuses on long document summarization in the scientific domain and achieved state-of-the-art results, slightly outperforming BigBird [58].

Despite the encouraging results, experiments with Longformer and LED are still restricted to the English language.

2.8 TREC Podcasts Track

The Text Retrieval Conference (TREC) is a series of workshops with the goal of encouraging research in the Information Retrieval field based on large test collections. Since 1992, it is co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense. Following the publication of the Spotify Podcasts Dataset [9], TREC hosted the Podcast track in 2020 and 2021 and asked researchers from the fields of information retrieval, NLP and speech analysis, to explore this data from the perspective of two challenge tasks: segment retrieval and podcast episode summarization.

The segment retrieval task was to retrieve relevant two-minute segments of podcast audio given a query. Different types of queries were defined in 2020 and 2021. This task is out of the scope of this Master thesis.

The podcast episode summarization task was to return a short description for each given podcast episode. This short description should capture the most important information in the content and help the user decide whether to listen to that episode or not. The description should be provided in the format of a short text snippet in grammatical standalone utterances of significantly shorter length than the input episode description. This task precisely defines the goal of this Master thesis project. However, while TREC's

[4]https://huggingface.co/docs/transformers/model_doc/led

Podcast Track focused only on the English language, we have explored the same task from a multilingual perspective.

In TREC 2020[25], 8 participants submitted a total of 22 experiments for the summarization task. All of the experiments were based on abstractive summarization techniques and used some form of deep learning model. Some of the participants used an extractive technique as a filtering step in their summarization process. The work which most resembles our proposal is the LongformerBART [27], a modified version of BART [30] model where the attention mechanism has been replaced by the Longformer attention mechanism[5]. Karlbom H. has also published his work in his Master thesis [26] from the Uppsala University. The LongformerBART model is, from an architecture and weights initialization standpoint, identical to the LED model [5] which was only later released. When compared against all of the experiments submitted, the LongformerBART achieved the second highest ROUGE-L precision.

In TREC 2021[28], we have submitted 2 experiments for the podcast summarization task. Both of our experiments used a multilingual model fine-tuned to English and Portuguese podcasts. Our methodology and results have been published in [51].

2.9 Summary

In this chapter, we review the literature on text summarization, the different models used and previous work on podcast summarization.

First, we revisit the different types of automatic text summarization while clarifying the differences between extractive and abstractive summaries, multilingual and monolingual summarization systems and others. We then propose a supervised multilingual abstractive single-document method for generic domain-specific (podcast transcripts) summarization system.

Next, we revisit the deep learning architectures used for abstractive summarization since 2014. We look into sequence-to-sequence learning, transformers, BERT, BART, mBART, XL-SUM and Longformer. We highlight the advantages of some of these works against its predecessors. For example, how transformers are able to process long sequences in a single step while sequence-to-sequence models took one step per token. Also, how longformers are able to process sequences much longer than transformers thanks to their new attention mechanism which dramatically reduces memory consumption.

Finally, we discuss how the TREC conference have explored podcast summarization. In 2020, Hannes Karlbom submitted the LongformerBART - a modified version of BART using the Longformer attention mechanism. In 2021, we submitted two variants of the LongMBART model - a modified version of mBART using the Longformer attention mechanism and capable of summarizing podcasts in English and Portuguese.

Chapter 3

Methodology

3.1 Datasets

In this section, we are going to describe the datasets used during this research project. We will discuss the methodology used to produce them and we will provide some basic statistics on the data.

3.1.1 English Podcasts Dataset

This English Podcasts Dataset[9] consists of 105,362 podcast episodes in English from 18,376 shows. This includes nearly 60,000 hours of audio and accompanying transcripts produced using Google Cloud Speech-to-Text API[1]. It also includes the following metadata for each of the episodes:

- Episode URI: uniquely identifies the podcast episode.
- Episode name: the name of the episode provided by the podcast creator.
- Episode description: a short description of the episode provided by the podcast creator.
- Show URI: uniquely identifies the show.
- Show name: the name of the show provided by the podcast creator.
- Show description: a short description of the show.
- Show language: the language of the show defined in BCP-47 format.
- RSS link: a URL to an RSS feed containing all of the published episodes for a given show and their respective metadata.
- Episode duration: the duration of the episode in milliseconds.
- Publisher: company or creator publishing the podcast.

Episode Name	Mini: Eau de Thrift Store
Episode Description	ELY gets to the bottom of a familiar aroma with cleaning expert Jolie Kerr. Guest: Jolie Kerr, of Ask a Clean Person. Thanks to listener Theresa.
Publisher	Gimlet
RSS Link	https://feeds.megaphone.fm/elt-spot

Table 3.1: Sample of episode metadata taken from [9]

Table 3.1 displays some of the metadata for a sample episode.

A wide range of topics are covered in the podcasts: lifestyle & culture, storytelling, sports & recreation, news, health, documentary, and commentary. Moreover, the podcasts are presented in various structural formats, number of speakers, and levels of formality. Some are scripted, others improvised, and presented in the forms of narrative, conversation, or debate. [9]

In Table 3.2, we have descriptive statistics over four metrics: count of words per episode transcript, count of words per episode description, sentences per episode description and episode duration in minutes. Looking at words per transcript, we can see that three quarters of the episodes have more than 2000 words per transcript which reinforces our thesis that traditional transformers with a full self-attention mechanism will incur significant data loss with inputs limited to 512 or 1024 tokens. In terms of episode duration, 50% of the episodes are less than 31 minutes long and 75% of the episodes are less than 50 minutes long. Looking at the ratio of words per transcript and duration for the 1st, 2nd and 3rd quartiles, speech rate ranged between 157 and 173 words-per-minute (wpm). This is aligned with [47] where the average speech rate for English was 167.54 wpm. This was based on a sample of 10 bulletins broadcast in the BBC radio station. The mean size of the episode descriptions is 83 words, giving us an indication of how long the automatically generated episode descriptions will be as we use them as target summaries during training.

Let us now compare the English Podcasts Dataset with with the CNN/DM dataset [38], more specifically compare the size of the documents offered in each dataset. Looking at the source documents, the average length in CNN/DM is 766 words while the average length in the podcasts dataset is 5728 words (transcripts). When we look at the target summaries, the average length in the CNN/DM dataset is 53 words while the average length in the podcasts dataset is 83 words (episode descriptions). The difference for the target summaries' lengths is of 30 words which does not really impact the method used to produce summaries. However, there is a large difference between the length of source documents in each dataset, i.e. episode transcripts compared against CNN/DM news articles. Transcripts are more than 7 times longer than articles in CNN/DM. This comparison exposes the importance of a solution such as the Longformer [5] to deal with long documents.

We also analyzed the distribution of episodes per genre (see Table 3.3). The most

[1] https://cloud.google.com/speech-to-text/docs/video-model

	Words per transcript	Words per description	Sentences per description	Episode duration (minutes)
mean	5728.22	83.58	4.81	33.64
std	4152.76	80.16	4.16	22.71
min	11	1	0	0.00
25%	2043	31	2	13.37
50%	5194	59	4	31.42
75%	8671	109	6	50.17
max	43504	2341	126	304.95

Table 3.2: English Podcasts dataset: statistics on words per episode transcript, words per episode description and episode duration in minutes.

common genres in this dataset are Education, Sports, Health and Fitness, Business and Comedy. These top 5 genres account for more than half (54%) of all episodes in the dataset.

Genre	Episodes count
Education	13308
Sports	12458
Health & Fitness	11617
Business	10947
Comedy	9427
Society & Culture	8310
Religion & Spirituality	8040
TV & Film	5953
Leisure	5947
Arts	5021
Kids & Family	3253
True Crime	2509
Music	2309
Science	1785
Technology	1686
History	962
Fiction	706
News	575
None	316
Government	233

Table 3.3: English Podcasts dataset: number of episodes per genre.

The transcripts data are delivered in JSON format. In listing 3.1, we can see a transcript snippet. The transcript text is provided as an array of words and each word is annotated with a start and end timestamp. This extra annotation allows data users to crop specific segments of the transcript by time range.

Listing 3.1: Transcript snippet


```
[{"words": [
    {"word":"Welcome","speakerTag":"1","startTime":"1.1","endTime":"1.9"},
    {"word":"to","speakerTag":"1","startTime":"1.9","endTime":"2.1"},
    {"word":"dissipation.","speakerTag":"1","startTime":"2.1","endTime":"3.1"},
    {"word":"Thank","speakerTag":"1","startTime":"3681.6","endTime":"3682.1"},
    {"word":"you","speakerTag":"1","startTime":"3682.1","endTime":"3682.4"},
    {"word":"very","speakerTag":"1","startTime":"3682.4","endTime":"3682.8"},
    {"word":"much","speakerTag":"1","startTime":"3682.8","endTime":"3683.3"},
    {"word":"for","speakerTag":"1","startTime":"3683.3","endTime":"3683.4"},
    {"word":"listening","speakerTag":"1","startTime":"3683.4","endTime":"3684.2"},
```


3.1.2 Portuguese Podcasts Dataset

This dataset consists of 123,054 podcast episodes in Portuguese from 16,131 shows. The episodes were sampled uniformly at random sampled from all episodes published between September 9, 2019 and March 31, 2022. In total, this dataset offers more than 76,000 hours of speech audio. The same metadata provided by the English Podcasts Dataset [9] is also available in the Portuguese Podcasts Dataset.

The process to build the Portuguese Podcasts dataset was the same by [9] with the following differences:

- The language of the show was Portuguese (pt-BR or pt-PT)
- Episodes were transcribed with Azure's speech-to-text service
- When requesting a transcription, we had to specify the target language, i.e. either pt-PT (Portuguese from Portugal) or pt-BR (Brazilian Portuguese). We used a number of heuristics based on the metadata to make this decision. In the end, 114,387 episodes were transcribed with pt-BR and 8667 were transcribed with pt-PT. We used pt-BR as a fallback option because the number of podcast producers in Brazil is larger.

Table 3.4 presents descriptive statistics regarding the number of words per episode transcript, the number of words per episode description, sentences per episode description and the episode duration. If we compare the statistics from English podcasts (Table 3.2) and Portuguese podcasts (Table 3.4, we notice that the median value of words per episode description and episode duration are very similar. However, the median number of words per episode transcript is 29% higher (English: 5194 | Portuguese: 6746). Given that the median duration is 31 minutes - the same as the English dataset - we can infer that the speech rate is higher for Portuguese podcasts. Looking at the ratio of words per transcript and duration for the 1st, 2nd and 3rd quartiles, speech rate ranges between 220 and 248 words-per-minute (wpm). We have to consider that automatic transcript have a word error rate so this speech rate is only an approximation.

In Table 3.5, we can see the distribution of episodes per genre. The top 5 genres account for 69% of all episodes. Business, Education, Sports and Comedy are part of the top 5 genres of both the English and Portuguese podcasts dataset.

The transcripts data are delivered in JSON format. In listing 3.2, we can see a transcript snippet. The transcript text is provided as an array of words and each word is annotated with a start and end timestamp. This extra annotation allows data users to crop specific segments of the transcript by time range.

	Words per transcript	Words per description	Sentences per description	Episode duration (minutes)
mean	9539.15	71.45	4.10	37.29
std	9976.02	64.99	3.49	32.78
min	0	1	1	0.24
25%	2203	28	2	10.87
50%	6746	54	3	31.23
75%	13692.75	92	5	55.04
max	205163	890	88	694.50

Table 3.4: Portuguese Podcasts dataset: statistics about words per episode transcript, words per episode description, sentences per episode description and episode duration in minutes.

Listing 3.2: Transcript snippet


```
{"words":[
  {"word":"alô","start_time_secs":6.39,"end_time_secs":6.94},
  {"word":"olá","start_time_secs":7.69,"end_time_secs":8.46},
  {"word":"andré","start_time_secs":8.47,"end_time_secs":8.93},
  {"word":"fran","start_time_secs":8.94,"end_time_secs":9.21},
  {"word":"tudo","start_time_secs":9.22,"end_time_secs":9.42},
  {"word":"bem","start_time_secs":9.43,"end_time_secs":9.57},
  {"word":"com","start_time_secs":9.58,"end_time_secs":9.69},
  {"word":"você","start_time_secs":9.7,"end_time_secs":10.32},
...
```


We have built this dataset as part of this Master thesis project and decided to release it as one of our contributions to the research community. We believe this dataset is a first step towards a multilingual perspective in the podcasts domain and that it will benefit all of the NLP researchers who lack text and audio datasets in Portuguese. It is important to note that the applicability of the Spotify Portuguese Podcasts dataset is not restricted to the podcast summarization problem. Besides search and summarization, explored in TREC 2020 [25] and 2021 [28], this data is valuable for tasks such as document segmentation or dialog modeling, and enables the exploration of new avenues in speech and language research. The same applies to the English Podcasts dataset [9].

3.1.3 Boilerplate annotations in episode descriptions in Portuguese

Episode descriptions are important to users when deciding if they want to listen to a given episode or not. These descriptions often mention the participants of the episode and the topic of discussion. However, podcast creators frequently use episodes descriptions to promote social media links or advertisements. We call such promotional content as "boilerplate" and automatically detecting such content has been already studied in [44]. This is an important task because we do not want our models to mimic this behavior of generating boilerplate which would most likely end up in hallucinating URLs or social media handles. Therefore, we have decided to remove boilerplate from episode descriptions

Genre	Episodes count
Business	26915
Education	23541
Sports	13467
Comedy	11211
Arts	9869
TV & Film	5451
Science	5379
Music	3952
Technology	3193
Society & Culture	3192
Kids & Family	3036
Leisure	2756
Health & Fitness	2337
History	2254
Fiction	2024
True Crime	1749
News	1098
Religion & Spirituality	1037
Government	531
None	62

Table 3.5: Portuguese Podcasts dataset: number of episodes per genre

as a pre-processing step. However, no dataset for podcast boilerplate was available in Portuguese, so we created it by manually annotating a random sample of 1000 episode descriptions. In order to increase variability in the descriptions, we also filtered out the top 5 most common repeated descriptions. For example, the Horoscope show releases episodes on a daily basis but always uses the same description "Ouça o episódio de hoje para saber tudo sobre o seu dia...".

