Text summarization using transfer learning

Extractive and abstractive summarization using BERT and GPT-2 on news and podcast data

Master's thesis in Computer science and engineering

VICTOR RISNE

ADÉLE SIITOVA

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Text summarization using transfer learning

Extractive and abstractive summarization using BERT and GPT-2
on news and podcast data

VICTOR RISNE
ADÉLE SIITOVA

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Text summarization using transfer learning
Extractive and abstractive summarization using BERT and GPT-2 on news and podcast data
VICTOR RISNE
ADÉLE SIITOVA

Company supervisor: Knut Nordin
Supervisor: Olof Mogren, Department of Computer Science and Engineering
Examiner: Richard Johansson, Department of Computer Science and Engineering

Master's Thesis 2019
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2019

Text summarization using transfer learning
Extractive and abstractive summarization using BERT and GPT-2 on news and podcast data
VICTOR RISNE
ADÉLE SIITOVA

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

A summary of a long text document enables people to easily grasp the information of the topic without having the need to read the whole document. This thesis aims to automate text summarization by using two approaches: extractive and abstractive. The former approach utilizes submodular functions and the language representation model BERT, while the latter uses the language model GPT-2. We operate on two types of datasets: CNN/DailyMail, a benchmarked news article dataset and Podcast, a dataset comprised of podcast episode transcripts. The results obtained using the GPT-2 on the CNN/DailyMail dataset are competitive to state-of-the-art. Besides the quantitative evaluation, we also perform a qualitative investigation in the form of a human evaluation, along with inspection of the trained model that demonstrates that it learns reasonable abstractions.

Keywords: transformer, BERT, GPT-2, text summarization, natural language processing.

Acknowledgements

We would like to thank our company supervisor Knut Nordin for his commitment and support throughout the project. Thanking the employees at the company for being inclusive and making us feel welcome. We would also like to thank our examiner Richard Johansson for his valuable input. Lastly, we thank our supervisor Olof Mogren for guiding us and providing useful feedback throughout the project.

Victor Risne, Adéle Siitova, Gothenburg, June 2019

Contents

List of Figures

List of Tables

1
Introduction

Text summarization is the task of creating an abbreviated version of a text document by extracting the most important information from the document. In today's society where we are faced with an enormous amount of data daily, it could be a great benefit to automatically retrieve the most salient parts of a text, which would enable us to easily gain knowledge of topics in long documents. Generally, humans can easily grasp the meaning of a text and formulate a coherent summary using their own words. For machines, the task becomes difficult, as it is harder for them to generate text of the same quality as humans. This is because an ideal summarizer has to form an abstraction of the meaning of an input. The model also needs to determine the most important parts of an input document. Lastly, a generated summary has to be in natural language, such that it can be on par with humans. In other words, a summary needs to be relevant to the document while simultaneously being readable to humans.

1.1 Background

The task of automatic text summarization is one of many tasks in the field of *natural language processing* (NLP). For many of the tasks in NLP, feature learning methods such as word2vec [18] and GloVe [24] have been the go to representation learning approaches. These representations are often referred to as word embeddings in NLP. Training the embeddings on large amounts of data enabled people to simply download the word representations and use them for any task. However, there is a disadvantage to traditional neural embeddings, such as GloVe and word2vec, which lies in the fact that each word has a fixed representation, no matter the context it is in. Take for instance the word *season*, which can both serve as a noun and a verb depending on the context.

In recent research, a number of applications have replaced traditional word embeddings in favor by new pre-trained language models resulting in significant improvements over many NLP tasks.

Text summarization is a well explored area in NLP. There are two main approaches when it comes to summarizing text: *extractive* and *abstractive*. The extractive approach extracts the most salient parts of a document and combines them to make a summary, while the goal of the abstractive approach is to generate a summary, of the quality of a human-written summary, by extracting and paraphrasing the input text. The performance of the extractive method is better compared to that of the abstractive method, mainly due to the challenging nature of the abstractive approach.

Recurrent neural networks (RNNs) are common to use for NLP tasks. A recurrent neural network with attention mechanisms works well for shorter inputs, such as in machine translation, but for summarization the network struggles with the input length due to its sequential processing nature. This gives rise to longer training times caused by difficulties in the ability to use parallelism while training the model.

Recently, a number of publications tackling various NLP tasks have utilized the Transformer [36], which is solely based on attention mechanisms eliminating the use of recurrence. Two of the most prominent releases were the language model GPT-2 [27] and the language representation model BERT [6].

1.2 Aim

In this thesis we aim to investigate whether using novel pre-trained models based on the Transformer, i.e., GPT-2 and BERT, improves the performance of existing methods on the task of text summarization. More specifically, the task can be divided in the following two stages:

- Extractive summarization using submodular functions, where BERT will be used for obtaining sentence embeddings. The extractive summary will serve as input to the abstractive method.

- Abstractive summarization by fine-tuning GPT-2 such that it can generate summaries.

Additionally, we aim to conduct a qualitative analysis of our results by conducting a human evaluation test, where humans will rate the summaries in terms of readability and relevance.

1.3 Problem Definition

There are major differences between the used datasets. First is CNN/DailyMail [10] which has been heavily used in text summarization. It is a dataset comprised of human-written news articles and several highlights, which serve as summaries for the articles. The second dataset, which we denote as Podcast, contains transcribed podcast episodes that were created with Google speech-to-text[1]. The datasets do not only differ in quality and structure, but also length. Articles in the CNN/DailyMail dataset are relatively short compared to the transcribed episodes in the Podcast

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

dataset, since a podcast episode may be up to an hour long its corresponding transcription may result in around 10 000 words. Given this, we can divide the task of generating a summary into two stages.

The long podcast episode transcripts raise an issue with the used model, GPT-2, having a context size of 1024 tokens, meaning that we cannot feed the entire text into the model. This presents the first stage of the summarization task, namely shortening the text such that it can serve as input to the generative GPT-2 model in the abstractive stage.

This will be achieved by investigating two methods:

- Utilizing the methods used in extractive summarization such as submodular optimization to obtain the most salient sentences of the original text.

- Splitting the text such that the most important information remains, which for a dataset with a clear structure, e.g., CNN/DailyMail would be in the beginning of the document. However, seeing as there is no general structure for every podcast we want to examine if we can find structures within each podcast show by analyzing what the model actually learns when training to generate a summary. More specifically, we aim to achieve this by visualizing attention heads.

Once we have an input of desirable length we enter the second stage of the task, which is generating a summary using GPT-2. The summaries will be evaluated with the ROUGE metric [15], which is a standard in text summarization, and the scores will be compared to current state-of-the-art.

To further evaluate the performance of our model, we will conduct two qualitative analyses of the generated summaries, one for each dataset, by letting humans rate the summaries on readability and relevance given the generated summary and the original text together with the original summary.

1.4 Related Work

We have collected a large amount of related work in both the domain of transfer learning as well as the task of text summarization. This section aims to give a brief history of both. First starting with recent advancements in transfer learning, going from computer vision to NLP, then giving a brief overview of relevant methods concerning text summarization.

1.4.1 Transfer Learning

Models trained on large amounts of data, e.g., computer vision models trained on ImageNet [5], can be utilized for cross domain knowledge. That is, we can transfer knowledge obtained in some way from a source domain to a target domain.

This is also known as *transfer learning* [22]. In the case of computer vision models trained on ImageNet, the idea of transfer learning is to use the existing weights of a state-of-the-art model to initialize models for different tasks. Thus, transfer learning enables one to save both time and computational power as well as utilizing transferred knowledge, e.g., structures in a language.

Peters et al. [25] presented Embeddings from Language Models (ELMo), a deep contextualized word representation, that unlike GloVe and word2vec represents words based on their context instead of as fixed representations. Although ELMo representations are an improvement over traditional word embeddings, they are treated as fixed parameters. Thus, a model for a specific task still needs to be trained from scratch.

Howard and Ruder [11] introduced Universal Language Model Fine-tuning (ULMFiT), a transfer learning method that can be applied on any NLP task, providing the opportunity of only fine-tuning a model for a downstream task instead of training it from scratch. This was similar to transfer learning in computer vision.

The fine-tuning strategy is also used by Radford et al. [26]. They presented the Generative Pre-trained Transformer (GPT), which uses minimal task-specific parameters and is trained on different tasks by only fine-tuning the pre-trained parameters. Following the same approach, Devlin et al. [6] introduced Bidirectional Encoder Representations from Transformers (BERT), which uses masked language models in order to enable pre-trained deep bidirectional representations.

In early 2019, Radford et al. [27] presented the GPT-2, which largely resembles their first model the GPT, with a few modifications. GPT-2 is a language model with 1.5 billion parameters, trained to predict the next word on 40 GB of text. However, due to several concerns of misuse, Radford et al. decided not to release the GPT-2 to the public [1]. Instead, they released a smaller version of the model with 117 million parameters, which in size is equivalent to their first published model the GPT. The smaller model, alongside with BERT, will be used in this thesis. To avoid confusion with the original GPT model we will denote the smaller GPT-2 model as GPT-2 throughout the rest of the thesis.

1.4.2 Summarization

There exist a great number of articles on text summarization both using the extractive and abstractive method. In this section we mainly focus on the work that have used the CNN/DailyMail dataset, which comes in two variants: *anonymized* [21] and *non-anonymized*. In the anonymized version, named entities are replaced by identifiers. Since many recent publications use the non-anonymized version, including this thesis, we will refer to that version of the CNN/DailyMail dataset throughout the rest of the section, unless stated otherwise. Additionally, the reported measures refer to ROUGE-1 F_1-score.

Extractive summarization Lin and Bilmes [16] use submodular functions for extractive document summarization, achieving a score of 38.9 in ROUGE-1 evaluated on the DUC 2004 dataset. Similarly, Kågebäck et al. [12] approach the task with submodular functions. While Lin and Bilmes use tf-idf representations of sentences, Kågebäck et al. use continuous vector representations to measure the similarity between sentences. They evaluate their model on the Opiniosis dataset [7] achieving 24.88 in ROUGE-1, compared to Lin and Bilmes who acquired a score of 20.57. Mogren et al. [19] achieved 39.35 in ROUGE-1 Recall on the DUC 2004 dataset using a proposed MULTSUM system that combines different similarity measures.

See et al. used a pointer generator network for abstractive summarization achieving a score of 39.53 on the CNN/DailyMail dataset. However, their extractive baseline model "lead-3", which extracts the three first sentences in a document, yielded a score of 40.34.

Current state-of-the-art in extractive summarization is held by Zhou et al. [39], who propose a neural network framework that jointly learns to score and select sentences. Their proposed method achieves a ROUGE-1 score of 41.71 on the CNN/DailyMail dataset.

Abstractive summarization Similarly to our proposed method, Liu et al [17] utilize an abstractive method with an extractive preprocessing stage to generate Wikipedia articles. The idea behind the preprocessing stage is to rank paragraphs, in order of importance, from the reference links in an article. From the ranked paragraphs they select a subset of tokens and use them as input to a Transformer decoder, where the article serves as the target.

Radford et al. tested their GPT-2 model on the task of summarization by adding the delimiter `TL;DR:`after the document to be summarize, with the idea that the model will start generating a summary based on the delimiter. They used the anonymized version of CNN/DailyMail, achieving a ROUGE-1 score of 29.34.

Zhang et al. [38] achieved a score of 41.71 in ROUGE-1 on the CNN/DailyMail dataset, which is state-of-the-art in abstractive summarization. They propose a two-stage model based on an encoder-decoder architecture, utilizing BERT on both sides.

1.5 Scope

This thesis has been conducted at a company in the audio-streaming industry, offering both music and podcasts. The company provided us with the Podcast dataset, comprised of English transcribed podcast episodes, which is the reason why we decided to limit our work to the English language. Moreover, there have been several publications lately where pre-trained language models have been used for transferring pre-trained knowledge to another domain. Because BERT has shown to be a good candidate for contextual feature extraction, and GPT-2 achieves a strong natural language understanding and is capable of text generation, we decided to limit our scope of investigation to these language models.

2

Theory

This chapter gives a description of the theory behind the concepts used in this thesis. The chapter begins with explaining the theory of *artificial neural networks* and the notion of training such networks with associated concepts, followed by a description of text representation and language models. Lastly, the chapter covers the theory behind the Transformer and the models used specifically in this thesis, namely BERT and GPT-2.

2.1 Artificial Neural Networks

An artificial neural network (ANN) is a computing system loosely inspired by the functions of a human brain. It is constituted of neurons in different layers and connections between them, where each connection holds a certain weight. The quintessential structure of an ANN includes an input layer, output layer and a number of hidden layers between the input and output layer.

An artificial neuron is a computational unit defined by an activation function and its connecting weights. The purpose of the activation function is to decide if a neuron should be active or not. It does this by taking as input a linear transformation of the outputs from neurons in the previous layer and the weights. The idea is that a neuron activates on certain input patterns that it has learned. Learning is done by altering the weights of the neurons.

The basic ANN is the *feedforward neural network* (FFNN), or *multilayer perceptron* (MLP). In an MLP, information flows forward from the input layer through the intermediate layers and to the output layer. For an input vector $\mathbf{x}$, the first hidden layer, $\mathbf{h}^{(1)}$, in a network computes $\mathbf{h}^{(1)} = g(\mathbf{w}^{(1)}\mathbf{x} + b)$, where g is some activation function, $\mathbf{w}^{(1)}$ is the weight vector in the first hidden layer and b is a bias term. These computations continue in the subsequent layers of a network until the output has been computed. An intermediate layer, such as $\mathbf{h}^{(1)}$, can be seen as a vector representation of the input $\mathbf{x}$ to an MLP. Neural word embeddings are an example of this.

Two common ANNs are *convolutional neural networks* (CNNs), often used in computer vision, and *recurrent neural networks* (RNNs), which we have seen more in the field of NLP.

2.1.1 Recurrent Neural Networks

A recurrent neural network [30] is a network suited for processing sequential data. Unlike a feedforward network where data flows forward, an RNN has feedback connections which means that the output is fed back into the model. More specifically, RNNs operate on a sequence containing vectors $\mathbf{x}_t$ with time step index t and in the range of 1 to τ. In figure 2.1 a basic RNN is depicted to the left with the input text *"I have a cat"* and to the right the same network represented as an unrolled computational graph, where each node is associated with one particular time step.

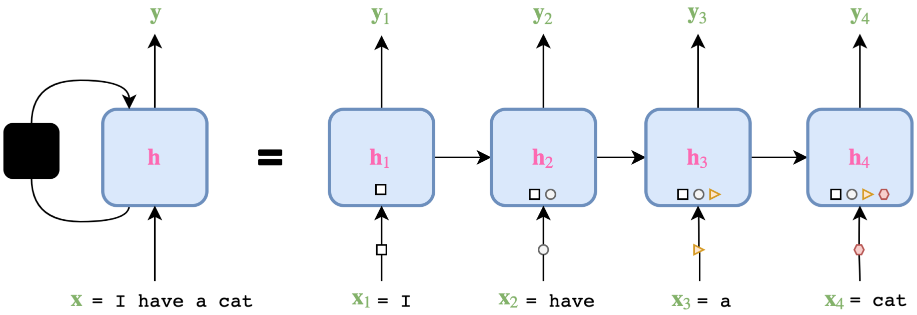


Figure 2.1: *Left-hand side:* A basic RNN with input $\mathbf{x}$, output $\mathbf{y}$ and hidden state $\mathbf{h}$ passed forward through time, where the black square indicates a delay of a timestep. *Right-hand side:* Same network as to the left, only unfolded.

2.2 Training neural networks

There are two main types of training in machine learning, namely *supervised* and *unsupervised*. Supervised learning utilizes labeled data such that for every input x there exists an output y, i.e., a *ground truth* or *target*. This makes supervised learning suitable for tasks such as classification. On the other hand, unsupervised learning lacks knowledge about the output and is more commonly used for finding structures in data making it suitable for, e.g., clustering data. In this thesis we utilize both learning techniques, unsupervised learning in the extractive approach and supervised learning in the abstractive approach.

In supervised learning, the goal is to train a network on classifying for a specific task, such that it learns to generalize for unseen data. This is done by iteratively updating the weights of a network until finding a set of optimal weights that minimize or maximize some *objective function*.