We used the tool doccano[2] to annotate the sentences containing boilerplate. We considered any of the following as boilerplate:

- Contact information such as "Send in voice message http://anchor.com/foobar"
- Social media promotion
- Advertisements
- Technical staff information such as producer, editor, sound technician.
- Hashtags to characterize or promote the content
- Credits to the soundtrack used during the episode
- Time marks such as "0:30 <topic 1> 1:25 <topic 2> 5:40 <topic 3>"
- License information such as Creative Commons license

[2]https://github.com/doccano/doccano

- Social media handles such as "@ladygaga"

The final dataset was comprised of 5145 sentences where 1404 were boilerplate sentences and 3741 were not. In table 3.6, we list a few examples of sentences annotated as boilerplate.

This dataset was later used to remove boilerplate from episode descriptions by training a binary sentence classifier which detects such content. In the next section, we will go into more details about the boilerplate detection method.

Boilerplate examples
Não deixe de nos seguir nas redes sociais através do @aceleraecarreira no Instagram e Facebook.
Para saber mais sobre o Projeto DEFCOM, acesse: __LINK__
O Toni esta no instigaram (@toni__vicente) e podem encontrar o canal dele __LINK__ | Falem connosco e deixem as vossas sugestões no twitter @techuntalked
Send in a voice message: __LINK__
Linkedin: __LINK__ Instagram: __LINK__ Canal no Telegram: t.me/productgurus
Ju & Ric - Divagando Pelo Mundo: __LINK__/divagandopelomundo *** Novas regras para visitar o Camboja: *** Mande um recado para o Viajão! __

Table 3.6: Examples of sentences annotated as boilerplate in Portuguese.

3.2 Pre-processing the data

The Spotify Podcast datasets described in section 3.1 were built with minimal processing in order to maximize its potential use in speech and natural language research. We carefully thought about every filtering step in order to avoid bias that would lead to results only reproducible in certain niches of podcasts. The downside is that the dataset come with data quality issues which may negatively affect the summarization models. One example of that is the advertisements found in episode descriptions as mentioned in subsection 3.1.3.

In order to clean the data, we have applied the following filters:

- We remove episodes with repeated descriptions (any description used in more than one episode). We applied a TF-IDF vectorization of the descriptions which were compared to each other using the cosine distance. Any data points with too similar descriptions (threshold 95%) were filtered out.
- We remove episodes where the episode description too similar to the show description (threshold 95%). This is an indication that the creator did not thoughtfully write the episode description but instead simply copied the show description to fill in this metadata.
- We remove any emails or URLs from episode descriptions as we did not want our trained models to hallucinate such information in the generated summaries.
- We remove boilerplate content from episode descriptions. We will describe this step in more details in section 3.3.
- We remove episodes where the creator descriptions are either too long or too short with the boundary conditions set to between 10 and 1300 characters.

After applying the filters above, we split the remaining data into 3 parts: train (90%), dev (5%) and test (5%). The train set was used to train our model, the dev set - sometimes called evaluation set - was used to evaluate the model after each N steps of training and the test was held out for a final evaluation once the model finished training. We report our results on the test set. We can see the number of episodes for each split after pre-processing in Table 3.2.

For the data split, we first grouped the episodes per podcast show and then assigned each group to a particular split. In other words, any two episodes of the same show were always in the same split. We believe this grouping strategy will prevent data leakage as the language style of episode descriptions in the same show tend to be similar.

	ratio	**EN**	**PT**
train	90%	80895	90859
dev	5%	4503	5073
test	5%	4511	5058

Table 3.7: Pre-processed data: split size by number of episodes

3.3 Boilerplate detection

Briefly speaking, boilerplate is any extraneous content which does not describe the episode in natural language text. Common cases of boilerate in podcasts are advertisements and promotional content for social media. [44] studied the problem of detecting boilerplate in both episode transcripts and episode descriptions. In this Master thesis, we focus only in detecting boilerplate in episode descriptions. We consider boilerplate detection and removal as an important step towards producing high quality podcast summaries because (1) a model trained with boilerplate material will reproduce such content and very likely hallucinate URLs, social media handles and emails, and (2) the mere presence of boilerplate hurts the purpose of a summary which is to surface the topic of an episode and bring context around it by adding unnecessary - and likely false - information.

For English podcasts, we leveraged an existing binary classifier trained to detect boilerplate on a sentence basis. The base model used was the *bert-base-cased*.

For Portuguese podcasts, we trained a new binary classifier from scratch following the same protocol as the one used to train the English boilerplate detector. Firstly, we manually annotated 1000 episode descriptions following the definition in subsection 3.1.3. Secondly, we broke down the episode descriptions into sentences using the Spacy library[3]. Lastly, we finetuned *bert-base-multilingual-cased* to classify each sentence as "contains boilerplate" or "does not contain boilerplate" using 1000 episode descriptions.

We used the doccano tool to manually annotated 1000 episode descriptions (see Figure 3.1).

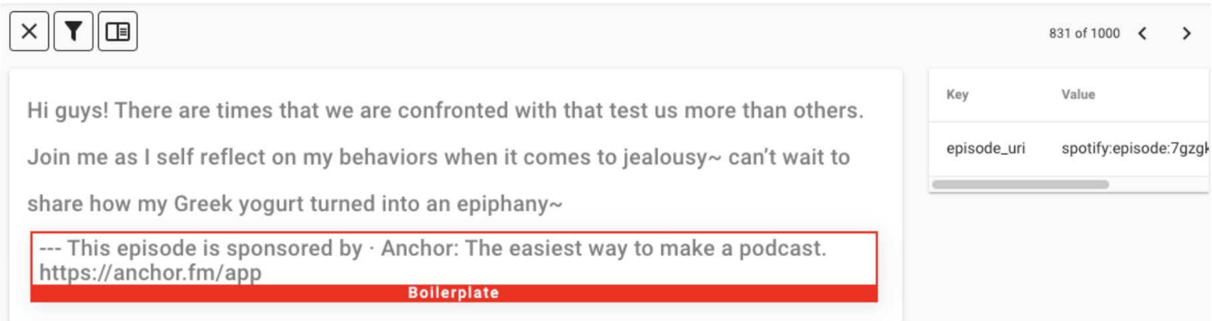


Figure 3.1: Doccano tool used to annotate boilerplate spans in 1000 episode descriptions.

We tested this model by removing any content classified as boilerplate in a held out set of 100 episode descriptions. We analyzed the episode descriptions before and after this removal and the results were:

- 95 episode descriptions were correctly cleaned, i.e. all boilerplate was removed without losing any legitimate non-boilerplate content.
- 2 episode descriptions were still left with some boilerplate.
- 3 episode descriptions lost legitimate non-boilerplate content.

[3]https://spacy.io/

Given that our boilerplate detector correctly cleaned episode descriptions in 95% of the cases, we then applied the same cleaning process to all of our train set. These descriptions are all in Portuguese.

Table 3.8 shows some examples before and after cleaning episode descriptions with the boilerplate detector.

Original episode description	**Description after removing boilerplate**
Como dito, faremos uma temporada especial, em parceria com a Directa Consultoria, para falar da pandemia. Nesse primeiro episódio, conversamos com o prefeito de Botucatu, Mario Pardini. — **Send in a voice message: https://anchor.fm/meiahoradomoro/message**	Como dito, faremos uma temporada especial, em parceria com a Directa Consultoria, para falar da pandemia. Nesse primeiro episódio, conversamos com o prefeito de Botucatu, Mario Pardini.
No episódio de hoje nossa bancada reuniu algumas dicas preciosas pra você melhorar a qualidade do seu treinamento baseados no BATMAN. Isso mesmo, no morcegão! Reza a lenda que o Batman mesmo sendo apenas humano é o único capaz de derrotar toda a liga da justiça, simplesmente pelo fato dele observar os pontos fortes e fracos dos seus companheiros de luta por justiça. Parece engraçado ou jocoso mas faz muito sentido. **Espero que gostem das dicas Não esqueça de nos seguir nas redes sociais: @laelrodrigues @kingarthurbjj @kellysaraf @plynnio Oss**	No episódio de hoje nossa bancada reuniu algumas dicas preciosas pra você melhorar a qualidade do seu treinamento baseados no BATMAN. Isso mesmo, no morcegão! Reza a lenda que o Batman mesmo sendo apenas humano é o único capaz de derrotar toda a liga da justiça, simplesmente pelo fato dele observar os pontos fortes e fracos dos seus companheiros de luta por justiça. Parece engraçado ou jocoso mas faz muito sentido.
As inúmeras baixas na equipe econômica do ministro Paulo Guedes deixam o setor produtivo e o mercado financeiro cautelosos. Eles temem que as mudanças estruturais que o Brasil tanto precisa não sejam colocadas em prática. Vamos conversar com o ex-secretário de Desestatização do governo Bolsonaro, Salim Mattar. Participam da conversa o CEO do Banco Genial, André Schwartz, e o economista-chefe da Genial Investimentos, José Márcio Camargo. **COMECE A INVESTIR AGORA, ABRA SUA CONTA GRATUITA NA GENIAL! ~https://genial.vc/abrasuaconta-yt**	As inúmeras baixas na equipe econômica do ministro Paulo Guedes deixam o setor produtivo e o mercado financeiro cautelosos. Eles temem que as mudanças estruturais que o Brasil tanto precisa não sejam colocadas em prática. Vamos conversar com o ex-secretário de Desestatização do governo Bolsonaro, Salim Mattar. Participam da conversa o CEO do Banco Genial, André Schwartz, e o economista-chefe da Genial Investimentos, José Márcio Camargo.

Table 3.8: Examples of episode descriptions in Portuguese after removing boilerplate with trained boilerplate detector.

3.4 Podcast Summarization methods

In this section, we will explain the set of baseline methods used for the summarization task and then go over each of the different methods used to train our own summarization models.

3.4.1 Baseline methods

First minute

We use a simple baseline which does not require any model training but relies on the fact that every word in a podcast transcripts is timestamped. For a given episode, we generate its summary by simply cropping the portion relative to the first minute of the episode's transcript. It is expected that these naive summaries will not meet the basic standards of proper written language because (1) the cropping will cut the last sentence in an arbitrary way and likely leave it unfinished, and (2) transcripts often contain mispunctuations and other errors inherent to automatic speech recognition (ASR). Nevertheless, this is a simple inexpensive method to establish a baseline and it has been reported previously in the literature [9] [27] [25] [28].

Previous works on text summarization often use the lead-N method as a baseline where the first N sentences are extracted and considered to be the summary of a document [30] [59] [40]. However, given that sentence boundaries are often of poor quality in transcripts, it is reasonable to use time, as opposed to sentences, as the capping limit to generate the summaries.

Other works in text summarization have used random selection of N sentences but this method has not been used in podcast summarization. We believe that random selection of sentences would most likely just generate a very poor summary due to the fact that transcripts are very long documents and again, sentences boundaries are often incorrect in transcripts. For these reasons, we have decided not to use such a baseline method.

TextRank

TextRank [37] is an extractive summarization model. It is a graph-based model which can be used as an unsupervised method to extract both keywords or key sentences.

We describe the steps of the TextRank algorithm while pointing out the peculiarities of our use case:

1. We break down the document into text units and add them as vertices in the graph. The granularity of the text unit depends on the task at hand (e.g. words or sentences). For extractive summarization, we will break down the document into sentences.

2. We identify relations between the text units and use edges to represent such relations in the graph. The edges can be directed or undirected, weighted or unweighted. For extractive summarization, the relation between every pair of sentences is defined as the number of common words divided by the sum of each sentence length (see Equation 3.1).

3. We iterate using a graph-based ranking algorithm until convergence. We used the sumy[4] implementation which leverages the PageRank algorithm [7]. At the end of this step, every vertex should have a score.

4. We sort the vertices according to their final score and select the desired cohort for the task at hand. In our experiments, we selected the top 4 and top 5 most relevant sentences. These parameters were chosen based on the median number of sentences per episode description as described in tables 3.2 and 3.4.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{3.1}$$

There are two downsides when using an extractive summarization method on transcripts: (1) Sentence boundaries are often incorrect in episode transcripts and as the model stitches the top key sentences, this may lead to summaries with half-sentence parts, (2) unlike some abstractive summarization methods which can denoise errors from the ASR process, extractive summarization methods simply copies and pastes pieces of text.

On the other hand, extractive summarization are less prone to hallucinations and allows us to compose a summary as a collage of the most important bits of audio in the episode.

XLSUM

In [22], the authors built a dataset containing 1 million article-summary pairs in 44 languages and finetuned mT5 [57] to experiment on the multilingual summarization task. The articles were extracted from BBC and the article-summary pairs were professionally annotated. Both the dataset and the models have been released[5] to the research community. English and Portuguese are listed among the 44 languages included in the XL-SUM dataset.

We leverage the finetuned mT5[6] without further finetuning as a baseline model for the podcast summarization task. This model is the one with the highest overlap with our work because it is an abstractive multilingual summarization model based on transformers.

MBART

MBART [52] is the multilingual version of BART [30]. We chose to use the MBART-50[32] model because it has been pre-trained in 50 languages (including Portuguese and English) and also because it is an encoder-decoder model, i.e. capable of generating text.

We experimented with two versions of MBART-50[7]: (1) vanilla MBART without any finetuning and (2) MBART finetuned on the podcasts data in English and Portuguese. MBART-50 is a translation model so the intention of using the vanilla MBART is just to compare it against the finetuned MBART.