2.2.1 Objective Function

When training a neural network, the goal is to maximize or minimize some *objective function*. A *loss function* is a type of objective function that is to be minimized [8]. The purpose of a loss function is to measure how well a model predicts the expected outcome for any data point in the training set. *Cost function* is the term for the performance measure evaluated on the whole training set [8].

A common loss function is the *cross-entropy error function* as seen in equation 2.1, where P is the target distribution and Q is the distribution of the network's predictions. It is a commonly used objective function in neural network s [29].

$$H(P, Q) = -\mathbb{E}_{x \sim P}[\log Q(x)] = -\sum_{i} P(i) \log Q(i) \tag{2.1}$$

Consequently, minimizing a negative term is the same as maximizing a positive term. This means that minimizing cross-entropy corresponds to maximizing the likelihood of the data, since *maximum likelihood estimation* (MLE) is defined as equation 2.2.

$$\hat{\theta}_{MLE} = \mathbb{E}_{x \sim P}[\log Q(x)] \tag{2.2}$$

2.2.2 Gradient Descent

Gradient descent is an optimization algorithm used to minimize some loss function $J(\theta)$, where θ denotes the parameters of the model [28]. The minimization is done by updating the parameters in the opposite direction of the gradient. The gradients are calculated by backpropagating the loss with respect to its parameters, i.e., backpropagation. The learning rate, η, determines the step size to take for the algorithm to reach a local minimum.

There are three different types of gradient descent, namely *batch gradient descent*, *stochastic gradient descent* and *mini-batch gradient descent*. The difference between them is how much data is used to compute the gradient of the loss function. In batch gradient descent the gradient is computed using the whole training dataset, see equation 2.3. Stochastic gradient descent updates the parameters for each data point in the training set, while mini-batch gradient descent updates the weights using a mini-batch of N data points in the training set.

$$\theta = \theta - \eta \nabla_{\theta} J(\theta) \tag{2.3}$$

2.3 Language Modelling

Language modelling is the task of predicting a sequence of text's probability, often conditioned on being the successor of a given text. People often use language models to predict the next word in a sequence given its predecessor. This can also operate on several levels, e.g., character, word or even sentence. This is generally considered to be an unsupervised task since the data required to train a language model is raw text.

2.3.1 Text Representation

When dealing with natural language, the textual information needs to be represented in some way such that it can be interpreted by a machine. A common way to represent text is by using word embeddings, such as word2vec [18] or GloVe [24], and large vocabularies. However, even with big vocabularies, there is a possibility that *out of vocabulary* (OOV) words occur. One way to overcome OOV was presented by Sennrich et al. [32]. They proposed to work with subwords instead of words using *byte pair encoding* (BPE). This method creates embeddings internally for these subword units.

2.3.2 Sequence-to-Sequence

The purpose of a sequence-to-sequence (seq2seq) model, or encoder-decoder model, is to map an input of variable length to an output of variable length where the input and output length may differ. Proposed by Sutskever et al. [33], a seq2seq model is comprised of two parts: an *encoder* and a *decoder*. Both components contain several recurrent units that take as input a single element. The encoder encodes the input sequence word by word by computing the hidden state h_i for each timestep i. It then passes the last hidden state h_n to the decoder which uses it as its initial state. The purpose of the final hidden state of the encoder is to give a representation of the whole input sequence in form of a fixed-length vector to the decoder. The decoder then uses the hidden state to generate output y_i and the next state s_i for each timestep. See figure 2.2 for a depiction of an encoder-decoder model. An issue arises when the input sequence is very long and the final hidden state of the encoder needs to encapsulate all of the information in a single fixed-length vector, which leads to loss of information.

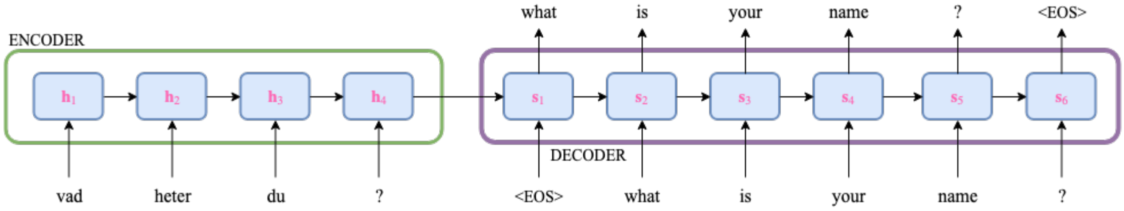


Figure 2.2: Sequence-to-sequence model translating the Swedish sentence to an English sentence.

2.3.3 Attention

The attention mechanism was introduced by Bahdanau et al. [3] as a way to circumvent the issue of handling long sequences. Instead of letting the encoder compress the information into a single fixed-length vector the decoder makes use of a context vector c_i together with the previous state s_{i-1} to generate the output for the current timestep i. The context vector is a linear combination of the hidden states, h_i, of the encoder and the attention weights α_{ij} for each timestep. The weight α_{ij}, seen in Equation 2.4 where $e_{ij} = a(s_{i-1}, h_j)$ and a is an alignment model parametrized as a FFNN, of each h_i is a probability that reflects how important the hidden state is with respect to the previous hidden state s_{i-1} in generating the output y_i.

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{2.4}$$

This attention mechanism enables the decoder to decide which parts of the source sequence to focus on, instead of forcing the encoder to compress all of the information into a single vector and passing it on to the decoder. See Figure 2.3.

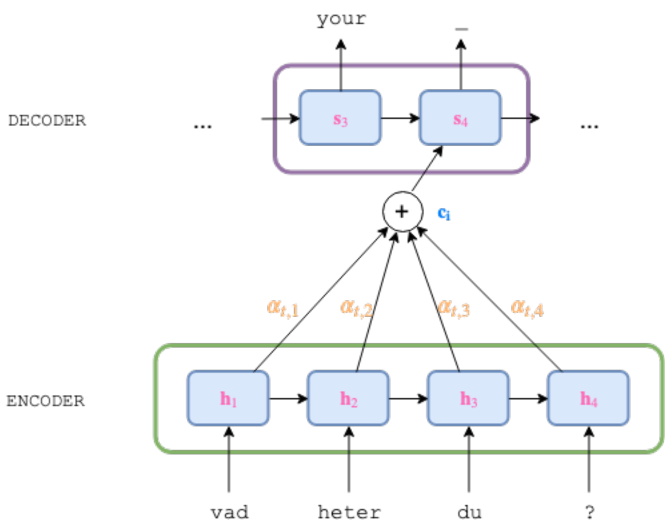


Figure 2.3: Attention mechanism illustrated with a translation task. Here, the model tries to predict the next word in the sequence, i.e, "name" provided the previous state and the context vector.

2.3.4 Training Sequence-to-Sequence Models

A common way to train a neural language model is to use a technique called *teacher forcing* [37]. It is utilized in architectures where the previously generated output is used as input to predict the next output, e.g, encoder-decoder networks. Teacher forcing is when the model produces an output $y(t)$ that is not desirable, thus, instead of using $y(t)$ as input to generate $y(t+1)$ we feed the model the ground truth at timestep $t+1$, i.e, the output we expected from the previous timestep.

2.4 Evaluation

When training the model on the task of text summarization we monitored the perplexity of the model and during evaluation of the produced summaries we utilized a set of metrics called ROUGE.

2.4.1 Perplexity

A common way to evaluate a neural language model is to calculate its perplexity, which is a measure of how accurately the model predicts unseen data. It is defined as 2^{-l}, where $l = \frac{1}{M} \sum_{i=1}^{m} \log p(s_i)$. Here, M denotes the total number of words in the test set and s_i is a sentence in the test set.

2.4.2 ROUGE

ROUGE (*recall-oriented understudy for gisting evaluation*) is an evaluation metric presented by Lin et al. [15]. The authors presented several ROUGE-measures namely ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. These measures automatically determine the quality of a generated summary by comparing it to the ground truth. This is done by counting the number of overlapping n-grams, word sequences and word pairs between the generated summary and ground truth. In this thesis we make use of ROUGE-1, ROUGE-2 and ROUGE-L.

ROUGE-N is an n-gram recall between a generated summary and a set of ground truth summaries, see equation 2.5, where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a generated summary and ground truth.