[4]https://github.com/miso-belica/sumy
[5]https://github.com/csebuetnlp/xl-sum
[6]https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
[7]https://huggingface.co/docs/transformers/main/model_doc/mbart

Parameter	**Value**
metric_for_best_model	ROUGE-2 F1 score
early_stopping_patience	3
truncate encoder	512 tokens
truncate decoder	128 tokens
batch size	5
eval_steps	2500
save_steps	5000
warmup_steps	1500

Table 3.9: Parameters used to finetune MBART

We experimented with MBART limited to 512 tokens and we trained on a machine with 1 GPU NVIDIA Tesla V100. The finetuning process was configured to stop whenever the ROUGE-2 F1 score did not increase after 3 evaluation rounds. All parameters used to finetune MBART are in Table 3.9.

3.4.2 LongMBART

In this experiment, we leverage the attention mechanism provided by Longformer [5]. We convert a MBART model into LongMBART by replacing the original full attention mechanism with Longformer's attention mechanism.

Converting MBART to LongMBART

The authors of [5] provided a notebook[8] in the Longformer github repo demonstrating how to convert the RoBERTA [33] model into a Longformer model. However, the code in this notebook was outdated with the Hugging Face's transformers version used by MBART-50 and also the LongformerEncoderDecoder model which contained the Longformer self attention component we needed. As a result, we had to spend quite a lot of time adapting the interfaces to match a newer version of the transformers library. Overall, converting MBART to LongMBART turned out to be a complex process and it required an entire week of full-time dedication.

From an architecture perspective, we start with the original MBART-50 model and then we extend the positional embeddings from 1024 to 4096 by copying the embeddings from MBART-50 multiple times. Although [5] mentions an encoder with a 16000 tokens limit, we had to choose 4096 tokens due to GPU memory constraints. Next, we replace the original full attention mechanism with Longformer's attention mechanism. We can see this change and the difference between the original MBART model (on the right) and the LongMBART model (on the left) in Figure 3.2. We replace the *MBARTAttention* with our own *LongformerSelfAttentionForMBART*. *LongformerSelfAttentionForMBART* is just a wrapper of the *LEDEncoderSelfAttention* which really contains the Longformer self attention mechanism. Since the encoder is a stack of self attention layers, we see *LongformerSelfAttentionForMBART* multiple times.

We apply no modifications to the decoder of MBART-50.

[8]https://colab.research.google.com/github/allenai/longformer/blob/master/scripts/convert_model_to_long.ipynb

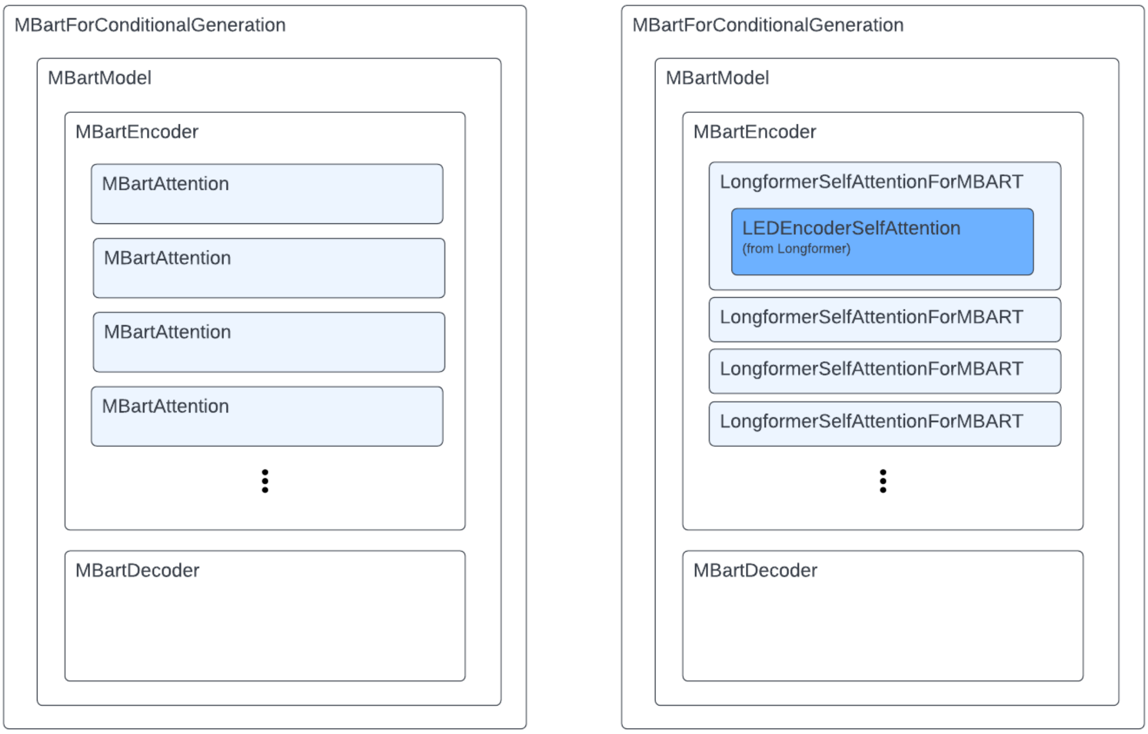


Figure 3.2: Converting MBART to LongMBART from an architecture perspective. We replace the original full attention mechanism (on the left) with Longformer's attention mechanism (on the right).

Finetuning

The next step was to finetune the model to the task of podcast summarization. We developed two distinct models with different finetuning strategies:

1. **One-round finetuning**: Finetuned only using episode descriptions and episode transcripts. Although mBART is a sequence-to-sequence model, it is initially trained to the machine translation task so we wanted to verify if finetuning it directly into the podcast summarization task would be successful. The training was set to early stop once the ROUGE-2 score didn't improve after 3 validation checkpoints.

2. **Two-rounds finetuning**: Finetuned initially on news articles from the XLSUM dataset[22] and then subsequently finetuned on episode descriptions and episode transcripts. The intuition here is that in the first round of finetuning, the model should learn how to summarize using high-quality news article-summary pairs. In other words, we expected mBART to transition from a neural machine translation model to a news summarization model. With the second round of finetuning, the model would then learn how to summarize podcast transcripts specifically. The training was set to early stop once the ROUGE-2 score didn't improve after 3 validation checkpoints.

In Figure 3.3, we can see the training and inference processes. For training, the model is trained on pairs of episode transcript-episode descriptions. Portuguese and English data are intermingled in order to avoid catastrophic forgetting[19]. During inference time, we pass either an episode transcript in Portuguese or English and the model generates a summary in the same language as the input. The source language and target language are specified as parameters when calling the inference function.

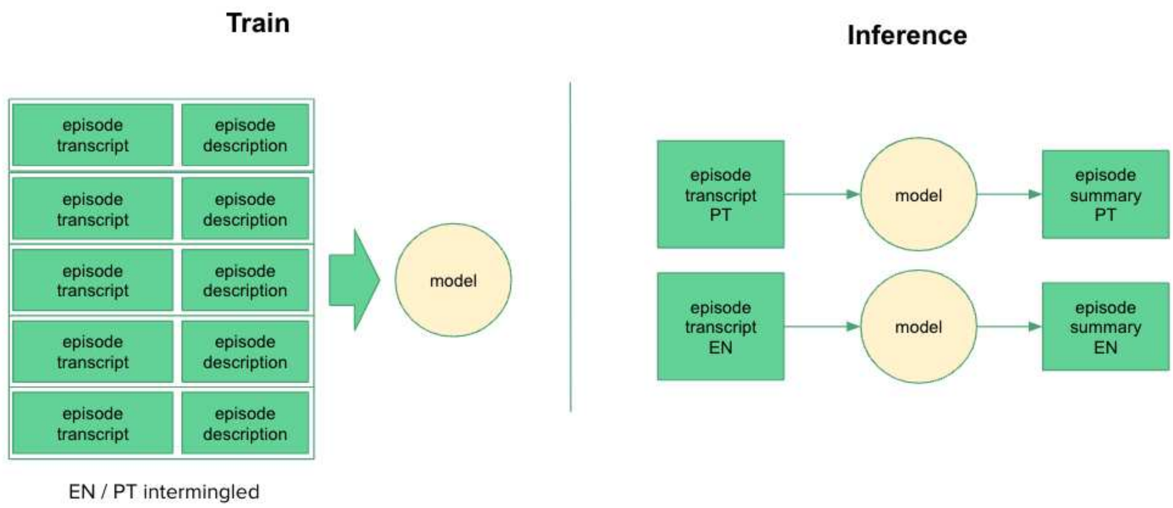


Figure 3.3: Training and Inference processes.

In Figure 3.4, we can see the full pipeline. Both Portuguese and English datasets go through the preprocessing step. We then use the preprocessed data to finetune a pre-trained model. Each finetuned model is then submitted to evaluation.

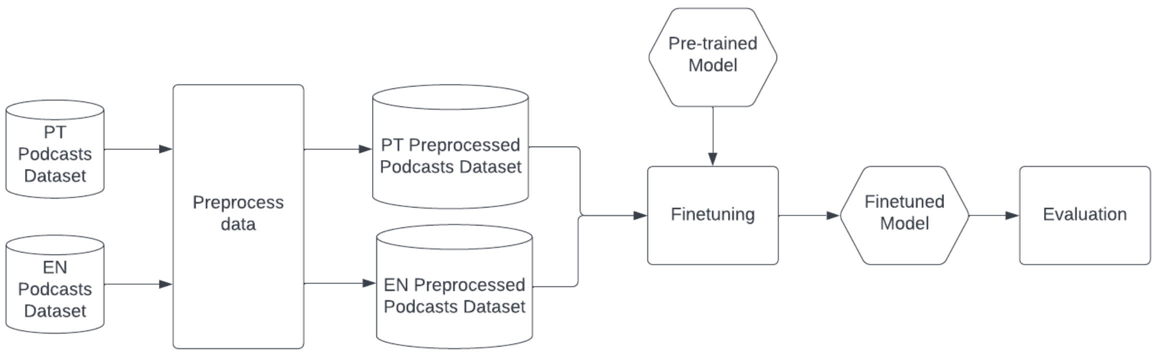


Figure 3.4: Full Pipeline.

3.4.3 Cross-lingual transfer learning

Cross-lingual transfer learning (CLTL) is a method used to build NLP models for low-resource target languages by leveraging labeled data from other (source) languages [8]. For example, we may train a model to summarize text in English but evaluate it to perform the same task on a different language. Most studies consider similar languages (e.g. English-German) and avoid distant languages (e.g. English-Japanese), since it is more challenging to conduct CLTL between distant languages than between similar language [55].

In this work, we explore CLTL for the task of podcast summarization. We finetune MBART and LongMBART using only data in English and later evaluate its performance when summarizing podcasts in Portuguese. We perform the same experiment but finetuning the model in Portuguese and evaluating it in English. Our hypothesis is that learning the task of podcast summarization is transferable across languages.

To evaluate this hypothesis, let us take two distinct languages A and B, where language A is the source language (i.e. language of the training data) and language B is the target language (i.e. language of the evaluation data). We will compare the performance of 4 models:

1. Model 1: the vanilla model
2. Model 2: model only finetuned to language A on podcast summarization
3. Model 3: model only finetuned to language B on podcast summarization
4. Model 4: model finetuned to languages A and B on podcast summarization

We use MBART and LongMBART as our models for this experiment. All of the variants we trained are illustrated in the model tree in Figure 3.5. We evaluate all of these variants on the English and the Portuguese podcast datasets.

Our expectation is that the model 3 or 4 will have the best performance since they were both trained on the target language. Moreover, we expect model 1 to provide the worst performance since both vanilla MBART and vanilla LongMBART are machine translation models and not summarization models. We could say that model 1 knows how to read and write the language but does not know how to summarize podcasts. And finally, we expect model 2 to perform better than model 1 thanks to cross-lingual transfer learning. In other words, model 2 should learn the task of summarization in language A and still perform fairly well when summarizing in language B.

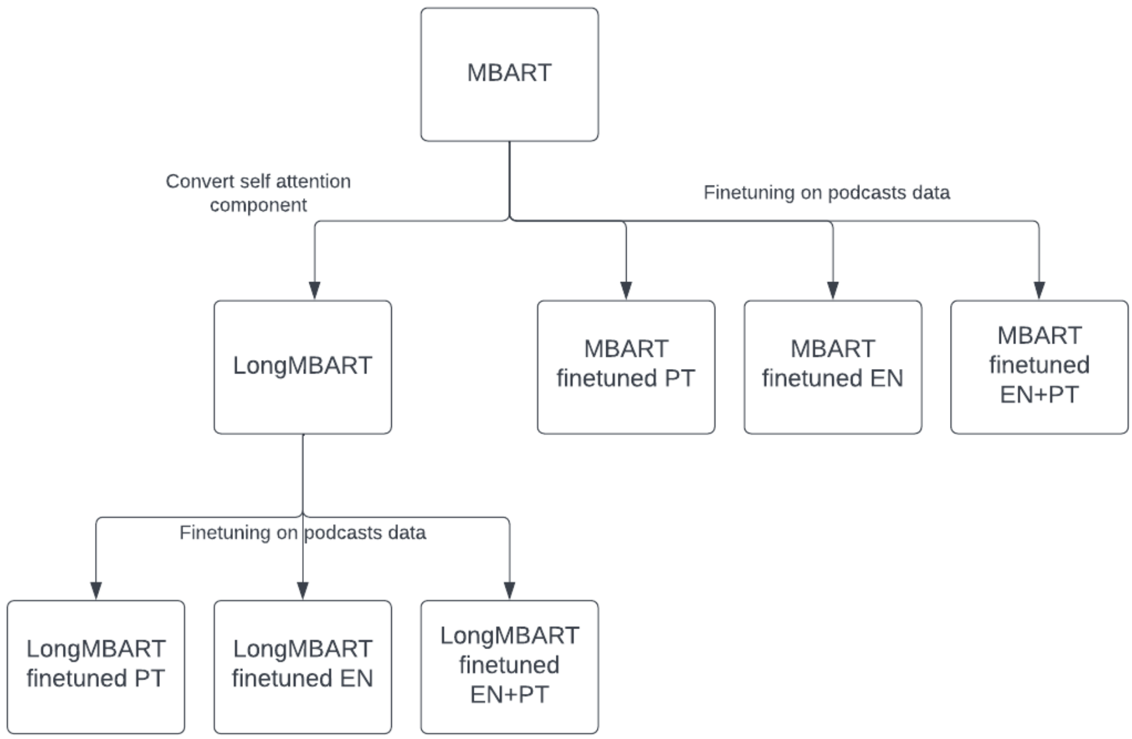
3.5 Evaluation

In this section, we will go through the evaluation process. Unlike supervised classification problems, there is not a single ground truth reference to be evaluated against. For a given episode, there are many different ways to summarize it. This plural characteristic of summarization is what makes its evaluation so challenging. For this Master thesis, we used two evaluation methods which we will detail in following subsections. In subsection 3.5.1, we discuss the ROUGE [31] metric and how it was used for intrinsic evaluation. In subsection 3.5.2, we discuss our human evaluation method and how specialists evaluated summaries during the TREC 2021 conference[28].

3.5.1 ROUGE scores

ROUGE [31] stands for Recall-Oriented Understudy for Gisting Evaluation. It measures the quality of a summary by counting the number of overlapping textual units against

Figure 3.5: All of the MBART and LongMBART variations used in this work. The arrows indicate a change applied on a base model to generate a new model.



another gold reference summary. Higher ROUGE scores indicates higher similarity between the candidate summary and the reference summary. Therefore, when it comes to ROUGE scores, the higher the better. Although other metrics have been developed to evaluate summarization [18], ROUGE still remains the most used. In a recent survey in Neural Abstractive Text Summarization[50], the most common evaluation metric was ROUGE by far.

For the Podcast Summarization task, as there are no gold reference summaries available, we resource to episode descriptions provided by the podcast creators. Even though these descriptions vary widely in scope and were not always written to act as summaries of an episode, we consider them the closest proxies available to reference summaries [25].

There are different flavors of ROUGE and each one uses a specific textual unit (e.g. unigrams, bigrams). We are going to look at two types of ROUGE scores: ROUGE-N and ROUGE-L.

ROUGE-N

ROUGE-N counts overlapping n-grams between the generated summary and the gold reference summary where N=1 (unigrams) and N=2 (bigrams) are commonly reported in text summarization. [31] defines ROUGE-N as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)} \tag{3.2}$$

Where n stands for the length of the n-gram $gram_n$, $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries and $Count_(gram_n)$ is the total count of n-grams.

Although the original paper [31] accounts for the scenario where there are multiple gold reference summaries for one summary being evaluated, this is not our case. We only have one reference summary for each episode so Equation 3.2 can be simplified to:

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{gram_n \in S} Count(gram_n)} \tag{3.3}$$

The denominator of equation 3.3 can change depending on the reference taken (precision or recall). The **precision** metric considers $Count_(gram_n)$ as the total number of n-grams in the candidate summary. The **recall** metric considers $Count_(gram_n)$ as the total number of n-grams in the reference summary. Finally, it is useful to calculate the **F1 score** which combines the precision and recall into a single metric by taking their harmonic mean. Let us assume that X is a reference summary with m n-grams, Y is a candidate summary with n n-grams and $Count_{match}(X, Y)$ is the number of overlapping n-grams in X and Y. Then, we can define precision, recall and F1 for ROUGE-N as follows:

$$Precision_{RougeN} = \frac{Count_{match}(X, Y)}{n} \tag{3.4}$$

$$Recall_{RougeN} = \frac{Count_{match}(X, Y)}{m} \tag{3.5}$$

$$F1_{RougeN} = \frac{2 * Precision_{RougeN} * Recall_{RougeN}}{Precision_{RougeN} + Recall_{RougeN}} \tag{3.6}$$

ROUGE-L

We also use another type of ROUGE called ROUGE-L, which measures the longest matching sequence of words using LCS (longest common sequence)[14] between the candidate and the reference summary. The longer the LCS, the more similar the candidate summary is to the reference summary. Instead of counting overlappin n-grams, ROUGE-L uses the number of words in the LCS as the measuring unit.

Let us now define ROUGE-L with a more precise notation. Assume that X is a reference summary of length m, Y is a candidate summary of length n and $LCS(X, Y)$ is the length of a longest common subsequence of X and Y. We can then define precision, recall and F1 score for ROUGE-L as follows:

$$Precision_{RougeL} = \frac{LCS(X, Y)}{n} \tag{3.7}$$

$$Recall_{RougeL} = \frac{LCS(X, Y)}{m} \tag{3.8}$$

$$F1_{RougeL} = \frac{2 * Precision_{RougeL} * Recall_{RougeL}}{Precision_{RougeL} + Recall_{RougeL}} \tag{3.9}$$

While ROUGE-1 and ROUGE-2 measures if a candidate summary is within the same topic as the reference summary (i.e. they both share the same vocabulary), ROUGE-L tries to capture similarity on a sentence level. One weakness of ROUGE is that a literal comparison of words or n-grams does not account for nuanced semantic similarity between two summaries. Simply changing the order of words or replacing words with synonyms may drastically reduce the ROUGE score of two sentences which convey the exact same idea. To address this shortcoming, new metrics such as BERTScore [60] and MoverScore [61] have been created while better methods for evaluation continues to be a topic of research in the NLP community.

Nevertheless, in [31], ROUGE-2 and ROUGE-L have demonstrated good correlation to human judgment scores for single document summaries. We observed this same correlation for podcast summaries and discuss it in section 4.2. Moreover, ROUGE-N and ROUGE-L have been the metrics used to reported past results in Podcast summarization so we have decided to do the same for comparability purposes.

3.5.2 Human evaluation

Our participation in TREC 2021 [28] for the Podcast Summarization task resulted in the evaluation of two of our trained models. Trained NIST[9] evaluators assessed 193 summaries for each model rated the quality of each summary using the EGFB rubric. EGFB stands for "Excellent, Good, Fair, Bad" and it follows this four-scale criteria:

- Excellent: the summary accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. In addition to giving an accurate representation of the content, it contains almost no redundant material which is not needed when deciding whether to listen. It is also coherent, comprehensible, and has no grammatical errors.

- Good: the summary conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains with little redundant material which is not needed when deciding whether to listen. Occasional grammatical or coherence errors are acceptable.

- Fair: the summary conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain redundant material which is not needed when deciding whether to listen and may contain repetitions or broken sentences.

- Bad: the summary does not convey any of the most important content items of the episode or gives the reader an incorrect or incomprehensible sense of what the episode contains. It may contain a large amount of redundant information that is not needed when deciding whether to listen to the episode.

[9] https://www.nist.gov/

Along with the EGFB grade, the following set of yes/no questions were answered for each summary:

- **Q1**: Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast?
- **Q2**: Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)?
- **Q3**: Does the summary include the main topic(s) of the podcast?
- **Q4**: Does the summary tell you anything about the format of the podcast; e.g. whether it's an interview, whether it's a chat between friends, a monologue, etc
- **Q5**: Does the summary give you more context on the title of the podcast?
- **Q6**: Does the summary contain redundant information?
- **Q7**: Is the summary written in good English?
- **Q8**: Are the start and end of the summary good sentence and paragraph start and end points?

The Podcast Summarization track of TREC 2021 only targeted to English podcasts. Human evaluation of podcast summaries is an expensive process as it requires listening to the entire episode before assessing each summary. Due to resources constraints, we were not able to perform human evaluations on the summaries in Portuguese. However, in section 4.2, we do analyze how the ROUGE scores correlated to human judgements using the results from all participants in TREC 2021. We believe that ROUGE score is a good proxy to be used to assess summary quality from a human perspective without the need to actually performing human evaluations.

3.6 Summary

In this chapter, we discuss the datasets and the methodology used to train our summarization models.

In the first part, we discuss the datasets. We provided a detailed description of the Spotify Podcast dataset including what metadata fields were available, the number of episodes, the number of shows, and how many hours of audio were included. We also provided some statistics regarding the number of words per transcript, the number of words per episode description, and what was the distribution if we slice the dataset per genre. Next, we define what boilerplate is, explain how the boilerplate dataset was built, and describe how accurate our boilerplate detection model is.

In the second part, we discuss the methodology. We describe the baseline methods used, how the LongMBART was built and trained, and lastly, what is the experimental setup for the cross-lingual transfer learning experiment. Next, we describe the two evaluation methods used: an automated evaluation using ROUGE scores and a human evaluation performed by NIST during the TREC conference.

Chapter 4

Results and Discussion

4.1 Human evaluation

As explained in section 2.8, the Text Retrieval Conference (TREC) is a series of workshops with the goal of encouraging research in the Information Retrieval field based on large test collections. One of the workshops in TREC 2021 [28] was the Podcast Summarization Track. Every team participating in the workshop was asked to submit summaries for a test set of 1000 episodes defined by the organizing committee. These summaries had to be automatically generated from the data provided by the Spotify Podcasts Dataset [9] or any other dataset. Teams could submit summaries from more than one system, where a system is defined as a program capable of automatically generating a summary using the episode's transcript and audio as inputs. The only constraint is that each system should output only one summary per episode.

We participated in the Podcast Summarization track by submitting the two variants of LongMBART as described in section 3.4.2. The first variant was LongMBART finetuned to episode transcripts and episode descriptions. The second variant was LongMBART finetuned with XL-SUM [22] article-summary pairs and also finetuned to episode transcripts and episode descriptions. For sections 4.1.1 and 4.1.2, the two variants are named as "Unicamp1" and "Unicamp2" respectively. We followed the conference's naming conventions for submitted systems.

Out of 1000 episodes in the test set, 193 were randomly selected to be evaluated by NIST evaluators following the methodology described in section 3.5.2. For comparability, all of the summaries submitted to the workshop were evaluated on the same set of 193 episodes.

4.1.1 Summary Quality

The first part of the human evaluation was an overall quality grade for each summary following EGFB scale: Excellent, Good, Fair, Bad.

In figure 4.1, we can see the absolute count of summaries for each grade. For the sake of simplicity, we will group Bad/Fair in one bucket and Good/Excellent in another. Both our models (Unicamp1 and Unicamp 2) produced 63% more Good/Excellent summaries than the 1st Minute Baseline while producing 13% less Bad/Fair summaries. These results

Systems compared	p-value	Is p-value <0.05?
1st Minute Baseline vs Unicamp1	0.0112	TRUE
1st Minute Baseline vs Unicamp2	0.0122	TRUE
Unicamp1 vs Unicamp2	0.7290	FALSE

Table 4.1: Wilcoxon signed rank test results to measure statistical significance of the difference of the summary quality score. We define the EGFB score numerically as E=3, G=2, F=1, B=0.

suggest that our models are outperforming the 1st Minute Baseline. Ideally, we would also like to compare our results against stronger baselines such as BART[30] or BART-PODCASTS[9] but unfortunately, such models were not evaluated in TREC 2021.

Figure 4.1: Overall quality scores. '1st Minute Baseline' refers to the TREC-provided baseline of the first minute of speech. Unicamp1 is LongMBART finetuned to podcasts data. Unicamp2 is LongMBART finetuned to XL-Sum and podcasts data.

In order to compare our models Unicamp1 and Unicamp2 with the baseline, we define EGFB scores numerically by assigning E=3, G=2, F=1, B=0. With that, we can calculate how statistically significant is the difference between our systems and the 1st Minute Baseline. We also compare one model against the other. For this statistical analysis, we used the Wilcoxon signed rank test - a non-parametric test for paired data - to calculate the p-value. As we can see in see Table 4.1, the difference is statistically significant for both models when compared against the 1st Minute Baseline. Also, the average quality score of our models is higher than the baseline (0.722 for the baseline, 0.969 for Unicamp 1, 0.943 for Unicamp2). When we compared Unicamp1 against Unicamp2, there was no statistically significant difference between the two.

4.1.2 Yes/No Questions

The second part of the human evaluation was a set of yes/no questions which capture a set of boolean attributes that a desirable podcast summary might contain:

- **Q1**: Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast?
- **Q2**: Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)?
- **Q3**: Does the summary include the main topic(s) of the podcast?
- **Q4**: Does the summary tell you anything about the format of the podcast; e.g. whether it's an interview, whether it's a chat between friends, a monologue, etc
- **Q5**: Does the summary give you more context on the title of the podcast?
- **Q6**: Does the summary contain redundant information?
- **Q7**: Is the summary written in good English?
- **Q8**: Are the start and end of the summary good sentence and paragraph start and end points?

In figure 4.2, we see the percentage of summaries where the answer was "yes" for each question. We consider the higher percentages better except for question 6. For question 6, the lower percentages are better.

For **question 1**, our models are worse than the 1st Minute baseline with a statistically significant difference. This result may be an indication that important names are lacking in the episode descriptions used during training. It could also be the case that the model is not trained to perform the task of identifying names of people and adding them to the summary. A recent publication[41] extracts named entities as a separate step in the summarization process and seems like a good option to solve this particular problem.

For **questions 2, 3 and 4,** there is no statistically significant difference between the models and the 1st Minute baseline. This means that either (1) the null hypothesis is true, i.e. there is no effect of using either of our models in comparison to using the 1st Minute Baseline or (2) there is an effect of using our models but this experiment did not have enough evidence to prove it [54]. Question 3 asks about the main topic of the episode which a fundamental feature of a good summary. Moreover, it has the highest Pearson correlation with the summary quality score among all yes/no questions (see Figure 4.3). Given its relevance, we consider that future work demonstrating an improvement over the baseline with a statistically significant different is much needed. Suggestions of future work would include running an evaluation on a larger sample of episodes or experimenting with other models. In regards to question 4, it is possible that most of the episode descriptions do not provide the podcast format in the description therefore not resulting in an improvement over the baseline. This hypothesis remains to be verified.