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{gram_n \in S} Count(gram_n)} \tag{2.5}$$

ROUGE-L stands for longest common subsequence and is calculated as in equation 2.8, where X and Y represent a generated summary and ground truth of length m and n respectively. $LCS(X, Y)$ represents the length of the longest common subsequence of X and Y and $\beta = \frac{P_{lcs}}{R_{lcs}}$ when $\frac{F_{lcs}}{R_{lcs}} = \frac{F_{lcs}}{P_{lcs}}$.

$$R_{lcs} = \frac{LCS(X, Y)}{m} \tag{2.6}$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \tag{2.7}$$

$$F_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{2.8}$$

2.5 Transformer

The Transformer was introduced by Vaswani et al. [36] and mainly developed for machine translation.

2.5.1 Architecture

The Transformer is composed of two larger building blocks, an encoder and a decoder, see Figure 2.4.

Encoder The encoder is comprised of two different modules. The first is multi-head attention (self-attention) with a residual connection [9], which is followed by layer normalization [2]. Thereafter, the result is fed into a feed forward neural network followed by the same residual with addition to a normalization technique.

Decoder In addition to the two modules from the encoder, the decoder inserts a third module right before the self-attention. This submodule modifies the existing self-attention to ignore subsequent tokens. Therefor, the decoder is considered unidirectional. The subsequent self-attention is also modified to be an encoder-decoder attention mechanism such that it attends over the encoder. The previously described residual connection and layer normalization are applied after each layer.

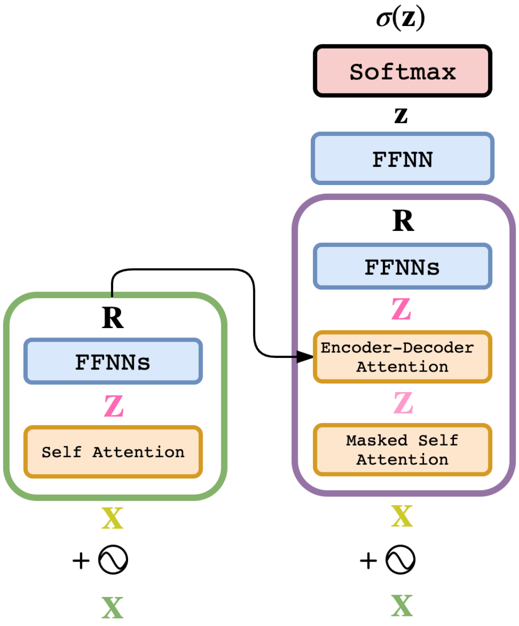


Figure 2.4: Transformer architecture [36]. Encoder (green) performs self-attention. Decoder (purple) modifies the self-attention to prevent from attending to subsequent positions, and is therefore unidirectional. The decoder also attends over the entire encoder using encoder decoder attention. The image omits the residual connection around each submodule followed by layer normalization.

Figure 2.5 depicts the larger building blocks of an encoder. Both input $\mathbf{x}_i$ and output $\mathbf{r}_i$ have the same dimension such that the encoder block can be stacked. The intermediate vectors $\mathbf{z}_i$ are the vectors calculated by the self-attention mechanism. The figure also includes residual connections and layer normalization.

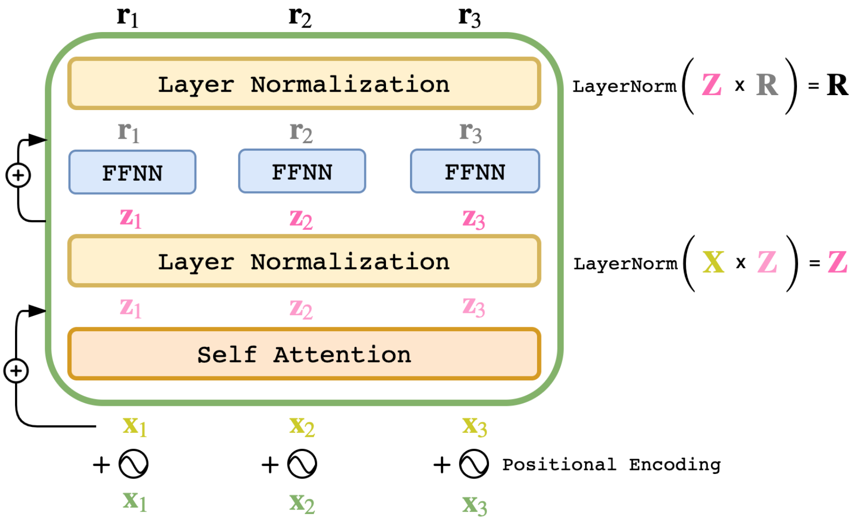


Figure 2.5: Inputs, intermediate calculations and outputs of the Transformer. In this example $\mathbf{Z}$ is the resulting matrix from self-attention which is a stack of $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. $\mathbf{R}$ is the resulting matrix which is served as the input to upcoming encoder layer. Left, arrows are going from $\mathbf{x} \to \mathbf{z}$ and $\mathbf{z} \to \mathbf{r}$ which depicts the residual connections. These are followed by layer normalization.

2.5.2 Self Attention

The self-attention mechanism can be seen as a three staged process. The stages are visualized in Figure 2.7 and described step by step below.

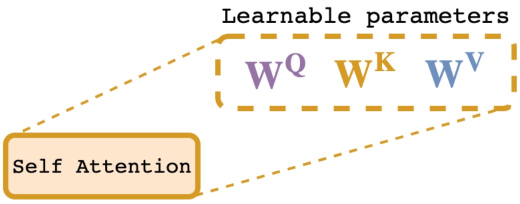
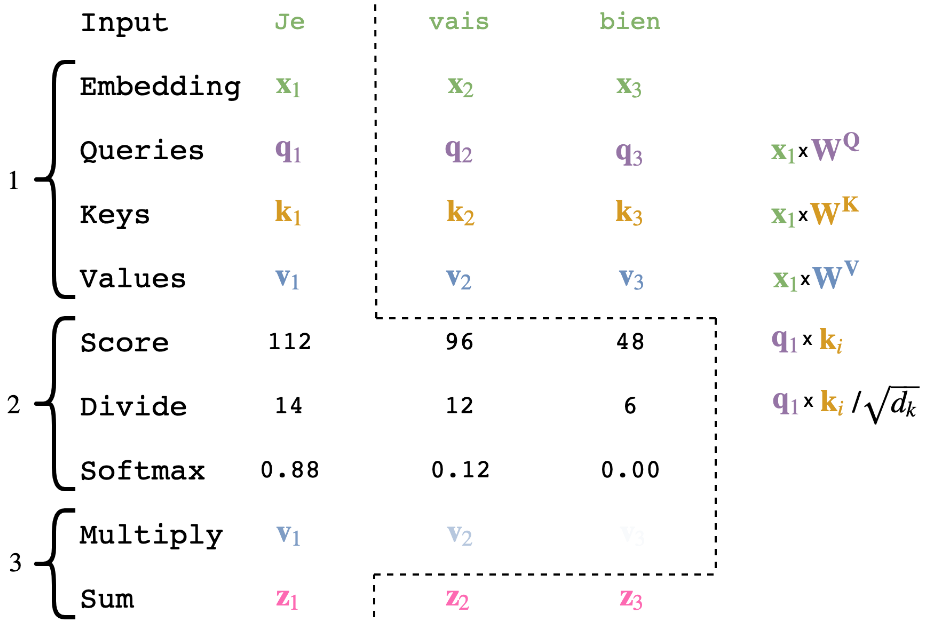


Figure 2.6: The learnable parameters in self-attention, masked self-attention and encoder-decoder attention.

1. Calculating Q, K, V Each embedded representation of the input is used, alongside the learnable parameters $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$ and $\mathbf{W}^{\mathbf{V}}$, to compute a query, key and value vector. These are computed by taking the dot product between corresponding matrix and the embedding vector.

2. Scoring This stage calculates the attention scoring. When scoring the attention for the input "Je" its corresponding query $\mathbf{q}_1$ is used to take the dot product with every key $\mathbf{k}_i$ in the sequence. The scores are then normalized by dividing each score by $\sqrt{d_k}$, where d_k is the dimensionality of the key vector. Lastly, a softmax is applied to all of the numbers to score their importance.

3. Weighted summation To produce the final output $\mathbf{z}_1$ the weighted sum of the softmax output with corresponding value is taken.



	Input	Je	vais	bien	
1	Embedding	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	
	Queries	$\mathbf{q}_1$	$\mathbf{q}_2$	$\mathbf{q}_3$	$\mathbf{x}_1 \times \mathbf{W}^{\mathbf{Q}}$
	Keys	$\mathbf{k}_1$	$\mathbf{k}_2$	$\mathbf{k}_3$	$\mathbf{x}_1 \times \mathbf{W}^{\mathbf{K}}$
	Values	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{x}_1 \times \mathbf{W}^{\mathbf{V}}$
2	Score	112	96	48	$\mathbf{q}_1 \times \mathbf{k}_i$
	Divide	14	12	6	$\mathbf{q}_1 \times \mathbf{k}_i / \sqrt{d_k}$
	Softmax	0.88	0.12	0.00	
3	Multiply	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	
	Sum	$\mathbf{z}_1$	$\mathbf{z}_2$	$\mathbf{z}_3$	

Figure 2.7: A three staged process of obtaining the attention output $\mathbf{Z}$ given the embeddings of an input sequence $\mathbf{X}$

However, in the actual implementation vectors are disregarded in favor of matrices which results in faster computations. Figure 2.8 describes the computations needed to acquire the self-attention output $\mathbf{Z}$ from input $\mathbf{X}$.

$$\begin{aligned} \mathbf{X} \times \mathbf{W}^{\mathbf{Q}} &= \mathbf{Q} \\ \mathbf{X} \times \mathbf{W}^{\mathbf{K}} &= \mathbf{K} \\ \mathbf{X} \times \mathbf{W}^{\mathbf{V}} &= \mathbf{V} \end{aligned} \qquad \operatorname{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^{\mathsf{T}}}{\sqrt{d_k}}\right)\mathbf{V} = \mathbf{Z}$$

Figure 2.8: Query, key and value matrix all calculated using corresponding weight matrix $\mathbf{W}^{\mathbf{Q}}, \mathbf{W}^{\mathbf{K}}$ and $\mathbf{W}^{\mathbf{V}}$ in conjunction with input matrix $\mathbf{X}$. The scaling factor $\sqrt{d_k}$ is the square root of the dimension of the key vectors.

2.5.3 Multi-head Attention

The Transformer does not only use one attention head per layer but eight of them. Doing this improves the model's capabilities of attending other parts of an input sequence. This gives each layer 24 learnable weight matrices $\mathbf{W^Q}_0, \mathbf{W^K}_0, \mathbf{W^V}_0 \dots \mathbf{W^Q}_7, \mathbf{W^K}_7, \mathbf{W^V}_7$. This in turn yields $\mathbf{Z}_0 \dots \mathbf{Z}_7$ which is multiplied by another weight matrix to give the final output $\mathbf{Z}$ of the multi-head attention layer.

2.6 GPT-2

GPT-2 with 117M trainable parameters is a Transformer based language model comprised of 12 slightly modified Transformer decoder blocks stacked upon each other, see Figure 2.9. Each block consists of layer normalization, residual connection around masked self-attention, layer normalization and lastly a residual connection around a feed forward neural network. After the 12 blocks, layer normalization is followed.

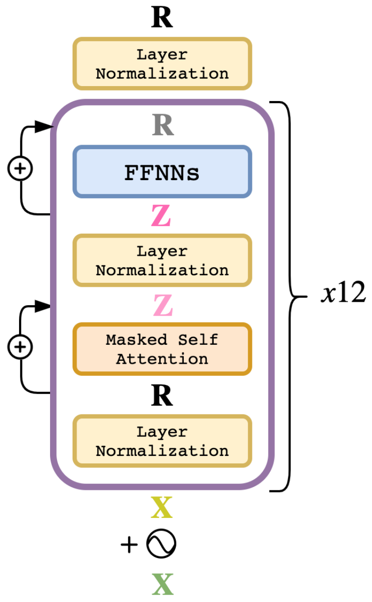


Figure 2.9: Architecture of the OpenAI GPT which is solely based on the Transformer decoder.

2.7 BERT

Devlin et al. [6] utilized the Transformer encoder for the architecture of their model BERT, which is comprised of 12 Transformer encoder stacked upon each other as depicted in Figure 2.10. This means that BERT is bidirectional, i.e., context aware about previous and upcoming tokens. Consider a language model that is unidirectional, such as GPT-2, which is efficiently trained by predicting the next word conditioned on previous information. Doing this with BERT would allow the model to indirectly see the upcoming word. Therefore, the model is trained on sentences where a certain percentage of the words are masked and the task is to predict the masked words, i.e., a cloze task [34]. Thus, BERT is trained as a masked language model. BERT is also trained to predict whether two sentences are adjacent or not.

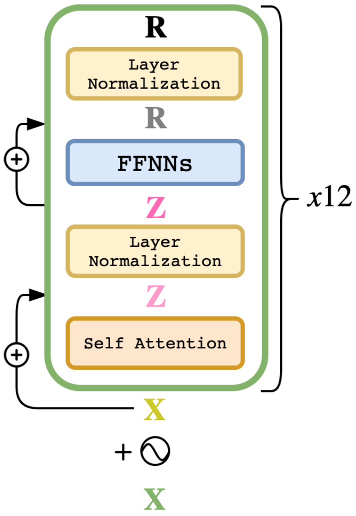


Figure 2.10: Architecture of BERT which is solely based on the Transformer encoder.

3
Methods

In this chapter we present the methods that were used in order to reach the aims of this thesis. Firstly, we give a description of the used datasets and preprocessing of the datasets. Secondly, we describe the two approaches taken on the task of text summarization: extractive and abstractive. Lastly, the evaluation metric ROGUE and the human evaluation method are explained.

3.1 Datasets

We have worked with two datasets in this thesis. The first dataset is the Podcast dataset, provided by the company. It is comprised of 12,983 transcribed podcast episodes with corresponding human-written summaries. Each episode originates from a certain podcast show. Table 3.1 presents the top four shows and their proportion in the dataset. The average word count in an episode transcript is 7,423 words. We hold out 10% of the transcribed podcast episodes to use as test data.

Show	**Episodes per show (%)**
My Brother, My Brother and Me	17.4
The Dollop	12.9
Last Podcast on the Left	12.3
My Favorite Murder	9.9

Table 3.1: Top 4 shows in the Podcast dataset.

The second dataset used is CNN/DailyMail [10, 20], since it is often used in summarization tasks. The dataset is comprised of online news articles with corresponding summaries. There exist two official versions of the dataset: *anonymized* and *non-anonymized*, where the anonymized version replaces named entities with unique identifiers while the non-anonymized version is simply the original article. Following See et al. [31], we operate on the non-anonymized version and use the same data split, i.e, 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs where a pair consists of a document and its summary.

3.1.1 Data Preprocessing

The Podcast dataset is comprised of 12,983 transcribed podcast episodes, where each episode has been transcribed with Google Speech-to-Text[1]. Since the dataset was created using Google Speech-to-Text, the transcribed text lacks punctuation which is needed to split the text into sentence vectors. To introduce punctuation we used a bi-directional RNN [35] to predict missing punctuation.

The CNN/DailyMail dataset has been preprocessed by lowercasing, removing all Unicode characters and introducing space between punctuation. This is because we want to use the same data as others who report on the CNN/DailyMail dataset. Before we used the preprocessed version of the dataset we operated on the original data without any preprocessing.

3.2 Extractive Summarization

Lin and Bilmes [16] created a class of submodular functions for the task of extractive summarization. A submodular function $\mathcal{F}$ is one that satisfies the property of *diminishing returns*: for any $A \subseteq B \subseteq V \setminus \{v\}$, where V is a set of sentences, $\mathcal{F}$ must satisfy:

$$\mathcal{F}(A + \{v\}) - \mathcal{F}(A) \geq \mathcal{F}(B + \{v\}) - \mathcal{F}(B) \tag{3.1}$$

That is, it is more valuable to add sentence v to the smaller set of sentences A than it is to B. They formulated the following objective function, which defines the quality of a summary:

$$\mathcal{F}(S) = \mathcal{L}(S) + \lambda \mathcal{R}(S) \tag{3.2}$$

where S is the summary and λ is a blending coefficient between the coverage $\mathcal{L}(S)$ and diversity $\mathcal{R}(S)$. The aim is to find a summary which maximizes diversity of the sentences chosen for the summary as well as coverage from the original document. A definition of both $\mathcal{L}(S)$ and $\mathcal{R}(S)$ can be found in Lin and Bilmes original paper [16].