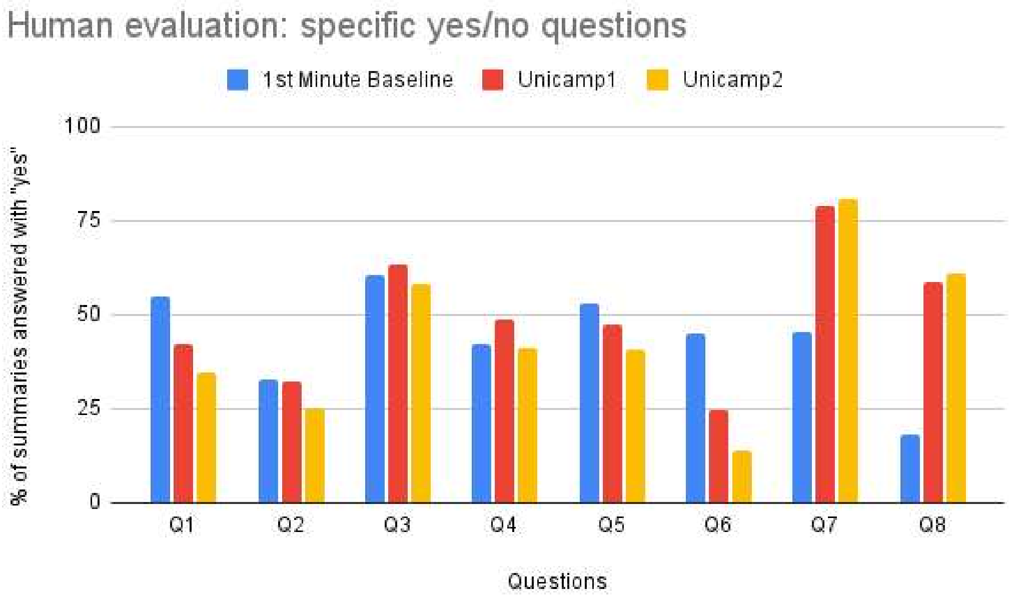


Figure 4.2: Averages per question. '1st Minute Baseline' refers to the TREC-provided baseline of the first minute of speech. Unicamp1 is LongMBART finetuned to podcasts data. Unicamp2 is LongMBART finetuned to XL-Sum and podcasts data.

Question 5 is the second most correlated question with the summary quality score (see Figure 4.3). It asks if the summary provides context on the title of the podcast. It is not surprising that the Pearson correlation coefficient with question 3 is 0.66, given that this question is somewhat similar to asking about the main topic of the podcast. Unfortunately, both our models performed worse than the baseline in terms of percentage of summaries answered with "yes" (see Figure 4.1.2). Although the difference is not statistically significant for Unicamp1, it is for Unicamp2. Again, future work in this particular aspect of summarization is needed.

Our models outperform the 1st Minute Baseline in **questions 6, 7 and 8** with statistically significant difference. Specifically for question 6, Unicamp2 performs better than Unicamp1 in regards to not producing summaries with redundant information (see Figure 4.4). Unlike the other questions, questions 7 and 8 are essentially assessing form and not content and we expected our models to perform better than the baseline in this area. Firstly, because state-of-the-art transformer-based models have been successful in generating fluent text and although the content of the summary may be at times inaccurate, the models will write coherent text. Secondly, the 1st Minute Baseline summary is simply a snippet of the transcript text which contains a certain amount of word errors. Thirdly, the end of the 1st Minute Baseline summary will be an arbitrary point with the 1-minute mark most likely resulting in the abrupt cut mid-sentence.

Question	**p-value**	**Is p-value <0.05?**
Q1	0.002256344241	TRUE
Q2	0.8927384009	FALSE
Q3	0.5287333251	FALSE
Q4	0.08508907282	FALSE
Q5	0.1590109785	FALSE
Q6	2.67E-06	TRUE
Q7	6.49E-11	TRUE
Q8	4.22E-16	TRUE

Table 4.2: Wilcoxon signed rank test results. For each yes/no question, we compare the 1st Minute Baseline against the Unicamp1 system.

Question	**p-value**	**Is p-value <0.05?**
Q1	3.68E-06	TRUE
Q2	0.06601967502	FALSE
Q3	0.5351434524	FALSE
Q4	0.902764825	FALSE
Q5	0.003798437441	TRUE
Q6	1.60E-10	TRUE
Q7	2.32E-12	TRUE
Q8	1.66E-17	TRUE

Table 4.3: Wilcoxon signed rank test results. For each yes/no question, we compare the 1st Minute Baseline against the Unicamp2 system.

Question	**p-value**	**Is p-value <0.05?**
Q1	0.03480847881	TRUE
Q2	0.03737298834	TRUE
Q3	0.1967056025	FALSE
Q4	0.06601967502	FALSE
Q5	0.08968602177	FALSE
Q6	0.003275897483	TRUE
Q7	0.592980098	FALSE
Q8	0.5150822787	FALSE

Table 4.4: Wilcoxon signed rank test results. For each yes/no question, we compare the model Unicamp1 against the model Unicamp2.

	egfb	audio	q1	q2	q3	q4	q5	q6	q7	q8
egfb	1.000000	0.224525	0.315758	0.308904	0.621226	0.339426	0.522264	0.293035	0.382040	0.379629
audio	0.224525	1.000000	0.049975	0.085139	0.140858	0.202880	0.110943	0.206044	0.292978	0.262828
q1	0.315758	0.049975	1.000000	0.598698	0.287084	0.301891	0.303172	0.087212	0.122881	0.037317
q2	0.308904	0.085139	0.598698	1.000000	0.231823	0.276614	0.241440	0.086879	0.135384	0.106010
q3	0.621226	0.140858	0.287084	0.231823	1.000000	0.308296	0.660961	0.162771	0.246238	0.230109
q4	0.339426	0.202880	0.301891	0.276614	0.308296	1.000000	0.307444	0.248549	0.287609	0.145186
q5	0.522264	0.110943	0.303172	0.241440	0.660961	0.307444	1.000000	0.140468	0.140593	0.139976
q6	0.293035	0.206044	0.087212	0.086879	0.162771	0.248549	0.140468	1.000000	0.185775	0.111610
q7	0.382040	0.292978	0.122881	0.135384	0.246238	0.287609	0.140593	0.185775	1.000000	0.616005
q8	0.379629	0.262828	0.037317	0.106010	0.230109	0.145186	0.139976	0.111610	0.616005	1.000000

Figure 4.3: Pearson correlation between the summary quality score EGFB and the yes/no questions across all submitted systems of TREC 2021.
Source: Extracted from [28].

4.2 Correlation between human evaluation and intrinsic metrics

Human evaluation of summarization is an expensive process [31]. For this reason, several metrics[31][4][60][18] have been proposed for automatically assessing the quality of a summary without human intervention. Ultimately, a good evaluation metric will correlate with human judgements as demonstrated with ROUGE[31] when applied against the datasets DUC 2001, 2002 and 2003 [42].

Although we had evaluators available during the TREC workshop, we needed a systematic way to evaluate our models in a frequent basis. For that, we measured the correlation between human judgements and the following intrinsic metrics:

- ROUGE-1: measures unigram overlap between reference summary and candidate summary [31]
- ROUGE-2: measures bigram overlap between reference summary and candidate summary [31]
- ROUGE-L: measures longest matching sequence of words using longest common sequence (LCS) [31]
- Meteor: measures unigram matches between the reference and candidate summaries. Unigrams can be matched based on their surface forms, stemmed forms, and meanings [4].
- BertScore: computes a similarity score for each token in the test summary with each token in the reference summary. Token similarity uses contextual embeddings as opposed to exact matching.

- description_summary_similarity_all_mpnet_base_v2: cosine similarity between SBERT[45] embeddings of the episode description and candidate summary.
- transcript_summary_similarity_all_mpnet_base_v2: cosine similarity between SBERT[45] embeddings of episode transcript and candidate summary (unsupervised metric).

For ROUGE and BertScore, we have analyzed precision, recall and F1-score.

For this analysis, we used summaries from all participating teams in TREC 2021 and we defined a quality score based on this numerical assignment of the EGFB scale: E=3, G=2, F=1, B=0. Given that we wanted to compare the performance of each system submitted and not necessarily individual summaries, we first averaged the summary quality score and each of the intrinsic metrics per system. Then, we calculated Pearson's correlation coefficient between the average quality score and each of the averaged metrics. Another reason for measuring correlation on a system level - as opposed to summary level - is the fact that ROUGE is very unstable as it can easily equal to zero when none of the n-grams match between the candidate and reference summaries. A summary can still surface the important aspects of an episode and be deemed as a good summary even when its ROUGE score is zero. Therefore, we did not expect ROUGE scores to correlate to human evaluation scores on a summary by summary basis. However, we did expect the average ROUGE score to correlate with the human impression of how good a summarization system is. The correlation between the average quality score and other averaged intrinsic metrics is in Figure 4.4.

We see high correlation for the F1-scores of ROUGE-1, ROUGE-2 and ROUGE-L as well as for description_summary_similarity_all_mpnet_base_v2. Overall, precision and recall demonstrated weaker correlation or even negative in some cases. BertScore and the unsupervised metric transcript_summary_similarity_all_mpnet_base_v2 do not seem to correlate with the average quality score either.

Nevertheless, the fact that Pearson's correlation coefficients are high for ROUGE F1-scores indicate that ROUGE is a good proxy to human judgments of summaries for podcast episodes and a good metric to compare different systems for the task of podcast summarization. Moreover, even though we are not able to perform human evaluation on summaries for episodes in Portuguese, this result gives us confidence that an intrinsic evaluation with ROUGE will lead to good summaries from a human's perspective.

4.3 Intrinsic evaluation using ROUGE scores

In this section, we evaluate our finetuned models and baselines using ROUGE scores [31]. Compared to human evaluations, intrinsic evaluations such as this one are much faster and cheaper to run which in turn result in more iterations during the research phase. We present our results using ROUGE-1, ROUGE-2 and ROUGE-L F1 scores. Table 4.5 presents the ROUGE scores when evaluating only on a test set of 4511 English podcasts. Table 4.6 presents the ROUGE scores when evaluating only on a test set of 5073 Portuguese podcasts. Lastly, Table 4.7 presents the ROUGE scores for the combination of both test sets in English and Portuguese.

	quality
quality	1.000000
egfb	0.999822
meteor	0.464324
description_summary_similarity_all_mpnet_base_v2	0.917660
transcript_summary_similarity_all_mpnet_base_v2	0.440032
bertscore-p	0.595510
bertscore-r	0.375171
bertscore-f1	0.516715
rouge-1:r	-0.170777
rouge-1:p	0.839715
rouge-1:f	0.929066
rouge-2:r	0.517585
rouge-2:p	0.854391
rouge-2:f	0.938316
rouge-l:r	-0.127953
rouge-l:p	0.844418
rouge-l:f	0.933793

Figure 4.4: Pearson correlation between the summary quality score EGFB and intrinsic metrics across all submitted systems of TREC 2021.

In order to make these results more intelligible, we will split the discussions per subsections. In subsection 4.3.1, we will contrast the MBART model and its Longformer counterpart. In subsection 4.3.2, we will discuss the LongMBART finetuned only on podcasts data and the LongMBART finetuned to XL-SUM data and podcasts data. In subsection 4.3.3, we will discuss the results from a cross-lingual transfer learning perspective. Finally, in subsection 4.3.4, we will talk about layout bias and why the first minute baseline presents competitive results.

4.3.1 MBART vs LongMBART

We converted the MBART model into LongMBART, a Longformer version which increased the input text size limit from 512 tokens to 4096 tokens. Our hypothesis was that passing more information (i.e. more transcript text) to the model would lead to higher ROUGE scores. A data analysis on a sample of episodes has demonstrated substantial information loss when reducing the input size from 4096 to 512 tokens. See results in Table 4.8. On average, we lose more than 80% of words per episode transcript when the input size is limited to 512 tokens. On the other hand, this number drops to 33.9% (Portuguese) and 35.6% (English) when the input size is limited to 4096 tokens.

Contrary to our beliefs, the LongMBART model did not lead to a higher ROUGE scores when compared to the MBART model. MBART has performed only slightly better than LongMBART for all ROUGE F1-scores in English and Portuguese podcasts.

When evaluating episodes in English (Table 4.5), the MBART finetuned monolingually to English (MBART + finetuned EN podcasts) provided the best scores of all experiments performed. Its Longformer counterpart (LongMBART + finetuned EN podcasts) performed practically on par but nevertheless with lower ROUGE scores. If we

	R1-F	R2-F	RL-F
First Minute baseline	0.1723	0.0303	0.1545
TextRank Top 5 sentences	0.1401	0.0161	0.1183
TextRank Top 2 sentences	0.1407	0.0145	0.1144
XLSUM vanilla	0.1174	0.0156	0.1036
MBART vanilla	0.1579	0.0272	0.1400
MBART + finetuned PT podcasts	0.0407	0.0046	0.0385
MBART + finetuned EN podcasts	**0.1862**	**0.0563**	**0.1663**
MBART + finetuned PT/EN podcasts	**0.1859**	0.0499	**0.1657**
LongMBART vanilla	0.1620	0.0280	0.1440
LongMBART + finetuned PT podcasts	0.0341	0.0043	0.0327
LongMBART + finetuned EN podcasts	0.1845	0.0521	0.1633
LongMBART + finetuned PT/EN podcasts	0.1812	0.0482	0.1620
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	0.1844	**0.0553**	0.1650

Table 4.5: ROUGE-1, ROUGE-2 and ROUGE-L F1 scores for internal test set of 4511 English podcast episodes. In bold, the top two highest ROUGE scores.