The first step in creating an extractive summary is to split the input text into sentences and finding vector representations of each sentence. We used BERT to extract the high dimensional context vector for every sentence. By doing this, every sentence in the entire text document becomes their own representation where similar context vectors represent similar sentences. See Figure 3.1 for a depiction of the procedure.

This extractive approach is similar to that of Kågebäck et al. [12]. The difference lies in the representation of sentences, where Kågebäck et al. used continuous vector representations but we chose to acquire those by using BERT. Like Kågebäck et al.,

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

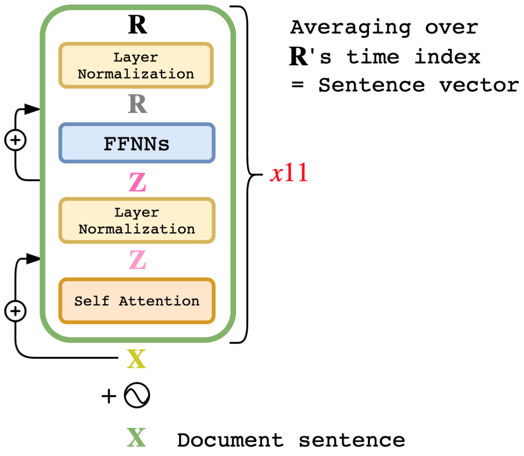


Figure 3.1: The first eleven encoders from BERT is used to acquire a sentence vector.

the similarity measure used in our work was the cosine similarity transformed to lie in the interval of $[0, 1]$ as:

$$\texttt{Sim}(i, j) = \left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} + 1 \right) / 2 \qquad (3.3)$$

The sentence vectors were computed by taking the output from the 11:th encoder and averaged over its time axis. We averaged over the outputs time index so our resulting vector is a sentence vector instead of contextualized word embeddings. Intuitively, the last encoder is close to training output so it could be biased towards pre-trained targets. Thus, using the second to last hidden state. This is visualized in Figure 3.2 where each subplot is a dimensionality reduction, principal component analysis [23], of data from UCI-News Aggregator Dataset[2] in every hidden state.

Since the extractive preprocessing stage was discarded, due to inefficiency discussed in Chapter 4, we had to experiment with which part of the transcript would be used as input to the abstractive model. We tried two approaches, namely, using the beginning of a transcript as input and the end of a transcript. These sliced versions of the data were also used to train the models. We tested these two approaches on the Podcast test set, i.e., 10% of the dataset.

3.2.1 Baseline Lead-x

The lead-3 baseline was proposed by Nallapti et al. [20] and was used on the anonymized version of the CNN/DailyMail dataset, achieving a ROUGE-1 F_1 score of 39.2. See et al. [31] tested the baseline on the non-anonymized version of the CNN/DailyMail dataset and achieved a ROUGE-1 F_1 score of 40.34. The method extracts the first three sentences of a document and uses it as a summary. Our baseline is similar to lead-3, only we extract the leading x words which corresponds

[2] https://www.kaggle.com/uciml/news-aggregator-dataset

[3] https://github.com/hanxiao/bert-as-service

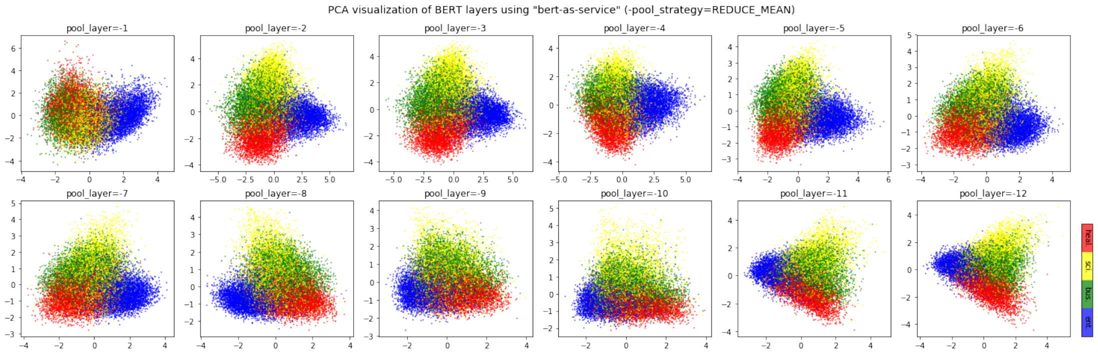


Figure 3.2: BERT as a Service[3] illustrating the four different classes in each of the hidden states in UCI-News Aggregator Dataset. Top left corner (pool_layer=-1) is the output from the last encoder and bottom right corner (pool_layer=-12) is the output from the first encoder.

to the length of the target summary. The reason to why we did not chose to extract the leading three sentences is because of how the Podcast dataset was created. Since we used a bidirectional RNN to predict the positioning of the punctuation, some sentences that were created consist of only two to three words which may have resulted in a very short summary.

3.3 Abstractive Summarization

Similar to Liu et al., we consider a Transformer decoder. The decoder is a good choice for generating text because it can be considered a generative language model, and can be sampled from sequentially. Since the Transformer decoder masks subsequent tokens the problem of self attending on future tokens is omitted. This enables using unsupervised learning by utilizing a big unlabeled dataset to predict the upcoming word. Additionally, it has been reported that the Transformer encoder-decoder architecture stops learning at longer sequences [17], which is why we chose to only utilize the decoder. More specifically we use GPT-2 with 117M parameters as our model, see Figure 3.3. This model layout is never altered and is used as is, fine-tuned to perform abstractive summarization. Our model was trained for two epochs and the complete fine-tuning process took roughly 14 hours.

To produce an abstract summary we only modified the input to our model. We follow the same method to generate a summary as proposed by Liu et al. [17] and Radford et al. [27]. The input to our model is then $(\mathbf{x}_1, \ldots, \mathbf{x}_n, \delta, \mathbf{y}_1, \ldots, \mathbf{y}_m)$ where δ is the vector representation of the `TL;DR:` delimiter, $\mathbf{x}_i$ and $\mathbf{y}_i$ is an integer representation of a token from the source document and generated summary respectively. Given this, the model can be trained to predict the next word conditioned on the antecedent sequence. Similar to Radford et al. [27], the model generated 100 tokens and considered this to be the abstract summary. Generating a token was done by using top-k truncated sampling, which is sampling from a multinomial distribution over the top-k candidates, where k is set to 2. This gave the model some stochastic

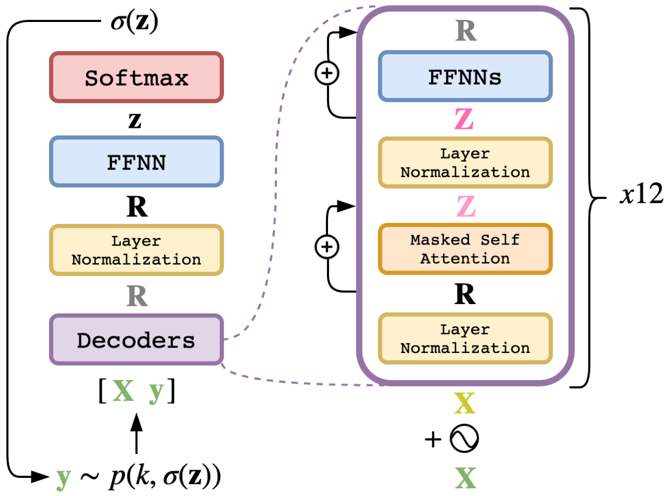


Figure 3.3: Architecture of OpenAI's GPT-2 which is solely based on the Transformer decoder, where $\mathbf{y} \sim p(k, \sigma(\mathbf{z}))$ indicates that generated token $\mathbf{y}$ is drawn from a categorical distribution which consists of the top-k candidates.

behavior. Training was done using teacher forcing and excluding $\mathbf{x}_1, \dots, \mathbf{x}_n$ and δ while computing the loss. Thus only teaching the model to perform abstractive summarization.