	R1-F	R2-F	RL-F
First Minute baseline	0.1674	0.0327	0.1397
TextRank Top 5 sentences	0.1169	0.0143	0.0959
TextRank Top 2 sentences	0.1335	0.0139	0.1058
XLSUM vanilla	0.1120	0.0159	0.0951
MBART vanilla	0.1586	0.0277	0.1342
MBART + finetuned PT podcasts	**0.1886**	**0.0516**	**0.1634**
MBART + finetuned EN podcasts	0.0393	0.0067	0.0369
MBART + finetuned PT/EN podcasts	**0.1835**	**0.0501**	**0.1598**
LongMBART vanilla	0.1136	0.0119	0.1021
LongMBART + finetuned PT podcasts	0.1826	**0.0501**	0.1598
LongMBART + finetuned EN podcasts	0.0280	0.0046	0.0266
LongMBART + finetuned PT/EN podcasts	0.1761	0.0491	0.1536
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	0.1764	0.0481	0.1535

Table 4.6: ROUGE-1, ROUGE-2 and ROUGE-L F1 scores for test set of 5073 Portuguese podcast episodes. In bold, the top two highest ROUGE scores.

	R1-F	R2-F	RL-F
First Minute baseline	0.1697	0.0316	0.1466
TextRank Top 5 sentences	0.1278	0.0152	0.1064
TextRank Top 2 sentences	0.1369	0.0142	0.1099
XLSUM vanilla	0.1146	0.0157	0.0991
MBART vanilla	0.1583	0.0275	0.1369
MBART + finetuned PT podcasts	0.1191	0.0295	0.1047
MBART + finetuned EN podcasts	0.1084	0.0300	0.0978
MBART + finetuned PT/EN podcasts	**0.1846**	**0.0500**	**0.1625**
LongMBART vanilla	0.1364	0.0195	0.1218
LongMBART + finetuned PT podcasts	0.1128	0.0286	0.1001
LongMBART + finetuned EN podcasts	0.1018	0.0270	0.0911
LongMBART + finetuned PT/EN podcasts	0.1785	0.0487	0.1576
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	**0.1802**	**0.0515**	**0.1589**

Table 4.7: ROUGE scores for internal test set of 5073 Portuguese podcast episodes combined with internal set of 4511 English podcast episodes. In bold, the top two highest ROUGE scores.

	Input 4096 tokens	Input 512 tokens
English	3248 words lost 35.6% of transcript words lost	5493 transcript words lost 82.9% of transcript words lost
Portuguese	3633 words lost 33.9% of transcript words lost	5954 transcript words lost 80.6% of transcript words lost

Table 4.8: We compare how much information is lost when input size is limited to 512 and 4096 tokens. Average of number of words lost per episode and average of percentage of words lost per episode. We present the results from a sample of 9133 episodes in Portuguese and 10251 episodes in English.

analyze the two models finetuned bilingually (MBART + finetuned PT/EN podcasts and LongMBART + finetuned PT/EN podcasts), we see very similar results where MBART resulted in slightly better ROUGE scores than LongMBART.

We can do the same comparison for the episodes in Portuguese and arrive at the same conclusions. MBART has slightly outperformed LongMBART in both the monolingually and bilingually finetuned versions. In Table 4.6, we can compare **MBART + finetuned PT podcasts** against **LongMBART + finetuned PT podcasts** and also compare **MBART + finetuned PT/EN podcasts** against **LongMBART + finetuned PT/EN podcasts**.

4.3.2 Adding finetuning on XL-SUM data

As explained in section 3.4.2, we experimented with two different finetuning strategies for LongMBART. We compare the models **LongMBART + finetuned PT/EN podcasts** and **LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts** in tables 4.5, 4.6 and 4.7. Our initial hypothesis was that finetuning the model with two summarization datasets (XL-SUM and podcasts) would lead to better summaries.

The double-finetuned model (LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts) performed slightly better than the single-finetuned one (LongMBART + finetuned PT/EN podcasts) in practically all 3 variants of ROUGE. For English and Portuguese combined, the double-finetuned model provided the 2nd best ROUGE scores out of all models experimented. Although this shows that the double-finetuned model is a good summarization model for podcasts, we did expect to see a larger difference between the single and double-finetuned models.

We conjecture that the fact that XL-SUM data is so different from podcasts data leads to no additional gain in ROUGE scores. In other words, summarizing BBC news articles (with XL-SUM data) is distinct enough of a task when compared to summarizing podcasts that learning both is not complementary. The fact that the vanilla XL-SUM model - trained only in XL-SUM data - performed poorly also seems to support this conjecture. ROUGE scores for vanilla XL-SUM model were worse than the 1st minute baseline.

Additionally, we conclude that having a preliminary round of finetuning in XL-SUM data is not worth the significant computing costs considering the minimal increase in ROUGE scores. This preliminary finetuning round took 3 weeks in total.

4.3.3 Cross lingual transfer learning

As explained in subsection 3.4.3, we experimented with all possible combinations of two models (MBART and LongMBART) and three finetuning variants (only English, only Portuguese and both languages intermingled). In this section, we analyze the results from the perspective of cross lingual transfer learning.

As expected, the models that were finetuned on the target language (i.e. language of the test set) produced the best results and outperformed the First Minute baseline. For English podcasts, these results can be seen in Table 4.5 for MBART and LongMBART finetuned to EN podcasts or PT/EN podcasts. For Portuguese podcasts, these results can be seen in Table 4.6 for MBART and LongMBART finetuned to PT podcasts or PT/EN podcasts. It is also worth noting that the difference between finetuning monolingually and bilingually is marginal which leads us to conclude that learning to summarize in an additional language does not come at the cost of worse performance in a first language.

The results for the models finetuned to a language other than the target language performed worse than the First Minute baseline and also worse than the vanilla version. We are looking at the models finetuned to PT podcasts in Table 4.5 and the models finetuned to EN podcasts in Table 4.6. These models tended to write summaries on the source language regardless of the input text's language or of the language parameters specified during inference time. In tables 4.10 and 4.9, we can see two examples of summaries produced with these models. We see that the language of the episode description - used as summary reference - is different from the language of the summary which explains why ROUGE scores were so low for these models. If we read the summaries, they seem to be on topic even if written on the wrong language. Another interesting fact is that some of the summaries mixed English and Portuguese in a sensible way (see table 4.9 where the summary says "reprodutive-age mulheres").

Episode description	Up to 30% to 50% of infertile women have endometriosis, based on laparoscopic diagnosis. Endometriosis affects fertility through a variety of mechanisms. In this session, we will review the ASRM position statement on endometriosis related infertility: work-up and patient management. — This episode is sponsored by · Anchor: The easiest way to make a podcast. https://anchor.fm/app
Summary generated	A endometriose é uma patologia do sistema reprodutor feminino, que pode afetar até 10% de todas as reprodutive-age mulheres. A endometriose é uma patologia do sistema reprodutor feminino. A endometriose é uma patologia do sistema reprodutor feminino, que pode afetar até 10% de todas as reprodutive-age mulheres.

Table 4.9: Example of summary generated by MBART finetuned to Portuguese podcasts only. The episode's language was English.

Episode description	Isso mesmo, 5 dias antes do lançamento do iPhone 13, vazaram imagens do iPhone 14 que vai ser lançado em 2022. E meu amigo, está animal!! Confere as imagens: https://www.instagram.com/applelogias/ — Send in a voice message: https://anchor.fm/applelogias/message Support this podcast: https://anchor.fm/applelogias/support
Summary generated	This week on the apologies podcast, we talk about Apple's decision to eliminate the iPhone from the market.

Table 4.10: Example of summary generated by MBART finetuned to English podcasts only. The episode's language was Portuguese.

In subsection 3.4.3, we stated a hypothesis that cross-lingual transfer learning was possible in the podcast summarization task. However, this hypothesis could not be confirmed with the given results. When testing on English podcasts, MBART + finetuned PT podcasts performed much worse than MBART vanilla. We saw very similar results when testing on Portuguese podcasts where MBART + finetuned EN podcasts performed much worse than MBART vanilla. These statements hold true for the LongMBART model as well. Although our initial assumption was that vanilla MBART did not know how to summarize at all, it just so happens that being a machine translation model, it is basically copying the initial portion of the transcript. This simple operation is a strong baseline as we can see with the First Minute baseline. We believe that more work still remains to be done to study CLTL on podcast summarization. A better comparison would be to replace MBART with a vanilla multilingual summarization model like XL-SUM and run the same experiments.

4.3.4 Layout bias in podcast transcripts

In [29], the authors discuss the concept of "layout bias" in news articles. News articles usually follow a writing format known in journalism as the "Inverted Pyramid" [56]. Ac-

Episode description	Top tips on how to stay healthy and be more environmentally friendly in your day-to-day life.Hosted by Daniel & Georgia
Summary generated	Hi and Welcome to our Healthcare and sustainability segment. I'm Daniel a second year Pharmacy student and I'm Georgia a second-year environmental science student and I'll segment is about how you can look after yourself and the environment.
Transcript	Hi and Welcome to our Healthcare and sustainability segment. I'm Daniel a second year Pharmacy student and I'm Georgia a second-year environmental science student and I'll segment is about how you can look after yourself and the environment. So today I'll be talking about different ways to maintain your mental health as it was Mental Health Awareness Week only a few weeks ago at the start of October. So firstly I wanted to give the definition of mental health according to the world help...

Table 4.11: Example of summary generated by vanilla MBART. We can see that the summary is a copy of the first few sentences of the transcript. This is the same as the First Minute baseline.

cording to this format, initial paragraphs contain the most newsworthy information, which is followed by details and background information. A human study in [29] analyzed 100 randomly sampled articled and concluded that nearly 60% of the important information was present in the first third of the news article. From a summarization perspective, that means that sentences in the beginning of an article are more relevant than sentences in the middle or towards the end.

We empirically observe this same layout bias in podcast transcripts. In Table 4.12, we can see an example of the naive First Minute baseline and how it can serve as a decent summary. It provides a number of important information for a good summary: names of the host, name of the guest, name of the podcast show and topic of the episode. This type of introduction is common specially for episodes with interviews.

Although we have not conducted a human study to measure layout bias in podcasts, we can see in Tables 4.6 4.5 4.7 that the First Minute baseline is very competitive and outperforms other summarization solutions such as TextRank and XLSUM vanilla. The weakness of the First Minute baseline summary lies in the poorly written text noted in the human evaluation in subsection 3.5.2.

We believe that this layout bias could also account for the fact that the Longformer-based model did not outperform the MBART model. In other words, it is possible that the initial 512 tokens of a transcript already cover the most important information needed to produce a summary and that extending the input size to 4096 tokens does not contribute much in terms of relevant information. This is only a conjecture at this point and remains to be studied further.

One unexpected outcome was the fact that the XL-SUM vanilla model performed worse than MBART vanilla. Given that XL-SUM is a model finetuned on the summarization

Episode description	In this **episode** you will **find out** about **Fashion** and **styling** - places to shop for **tweens** and top tips for everyday styling. all by yours truly **Amelia** — Send in a voice message: https://anchor.fm/tweentalk/message
First minute transcript	Hi there, welcome to **Tween** talk. Today's **episode** is so great. I got to **find out** more about **fashion** and **styling** from Fashion Stylist Siobhan Baxter and you will too. I hope you enjoy this episode. Hi there. My name is **Amelia** and I'm host between doc here will be talking all things swing Fashion Beauty food and lots more. Hi there, welcome back to twinkle today. We've got a very special guest here with me **Fashion** Stylist schiavone Baxter who just happens to be my mum. Hi, Mom. Hi, Millie, very happy to be here on your podcast. Okay. So today I'm gonna ask you a few questions. Is that all right? That would be amazing. All right, let's get into it. So what is **styling** and where do you do it? So What I do every day is different. So I style

Table 4.12: Example of summary generated by the First Minute baseline. This summary is simply the transcription of the first minute of the episode.

task, our expectation was that it would lead to higher ROUGE scores when compared to MBART which is a neural machine translation model. As observed in Table 4.11, we noticed that MBART is mostly copying the beginning of a transcript and therefore producing a summary similar to the First Minute baseline. MBART in this case is simply acting as a translation model where the source and target language are the same. Thus, we surmise that the competitive performance of the vanilla MBART model is again an artifact of the layout bias previously discussed, and would not necessarily generalize to other datasets.

4.3.5 Common errors in summaries

In this subsection, we analyzed a random sample of 100 summaries in English and another random sample of 100 summaries in Portuguese. This analysis focused mostly syntactic errors as opposed to semantic errors. That means we did not listen to the entire episode before inspecting the episode description and neither did we do a thorough check of misinformation in the descriptions. Nevertheless, we were able to spot one obvious case of misinformation in the summaries in Portuguese.

The most common errors found were: repetition, missing punctuation, lack of quotes and abrupt ending. In Table 4.13, you can find how many occurrences of each error was found in the two samples of summaries. We discuss each error in the following sub-sections.

Repetition

Repetition manifests in different levels of granularity. For example, we may find repeated words:

Error	**Occurrences in sample of summaries in EN**	**Occurrences in sample of summaries in PT**
Repetition	15	7
Abrupt ending	5	3
Incorrect punctuation	3	4
Lack of quotes for named entity	2	3

Table 4.13: Number of errors found in each random sample of 100 summaries.

> "I've been **lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy lazy**"

> "Nesse episódio falamos um pouco sobre **moda**, **moda** e o que não gostamos de cair numa ilha."

We may also find cases of repeated sentences or n-grams:

> "In this week's episode, the Brothers and Sisters team discuss the final week of the year, **what it means to be generous**, and **what it means to be generous**."

> "No episódio de hoje, falamos sobre o novo álbum da cantora Britney Spears **e o que está fazendo com o The Glee Project**. Também falamos sobre o movimento Black Lives Matter **e o que está fazendo com o The Glee Project**."

Repetition was the most common error found and it is most likely related to the use of ROUGE as the target metric used during finetuning. When the summarization model needs to continue generating more text - such as after the conjunction "and" as seen in the examples above - it will try to maximize the ROUGE. In some cases, the model may not have any additional information to add to the generated summary so it simply repeats n-grams likely to be present in the reference summary.

Missing or incorrect punctuation

We have found cases of missing punctuation (commas, periods, question marks and colons) or incorrect use of them. In most cases, we recognized many elements of spoken language as opposed to written language. In other words, these summaries still resembled a transcript and with that, its common problem of missing punctuation.

To the reader of these summaries, missing punctuations (specially missing periods) add a major obstacle for interpretation. The cognitive burden includes finding the different ways to separate sentences or ideas and also decoding the text into a possible speech with intonations.

Here is one example of a summary missing a question mark and a corrected version in English:

> "This week's topic is a slightly old fashioned word, maybe a slightly unpopular word discipline. **I've got good news and bad news horses don't need discipline**. Horses need routine."
>
> [Corrected] "This week's topic is a slightly old fashioned word, maybe a slightly unpopular word discipline. **I've got good news and bad news. Horses don't need discipline.** Horses need routine."

Here is another example in Portuguese:

> "**Escolhas tem suas consequências e renúncias, pois é que tudo que acontece com a gente nem sempre está sob nosso controle. Mas a forma que nós vamos reagir a esses acontecimentos sim, tudo isso são escolhas.** As escolhas vão terminar o seu sucesso, o seu destino e quais são as escolhas que temos o poder de fazer nesse tipo de Cassidy. Hoje eu te convido a acompanhar um trecho de uma conversa que fiz para um grupo de Telegram. Esse trecho foi editado e vem aqui com as escolhas essenciais que nós devemos fazer para ter sucesso e felicidade. Espero que você goste."
>
> [Corrected] "**Escolhas tem suas consequências e renúncias. Pois é. Tudo que acontece com a gente nem sempre está sob nosso controle. Mas a forma que nós vamos reagir a esses acontecimentos sim. Tudo isso são escolhas.** As escolhas vão (de)terminar o seu sucesso, o seu destino e quais são as escolhas que temos o poder de fazer nesse tipo de Cassidy. Hoje eu te convido a acompanhar um trecho de uma conversa que fiz para um grupo de Telegram. Esse trecho foi editado e vem aqui com as escolhas essenciais que nós devemos fazer para ter sucesso e felicidade. Espero que você goste."

Lack of quotes for named entities

Many of the summaries mention named entities such as the name of a book or of another podcast show. Sometimes, these mentions are not quoted which makes it hard to interpret the text. Here are some examples:

> "In this episode of the Gratitude Podcast, Georgian Benta sits down with thought leader and **author of The 10 Worlds**, A.D. Freud. A.D. is a thought leader in clinical social and consumer psychology. He recently published his **book The 10 Worlds: The New Psychology of Happiness**, which he and his co-authors spent over two decades researching the answer to this question: What is Happiness after transitioning to a career in marketing? A.D. has held senior positions at companies including Google, McKinsey, Weekly Go Go Go, and currently Red Box. He's been featured in the Economist, Forbes,"

> "Muitos anos depois, diante do pelotão de fuzilamento, o Coronel Aureliano buendía havia de recordar aquela tarde remota em que seu pai o levou para conhecer o gelo. É assim que o autor colombiano Gabriel García Márquez começa a sua obra mais conhecida, que é **100 anos de solidão**. Meu nome é Matheus e esse é o Fictionário. Um podcast sobre livros."

Named entity detection is a problem long studied in NLP and using some of its methods to correct this problem may be an alternative. Another alternative to explore would be to use a knowledge graph where these entities (books, podcast shows, movies) have already been detected.

Like incorrect punctuation, the lack of quotes is another example of error propagated from noisy transcripts as the input documents.

Abrupt ending

We have found some summaries to end abruptly, meaning they ended mid-sentence. Here is an example in English:

> "In this week's Afterbuzz After Show, hosts Carla, Amanda, and Shaun T break down the latest episode of the Real Housewives of Orange County season 14, "Dance, Like, No One's Watching." ABOUT REAL HOUSEWIVES OF ORANGE CITY: Real Housewives of Orange County is an upcoming American reality competition television series based on the characters by Archie Comics. It has been ordered to series at The CW and is scheduled to air during the 2016–17 television season, coinciding with the Archie character's 75th anniversary. **The series focuses on the Archie character's search for "America's next"**

Most of the cases of abrupt ending found were in long summaries (100-122 words) such as the one above and this can be attributed to the fact that when tokening the episode descriptions for the training set, we used the parameter max_length=150. So the longest reference summaries seen by the model during training were 150 tokens long.

Hallucinations and misinformation

Although we did not analyze the summaries on a semantic level, it is important to at least mention here that hallucinations are still a major problem to be solved in the summarization area. Hallucinations are facts present in the summary but that cannot be backed up by the input document. This is different from misinformation which means a false statement regardless of what the input document says.

During our analysis, we found one case which caught our attention:

> No episódio de hoje, conversamos com Mario Avelar, trader de opções e presidente do Banco Central.

The name of the guest was Mario Avellar and although he was a trader, he was never the president of the Central Bank of Brazil. Analyzing the transcript, we can find this bit which explains the mistaken summary. In this case, the transcript is very noisy so we are also providing a corrected version as well:

> "...para você que quer entender melhor sobre o assunto tão comentado nas redes sociais conversamos mais à frente com o **mar Avelar é de trader de opções destaque do noticiário presidente do banco central Roberto.** Leite que embora não seja uma meta explícita da..."
>
> [Corrected]"...para você que quer entender melhor sobre o assunto tão comentado nas redes sociais, conversamos mais à frente com **Mario Avellar, head trader de opções. Destaque do noticiário: o presidente do Banco Central, Roberto** Campos Neto, disse ontem a noite que embora não seja uma meta explícita da..."

Comparing the original transcript and the corrected version, we can notice that the noise introduced in the original transcript is significant and practically impedes someone from reaching the correct interpretation. The lack of a period before "Destaque do noticiário" indicating a new separate idea, misled the model to attribute "presidente do banco central" to the person Mario Avellar. So although, this is not a case of hallucination, it is still a case of misinformation.

4.4 Summary

In this chapter, we analyze the results of our experiments through the lens of two evaluation methods: a human evaluation and an automated evaluation using the ROUGE metric.

Firstly, we analyzed the results of the human evaluation. The summary quality of our models outperformed the first minute baseline with a statistically significant difference. However, when it came to the Yes/No questions, our models performed better or worse than the baseline depending on the question. Our models performed well in the questions pertaining good written English and less redundant information.

Next, we found a strong correlation between ROUGE scores and the summary quality score used during the human evaluation. This supports our decision to use ROUGE as the metric to evaluate podcast summaries.

In the subsequent sections, we evaluated our initial hypothesis against the ROUGE scores of our experiments. Although we assumed a gain in ROUGE score with different methods, these gains were not confirmed. We reflected on these results by a set of conjectures to be investigated in future work.

Finally, we provide an interesting analysis of the most common errors found in podcast summaries. We present examples of summaries for each one of the common errors: repetition, incorrect punctuation, lack of quotes for named entities and abrupt ending. We also analyze why these errors occur considering the input documents and the summarization model used.

Chapter 5

Final considerations

In this chapter, we review our findings and major topics discussed in this Master thesis. In addition, we propose future work in the domain of multilingual podcast summarization.

5.1 Contributions

- **Podcast Dataset in Portuguese**: We have published to the research community a dataset consisting of 123,054 podcast episodes in Portuguese from 16,131 shows and encompassing more than 76,000 hours of speech audio. The addition of a second language to the existing Spotify Podcast Dataset is a first step towards the exploration of Podcast Summarization from a multilingual perspective. We believe this dataset will be valuable to researchers NLP community studying spoken language based on audio or text data.

- **Multilingual Podcast Summarization**: To the best of our knowledge, this is the first work on podcast summarization using multilingual transformer-based models. We have evaluated our results with humans and intrinsic metrics. We have compared these results against a set of different baselines. We concluded that it is possible to train a single summarization model in two languages without loss in performance when compared to a model trained monolingually.

- **Cross lingual transfer learning**: We have studied cross lingual transfer learning when applying multilingual transformer-based models to summarize podcast episodes. We experimented with two variants of finetuning (only English podcasts and only Portuguese podcasts) and two models (MBART and LongMBART). It was not possible to confirm cross lingual transfer learning in any of finetuned models. We believe that the layout bias in podcasts is a confounding factor which prevented a fair comparison of a vanilla model against one of the finetuned models.

- **LongMBART experiments**: We have converted the MBART model into its Longformer version called LongMBART. We have studied the impact of using Longformer's attention mechanism when summarizing podcast episodes multilingually. We have contrasted MBART with LongMBART and have concluded that LongM-

BART with longer input text size does not provide significant improvements over the full-attention model MBART.

- **Finetuning with XL-SUM and podcasts**: We have finetuned LongMBART with XL-SUM data followed by podcasts data. We did not observe any gain in ROUGE scores by adding a preliminary round of finetuning on XL-SUM data. We conjecture that the tasks of summarizing news articles and summarizing podcasts are so distinct that the two rounds of finetuning do not complement each other.

- **Correlation between human and intrinsic evaluation**: We have concluded that there is a high correlation between the summary quality perceived by humans and ROUGE F1-scores when averaging these metrics on a per-system level. This correlation is an indication that ROUGE scores can be used to compare different summarization systems while being faithful to human judgement of summaries.

5.2 Future work

We believe that multilingual podcast summarization should not be limited to English and Portuguese. MBART has been finetuned to 50 different languages so future work could experiment expanding this summarization to any of the other languages in this list. As we add more languages to this single summarization, does ROUGE scores start to decrease? It is also important to include low-resource languages and study cross lingual transfer learning for those cases. Can these low-resource languages benefit from other high-resource languages in a podcast summarization model?

The XL-SUM model open another avenue of interesting experiments. We believe that finetuning the XL-SUM model on podcasts data would be a straightforward and important experiment to run. Although we have not seen gains in using the XL-SUM dataset to finetune our models, would the results be different using the XL-SUM model directly? There is an important difference between the two: the XL-SUM model has been trained on 44 and in our experiments, we were only using the English and Portuguese sub-sets of the XL-SUM dataset.

We have also discussed some errors on the syntactic and semantic level. On the syntactic level, we noted problems such as incorrect punctuation, abrupt ending and repetition. Using language models and existing technology for automatic grammar checking, we should be able to at least detect errors and apply a post-processing step to either delete or fix problematic sentences. On the semantic level, hallucinations is a problem currently studied in the NLP area and it would be important to investigate if solutions tested on well structure text would also work on podcast transcripts.

Lastly, a much more challenging avenue of exploration would be generating summaries directly from audio data. Facebook AI has been investing a lot on the field of Textless NLP and promising results have been demonstrated on the task of translation where the input is in the audio format. Given that translation and summarization are both sequence-to-sequence problems, using audio for generating text summaries would be a very interesting line of research to explore. The advantage of using audio directly is the fact that we do not introduce noise early in the process with transcripts.

Summarization is one of the most challenging problems in NLP and here we only provide a few options of future work. I am sure that our journey on exploring podcast summarization is just beginning.

Bibliography

[1] Podcast statistics & industry trends: Avg. listener numbers, gear, formats & more. https://www.thepodcasthost.com/listening/podcast-industry-stats/#How_Many_Podcasts_Are_There_Right_Now. Accessed: 2022-07-06.

[2] Jay Allamar. The illustrated transformer. https://jalammar.github.io/illustrated-transformer/. Accessed: 2022-05-21.

[3] Deepa Anand and Rupali Wagh. Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University - Computer and Information Sciences*, 34(5):2141–2150, 2022.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[6] Pratik K. Biswas and Aleksandr Iakubovich. Extractive summarization of call transcripts. *CoRR*, abs/2103.10599, 2021.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[8] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[10] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents, 2018.

[11] Arman Cohan and Nazli Goharian. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, 19:1–17, 09 2018.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

[14] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, 3rd Edition.* MIT Press, 2009.

[15] Selvani Deepthi, Vijayasanthi Puli, G. Raju, and Rajesh Bandaru. *An Automatic Legal Document Summarization and Search Using Hybrid System*, pages 229–236. 01 2013.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.

[18] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021.

[19] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[20] Apar Garg, Saiteja Adusumilli, Shanmukha Yenneti, Tapas Badal, Deepak Garg, Vivek Pandey, Abhishek Nigam, Yashu Gupta, Gyan Mittal, and Rahul Agarwal. *NEWS Article Summarization with Pretrained Transformer*, pages 203–211. 02 2021.

[21] Carlos-Emiliano González-Gallardo, Romain Deveaud, Eric SanJuan, and Juan-Manuel Torres-Moreno. Audio summarization with audio features and probability distribution divergence, 2020.

[22] Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages, 2021.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[25] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. Trec 2020 podcasts track overview, 2021.

[26] Hannes Karlbom. Abstractive summarization of podcast transcriptions. Master's thesis, Uppsala University, Artificial Intelligence, 2021.

[27] Hannes Karlbom and A. Clifton. Abstract podcast summarization using bart with longformer attention. In *TREC*, 2020.

[28] Jussi Karlgren, Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Sravana Reddy, Edgar Tanaka, and Md Iftekhar Tanveer. Trec 2021 podcasts track overview, 2022.

[29] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics.

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[31] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[32] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[34] Elena Lloret, María Romá-Ferri, and Manuel Sanz. Compendium: A text summarization system for generating abstracts of research papers. pages 3–14, 06 2011.