3.4 Evaluation

Each experiment was evaluated with ROUGE (ROUGE-1, ROUGE-2 and ROUGE-L) [15]. We monitored the perplexity to ensure that our model did not overfit to our training data.

3.4.1 Human Evaluation

To further evaluate the performance of our models we conducted two surveys for each of the datasets where humans rated a subset of the summaries according to their readability and relevance with respect to the source document[4,5]. We randomly sampled 50 generated summaries from the test set of CNN/DailyMail and presented them together with ground truth summaries and the source document. The same procedure followed for the Podcast dataset, but with a filter that removed generated summaries which scored below 30 in ROUGE-1 F_1. This decision was based on the fact that there were many generated summaries that had a poor quality. Therefore, we only wanted to test those that received a ROUGE-1 F_1 score of at least 30. The Podcast survey was taken by the company employees, while the CNN/DailyMail survey was taken by students at Chalmers University of Technology. We compared these results with Kryściński et al. [14] which did the same experiment and chose

[4] https://podcast-summary-thesis.firebaseapp.com/

[5] https://news-summary-thesis.firebaseapp.com/

to report both readability and relevancy with a 95% confidence interval.

4
Results and Discussion

In this chapter, we present the results obtained from our experiments along with a discussion concerning the results. We also discuss the quality of the datasets used in this thesis. Lastly, we discuss the ethics around large language models, specifically GPT-2.

4.1 Extractive Summarization

Table 4.1 reports the ROUGE F_1-scores we obtained on the Podcast dataset from two different methods: submodular optimization and lead-x baseline. Both methods (lead-3 in previous work) have shown to yield high ROUGE scores on benchmarked datasets, such as DUC 2004 and CNN/DailyMail. However, on this dataset the results are quite low. We believe that this is caused by the quality of the Podcast dataset, which we discuss in depth in Section 4.4. Table 4.2 shows an episode transcript/summary sample together with summaries from submodular optimization and our baseline.

We did want to have an extractive preprocessing stage to determine salient parts of podcast episodes which could not fit into our abstractive model. Liu et al. [17] showed that an extractive preprocessing stage greatly improved their generated summaries. However, seeing that it was not the case for us as we received poor ROUGE scores using submodular optimization we decided to exclude the preprocessing stage and instead feed as much as could fit of the transcript into our abstractive model. Another reason as to why we decided to move on from the extractive stage is that it was time consuming. Extracting a summary consisting of ten sentences took 48 hours for the Podcast dataset.

Model	**R-1**	**R-2**	**R-L**
Submodular optimization	9.62	0.70	6.34
Lead-x baseline	13.00	2.14	11.34

Table 4.1: ROUGE F_1-scores on the Podcast test set.

GPT-2 (ROUGE-1 F_1 26.9): President Trump says he'll be following his nominee and his accuser, and you know what I could be persuaded also what our Christine Blasey Ford and Brett Kavanaugh telling a Senate committee. Key Republicans defend the nominee when a woman says something
Submodular optimization (ROUGE-1 F_1 10.5): I'M Rachel Martin, along with Steve inskeep, and this is a first from NPR news plan to increase pressure on Iran, American diplomats hope to stop Iran's oil exports, whether other world powers like it or not? Stay with us we'll give you the news He gave a speech on the senate floor yesterday Thank you we'll be covering that hearing throughout the day now, when President Trump spoke to reporters last night, he was in New York, he's been there for the United Nations meetings He'S been telling other countries to stop doing business with Iran, or else this all goes back to the 2015 nuclear agreement, which lifted a lot of sanctions against Iran in exchange for one than had to limit its nuclear program We'Ve had many presidents were unable to do anything anything at all with North Korea We have a good relationship and, of course, Brett Kavanaugh, when I look at what's happened to the reputation of a great gentleman, a great intellect, a brilliant man, somebody that has a chance to be one of our great Supreme Court justices in history What did the president say about his Deputy attorney general Rod, Rosenstein, who seemed about to be fired or forced to resign? Just a few days ago, I asked him whether he was going to fire
Lead-x baseline (ROUGE-1 F_1 14.9): President Trump says he'll be following today's testimony from a supreme court nominee and his accuser, and you know what I could be persuaded also what our Christine blasi Ford and Brett Kavanaugh, telling a senate committee. I'M Rachel Martin, al
Target: We discuss President Trump's remarkable press conference and preview the testimony of Brett Kavanaugh and Christine Blasey Ford in front of the Senate Judiciary Committee. Also, the U.S. is reimposing sanctions on Iran, despite European opposition.
Article: President Trump says he'll be following today's testimony from a supreme court nominee and his accuser, and you know what I could be persuaded also what our Christine blasi Ford and Brett Kavanaugh, telling a senate committee. I'M Rachel Martin, along with Steve inskeep, and this is a first from NPR news. Key Republicans defend the nominee when a woman says something she needs to be heard. But when you accuse any person of the crime you can use your needs to be tested. Another Republican described President Trump, as quote uninformed and uncaring, for questioning women who wait gears to report abuse. How open are Senators months and what's the u.s. plan to increase pressure on Iran, American diplomats hope to stop Iran's oil exports, whether other world powers like it or not? (...). The president took reporters questions yesterday for well over an hour. The move refocused attention on the president himself, just as his nominee Brett Kavanaugh, prepares for even more intense scrutiny. Today the president touched on the economy and our economy now is hotter than it's ever been (...) We have a good relationship and, of course, Brett Kavanaugh, when I look at what's happened to the reputation of a great gentleman, a great intellect, a brilliant man, somebody that has a chance to be one of our great Supreme Court justices in history. Intellectually, I think it's a shame. The Supreme Court nominee is now facing multiple allegations of sexual misconduct or assault, as he prepares to testify before the Senate Judiciary Committee at 10 a.m. eastern time today and PR White House reporter Ayesha, Roscoe and NPR Congressional reporter Kelsey Snell are both with us good morning (...)

Table 4.2: Sample article/summary pair from the Podcast test set together with summaries produced by GPT-2, submodular optimization and lead-x.

4.2 Abstractive Summarization

Here we present the results obtained on the Podcast dataset and CNN/DailyMail dataset using the abstractive method. We present two different results on the CNN/DailyMail dataset, one where we used the official version which has predefined dataset splits, e.g., training, validation and test. It is also the version that is preprocessed by lowercasing and the removal of Unicode characters. The other version was downloaded directly as it was created by Hermann et al. [10] and contains the original dataset unprocessed with no predefined dataset splits.

Our experimental setup follows the one used by Radford et al. [27]. This means that we reuse the hyperparameter settings. We tried altering the learning rate, but both decreasing and increasing hurt the performance of the model. However, we did not include dropout since it was only used for classification tasks in the GPT-2 model.

The model was fine-tuned for text summarization using teacher forcing [37] where maximizing the likelihood was our objective. The Adam optimization algorithm [13] was used to minimize the negative log-likelihood on a single NVIDIA Tesla V100 GPU with 16GB of memory. We chose a batch size of one with accumulating gradients of size 128. Increasing the batch size was not possible due to memory constraints of used hardware. When we sequentially sampled our model we chose k equal to 2 in top-k truncated sampling.

4.2.1 Podcast

Table 4.3 reports the ROUGE-1 F_1 scores obtained on the Podcast test set which is denoted as *All*. As we can observe from the table, the ROUGE scores improved significantly when the model was given the beginning of the transcript. Some summaries that were generated when the model was fed with the end of the transcript, were heavily influenced by commercials/promotions which often appear at the end of a podcast episode.

Our hypothesis is that the results may have looked different if the model had a bigger context size, whereas now it is only exposed to a fraction of the transcript and thus not being given all of the information needed to generate a summary of the whole transcript.

We performed additional experiments in which we extracted show-specific episodes and treated them as separate datasets. These results can be found in Table 4.3 where we experimented with the two shows *My Brother, My Brother and Me* and *My Favorite Murder* which make up 17.4% and 9.9% of the dataset respectively. The model was trained for 10 epochs and tested on 10% on each dataset and took roughly three hours.