[35] H. P. Luhn. A business intelligence system. *IBM J. Res. Dev.*, 2(4):314–319, oct 1958.

[36] Shilpi Malhotra and Ashutosh Dixit. An effective approach for news article summarization. *International Journal of Computer Applications*, 76:5–10, 08 2013.

[37] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[38] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond, 2016.

[39] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018.

[40] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745, 2018.

[41] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, and Ryan T. McDonald. Planning with entity chains for abstractive summarization. *CoRR*, abs/2104.07606, 2021.

[42] P. Over and J. Yen. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems. `http://www-nlpir.nist.gov/projects/duc/pubs/2003slides/duc2003intro.pdf`, 2003.

[43] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A., June 2014. Association for Computational Linguistics.

[44] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. Detecting extraneous content in podcasts, 2021.

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[46] Edison Research. The infinite dial 2022, Mar 2022.

[47] Emma Rodero. A comparative analysis of speech rate and perception in radio bulletins. *Text and Talk*, 32:391–411, 05 2012.

[48] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November 2020. Association for Computational Linguistics.

[49] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proc. NIPS*, Montreal, CA, 2014.

[50] Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265, 2021.

[51] Edgar Tanaka, Ann Clifton, and Md. Iftekhar Tanveer. Multilingual podcast summarization using longformers. 2022.

[52] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[54] Denis Visentin, Cleary Michelle, and Glenn Hunt. The earnestness of being important : Reporting non-significant statistical results. *Journal of Advanced Nursing*, 76, 04 2020.

[55] Chao Wang, Judith Gaspers, Thi Ngoc Quynh Do, and Hui Jiang. Exploring cross-lingual transfer learning with unsupervised machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2011–2020, Online, August 2021. Association for Computational Linguistics.

[56] Wikipedia. Inverted pyramid (journalism) — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Inverted%20pyramid%20(journalism)&oldid=1095643247`, 2022. [Online; accessed 09-July-2022].

[57] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[58] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.

[59] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777, 2019.

[60] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

[61] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *CoRR*, abs/1909.02622, 2019.

.1 Appendix 1: Examples of summaries for an episode in Portuguese

Model	Predicted summary
Episode description by creator	Andrei Spacov, sócio e economista chefe da Exploritas, gestora de recursos independentes com foco em renda variável e fixa no Brasil, falou sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global, eleições americanas e também os principais riscos para o investidor se atentar neste ano.
First Minute baseline	hoje eu converso com andreas pacote economista chefe da explore tas uma gestora de recursos independentes com foco em renda fixa e variável no brasil e na américa latina a gente vai falar sobre as expectativas do mercado doméstico e internacional para dois mil e vinte os desafios fiscais no brasil e na argentina o cenário para o crescimento global e as eleições americanas e por aí vai se inscreva no nosso canal e fique conosco andrei recentemente você se tornou sócio da explorer tas com você também entrou o edson sarti que é um gestor e um trader bastante experiente né é foi um ano bastante turbulento para as portas no ano passado principalmente por conta dos choques com argentina conta para mim o que que muda com a entrada de vocês bom obrigado pela oportunidade de estar aqui é compartilhando com vocês a nossa visão é o explore todas desde o início ele ele é um fundo que tem
TextRank Top 2 sentences	O consenso né entre os economistas é que é esse tipo de imposto ele é regressivo é no sentido de que é por exemplo pobre faz muita ele consome bastante para parte da sua renda ele faz muito a transação é financeira EE ele acabaria pagando é mais do que o rico não é é por conta disso apesar de ser em valores bem pequenininhos não é é tem essa questão da agressividade do lado bom da do imposto como de transações financeiras é. É essa tendência mundial já vem de muito tempo né ela é muito estudada tem a ver com é questões é democrata demográficas não é desculpa é as pessoas é tão vivendo mais aumenta expectativa de vida então as pessoas tem que aumentar a poupança ao longo da vida é então é é um fenômeno longo e está muito difícil ver isso se revertendo rapidamente né se a gente olhar para para as inflações nos países desenvolvidos.

TextRank Top 5 sentences	Ou uma eleição com mais sal não é com um cara mais esquerda como Sanders ou é Elizabeth Warren isso isso é coisa rápida é entre fevereiro e meados de março que a gente vai saber é isso vai determinar como é que se comporta o resto do ano né é mas certamente vai ser um tema que vai vai vai acompanhar os mercados aí esse ano não é difícil não ser diferente agora falando em economia doméstica é o crescimento do PIB brasileiro tu achas que ele vai conseguir se descolar? É sumido não é isso é o Chile especificamente é em relação a todos né por exemplo na Colômbia agora dia 21 a gente tem de 21 de janeiro é tem marcada e uma já 11 manifestação grande tudo mais que a gente tem que acompanhar de perto como é que vai ser é mas o fato é que acalmaram bem né e principalmente no Chile em que que isso ganhou proporções muito grandes né é para para para as nossas teses de investimento né o histórico do nosso fun? É está é atacando uma parte desses gastos obrigatórios a outra parte é que é muito importante atacar é os gastos com pessoal é funcionalismo público que é exatamente o que essas essa reforma administrativa essa PEC emergencial fazem né então é é eu acho que assim para efeito de mercado essas reformas podem ter um impacto parecido com o que teve o teto de gastos lá atrás né que foi bem bastante positivo? O consenso né entre os economistas é que é esse tipo de imposto ele é regressivo é no sentido de que é por exemplo pobre faz muita ele consome bastante para parte da sua renda ele faz muito a transação é financeira EE ele acabaria pagando é mais do que o rico não é é por conta disso apesar de ser em valores bem pequenininhos não é é tem essa questão da agressividade do lado bom da do imposto como de transações financeiras é. É essa tendência mundial já vem de muito tempo né ela é muito estudada tem a ver com é questões é democrata demográficas não é desculpa é as pessoas é tão vivendo mais aumenta expectativa de vida então as pessoas tem que aumentar a poupança ao longo da vida é então é é um fenômeno longo e está muito difícil ver isso se revertendo rapidamente né se a gente olhar para para as inflações nos países desenvolvidos.

MBART vanilla	Hoje Hoje eu converso com Andreas pacote economista chefe da explore TAS uma gestora de recursos independentes com foco em renda fixa e variável no Brasil e na América Latina a gente vai falar sobre as expectativas do mercado doméstico e Internacional para 2020 os desafios fiscais no Brasil e na Argentina o cenário para o crescimento global e as eleições americanas e por aí vai se inscreva no nosso canal e fique conosco. Andrei recentemente você se tornou sócio da Explorer TAS com você também entrou o Edson sarti que é um gestor e um trader bastante experiente né é foi um ano bastante turbulento para as portas no ano passado principalmente por conta dos choques com Argentina conta conta para mim o que que que muda com a entrada de vocês. Bom obrigado pela oportunidade de estar aqui é. Compartilhando com vocês a nossa visão.
XLSUM vanilla	O ano passado foi um ano bastante turbulento para as portas.
Episode description by creator	Andrei Spacov, sócio e economista chefe da Exploritas, gestora de recursos independentes com foco em renda variável e fixa no Brasil, falou sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global, eleições americanas e também os principais riscos para o investidor se atentar neste ano.
MBART + fine-tuned PT podcasts	O economista-chefe da Exploritas, Andrea Pacote, conversou com o economista-chefe da Exploritas, Edson Sarti, sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.
MBART + fine-tuned EN podcasts	Andreas pacote economista and Andreas pacote economista chefe Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Exploration TAS, uma gestora de recursos independentes with a Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Explo Exploration TAS, is a co co co co co co co co Gabriel Andreas pacote economista and Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Explo Explo Exploration TAS, a gestora of ER Andreas pacote economista and Andreas pacote economista Chefe da Explo Explo Explo Explo Exploration TAS, a gestora of explore explore explore TAS, a gestora de recursos independents, uma gestora de recursos independientes, with a geo

MBART + fine-tuned PT/EN podcasts	Neste episódio, o economista-chefe da Explorer TAS, Andreas Pacote, fala sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.
LongMBART + finetuned PT podcasts	No episódio de hoje, eu converso com o Andreas Paquito, Economista-chefe da Exploritas, uma gestora de recursos independentes com foco em renda fixa e variável no Brasil e na América Latina. Falamos sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.
LongMBART + finetuned EN podcasts	In this episode, I spoke with Andreas Pak economist, chefe da Explore TAS, about the expectations of mercado doméstico e Internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crecimiento global e as eleições americanas.
LongMBART + finetuned PT/EN podcasts	Andreas Paco, Economista-Chefe da Explorer TAS, fala sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, cenário para o crescimento global e eleições americanas.

Table 1: Summaries of a podcast episode in Portuguese.

.2 Appendix 2: Examples of summaries for an episode in English

Model	Predicted summary
Episode description by creator	In this episode, Dr. Lisa and Tom Gleason introduce listeners to the fundamentals of sound healing, including the origins and benefits of this therapeutic practice. — This episode is sponsored by · Anchor: The easiest way to make a podcast. https://anchor.fm/app
First Minute baseline	We've been having so much fun making these podcasts. If you're thinking about making a podcast you should think about anchor anchor is the easiest way to make a podcast. Let me explain a little bit about this creation tool. It's free these tools allow you to record and edit your podcast right from your phone or your computer and then anchor distributes your podcast for you, so it can be heard on Spotify Apple podcast and all other major podcasting platforms. And here's the best part you can make money from your Podcast with no minimum listenership. It's everything you need to podcast in one place. Just go ahead and download the free anchor app or go to Anchor dot f m– to get started. Hello everyone, and thank you for tuning in to Good Vibration sound healing the Art and Science of vibro acoustic sound therapy. I really appreciate
TextRank Top 2 sentences	So a sound healing was just kind of a natural progression from my music and I started to experiment a little with sound frequencies in songs and things of that nature and I just got really excited about the power of sound and we all know that a song can certainly touch Us in such a deep way and it's the same thing for for for sound healing as well. In fact, we were trying not to he ate music and not trying to organize sound but to distill sound down to notes and use those very intentionally and specifically and what I thought was really interesting was the fact that sometimes what didn't sound necessarily musical had some of the highest healing properties right such as gongs and things like that that just reverberates so deeply, but it was very difficult for me because I immediately when I hear Sam Sound of any kind, I immediately tried to give it Melody and Harmony and I create around that in a musical way.

TextRank Top 5 sentences	We'll talk a little bit more about myths the tools the benefits and of course the science behind this amazing therapy before we get into talking more about sound healing or sound therapy. So a sound healing was just kind of a natural progression from my music and I started to experiment a little with sound frequencies in songs and things of that nature and I just got really excited about the power of sound and we all know that a song can certainly touch Us in such a deep way and it's the same thing for for for sound healing as well. In fact, we were trying not to he ate music and not trying to organize sound but to distill sound down to notes and use those very intentionally and specifically and what I thought was really interesting was the fact that sometimes what didn't sound necessarily musical had some of the highest healing properties right such as gongs and things like that that just reverberates so deeply, but it was very difficult for me because I immediately when I hear Sam Sound of any kind, I immediately tried to give it Melody and Harmony and I create around that in a musical way. We looked at a variety of training organizations from organizations in California to New York City and we kept coming back to a smaller training organization in New York state and in particular in the area that we spend a lot of time up in the Catskills and we Sound Stage Academy of sound healing and Tom and I that's kind of our happy place going up to the Catskills. I think it's the only practice that I can think of that is fully reciprocal right like while we are practicing and Tom and I do sessions one-on-one with clients or we'll do a session where Tom and I worked together on a client which is, you know, really a like a full bonus kind of session and then we also work in group sound therapy or what we call sound baths where a large group of clients will come together.
MBART vanilla	We've been having so much fun making these podcasts. If you're thinking about making a podcast you should think about anchor anchor is the easiest way to make a podcast. Let me explain a little bit about this creation tool.
XLSUM vanilla	This is a full transcript of BBC Radio 4's Welcome to the Sound of Music series.
Episode description by creator	In this episode, Dr. Lisa and Tom Gleason introduce listeners to the fundamentals of sound healing, including the origins and benefits of this therapeutic practice. — This episode is sponsored by · Anchor: The easiest way to make a podcast. https://anchor.fm/app

MBART + finetuned PT podcasts	Neste episódio, Tom Gleason e Lisa Gleason falam sobre a importância do vibracional sound para a saúde e bem-being.
MBART + finetuned EN podcasts	Welcome to Good Vibration Sound Healing, the Art and Science of Vibrational Sound Therapy. I really appreciate you spending some time with us today. My name is Tom Gleason, and I'm here with my wife Dr. Lisa Gleason, and we are both certified sound practitioners and extremely passionate about this subject. Check in these shows we're going to be discussing in depth all aspects of vibrational sound in healing and try to help you gain an understanding of how to achieve personally a body in balance and overall sense of wellbeing.
MBART + finetuned PT/EN podcasts	Welcome to the first episode of Good Vibration Sound Healing with Dr. Tom Gleason and Dr. Lisa Gleason. In this episode, we introduce the concept of sound and vibration ultherapy, and talk about the various applications over time. We talk a little bit more about myths, tools, benefits, and of course the science behind this amazing therapy. Before we get into talking more about sound healing or sound therapy, we talk a little bit about how we came into this practice.
LongMBART + finetuned PT podcasts	O que é Sound Healing? O
LongMBART + finetuned EN podcasts	Welcome to Good Vibration, Sound Healing, the Art and Science of Vibro Acoustic Sound Therapy. In this episode, Dr. Lisa Gleason and Dr. Tom Gleason discuss all aspects of vibrational sound healing and try to help you gain an understanding of how to achieve personally a body in balance and an overall sense of well-being.
LongMBART + finetuned PT/EN podcasts	In our first episode, Tom and Dr. Lisa discuss the science behind Sound Healing and the use of vibrational sound healing.

Table 2: Summaries for a podcast episode in English.