Dataset	**R-1**	**R-2**	**R-L**
All	16.28	4.09	12.80
All (end)	10.52	1.69	7.92
My Brother, My Brother and Me	19.01	6.23	14.87
My Favorite Murder	23.41	8.62	16.51

Table 4.3: ROUGE F_1-scores on the Podcast test set, denoted as *All* and show specific results.

In Table 4.4 a podcast summary generated from GPT-2 is presented. This summary is based on an episode from the TED Radio show. This example demonstrates that the model is interpreting the podcast episode in its own way. It is trying to paraphrase the title forest ecologist to evolutionary biologist. This could be one effect from letting the model sample at generation instead of always picking the top candidate. We do not however have any evidence supporting this theory, and might as well be an inherited property from the pre-trained language model.

Generated (ROUGE-1 F_1 42.9): How do we shape the world around us through our networks and the pathways that connect us online, and in our bodies and in our relationships? This hour, TED speakers explore how our digital lives shape our bodies, minds and the lives of our loved ones. Guests include evolutionary biologist Suzanne Simard, social psychologist Dr. Robin D'Ambrosio, evolutionary biologist Dr. James Hansen, and environmental activist Jane
Target: Networks surround and sustain us, in nature, in our bodies, in relationships, in the digital world. This hour, TED speakers explore how we rely on networks and how we have the power to shape them. Guests include ecologist Suzanne Simard, UPS executive Wanis Kabbaj, computer scientist Avi Rubin and anthropologist Robin Dunbar. (Original broadcast date: January 13, 2017)
Article: Support for Ted radio hour and the following message come from rocket mortgage by Quicken Loans, introducing rate Shield approval if you qualify and if rates go up, your rate stays the same, but if rates go down, your rate also drops to learn more go to Rocket mortgage.com Ideas - hey it's guy here, so have you ever considered just how many of your Facebook friends are actually friends of yours in real life? Well, in this episode, we explore ideas about the connections that sustain us in the digital world and nature in our bodies and in our relationships. Today'S Show is called networks and it originally aired in January of 2017. This is the TED Radio Hour each week, groundbreaking TED Talks, pet technology, entertainment design design. Is that really what's 10 function? I'Ve never known that delivered a Ted conferences around the world. If the human imagination we've had to believe in impossible thing, the true nature of reality beckons from just beyond those talks. Those ideas adapted for radio, from NPR and Guy Roz, and on today's show ideas about the power of networks. How those connections and those Pathways Define the world around us in our cities, in our relationships in our bodies and especially in nature, so about 25 years ago, Forest ecologist, Suzanne Simard had a hunch. Yes, that's right. She thought that trees could talk. Imagine like when you're walking through the forest, you might you hear the crunching of the Twigs under your feet in the rustling of the leaves, but you thought what, if there's more going on because big chattering going on that, we can't hear that they're attuned to each Other now, at the time and again, this is about 25 years ago, a team of scientists in England were wrapping up an experiment where they'd grown in the laboratory. (...)

Table 4.4: Sample transcript/summary pair together with a generated summary from the Podcast dataset.

4.2.2 CNN/DailyMail

Table 4.5 reports our results on the CNN/DailyMail dataset along with other recent approaches in abstractive summarization. Our model was trained for two epochs and the complete fine-tuning process took roughly 14 hours. Radford et al. [27] tested all four of their models on the anonymized version of CNN/DailyMail test set without having trained them for the task. In Table 4.5 we report the ROUGE scores they obtained on their largest model GPT-2 with 1.5B parameters. Figure 4.1 depicts the distribution over the ROUGE-1 F_1 scores on the official CNN/DailyMail test set.

Model	**R-1**	**R-2**	**R-L**
Radford et al. [27]*	29.34	8.27	26.58
Zhang et al. [38]	**41.71**	**19.49**	**38.79**
Our model	39.38	14.64	35.38
Our model†	33.88	10.82	30.95

Table 4.5: ROUGE F_1-scores on the CNN/DailyMail test set. The $*$ mark indicates that the authors evaluated their model using the anonymized version of CNN/DailyMail. The † indicates that the model was trained and evaluated on the unprocessed version of CNN/DailyMail for one epoch, where 10% of the data was used for the test set.

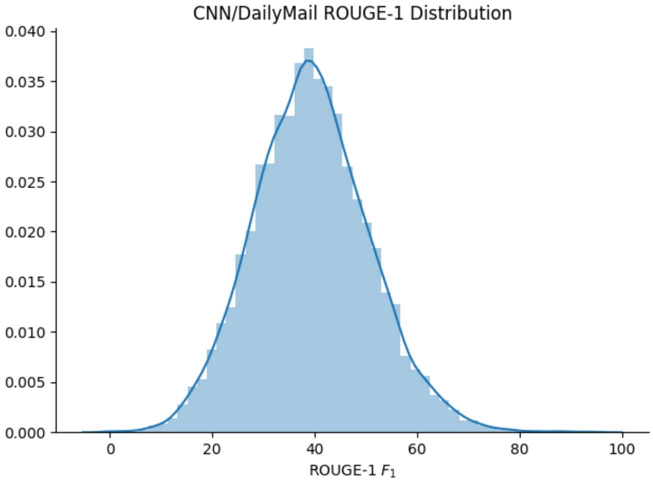


Figure 4.1: Distribution over the articles ROUGE-1 F_1 score on the test set in the official CNN/DailyMail dataset.

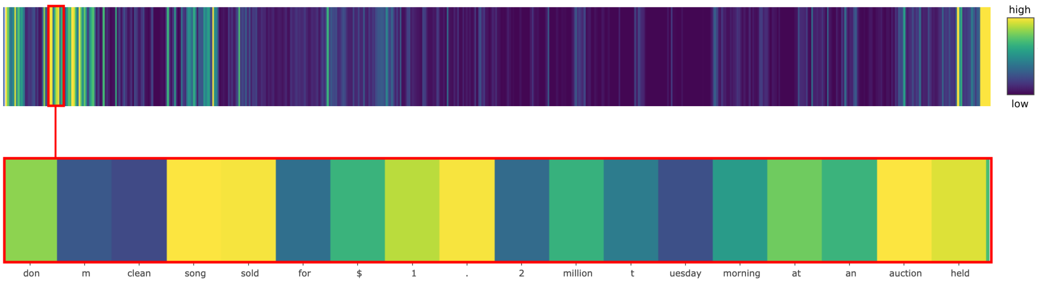
Figure 4.2 shows example outputs from the model for two different input documents. In most example documents we have inspected, the resulting summary is both consistent with the content of the document, and has a high degree of fluency. In Appendix A we further include a few summaries generated by our model.

Figure 4.2 also shows visualizations of the attention pointers during prediction corresponding to each example of the two summaries. Each token of the input document is represented by one slice in the picture, and brighter color means higher attention weight. One can see that the model decides to put its attention in different locations of the input documents depending on the actual content, and not necessarily only in the beginning of the document (which has previously been considered as a useful strategy when creating summaries [31]). Our analysis of the attention pointers during summary prediction demonstrates that the system attends not only to certain positions in the documents (such as the beginning which is usually considered informative in news articles), but dynamically determines where to put its focus for each input document. This is an interesting observation and also a sanity check that our system does not solve this problem trivially by extracting the leading sentences.

Generated (ROUGE-1 F_1 41.2): the lyrics to the famous **don mclean song sold for $ 1.2 million tuesday morning at an auction held** by christie 's . " mclean said it was time to part with the manuscript . "

Target: don mclean 's “ american pie ” lyrics auctioned for $ 1.2 million . the song is dense with symbolism ; mclean says lyrics , notes will reveal meaning . “ pie ” is mclean 's biggest hit , was no. 1 in 1972 .

Article: (cnn) that's some rich “ american pie. ” the lyrics to the famed **don mclean song sold for $ 1.2 million tuesday morning at an auction held** by christie's . “ don mclean's manuscript of ‘ american pie ’ achieved the 3rd highest auction price for an american literary manuscript, a fitting tribute to one the foremost singer-songwriters of his generation, ” christie's tom lecky said in a statement. mclean told rolling stone that it was time to part with the manuscript (...) the record for a popular music manuscript is held by bob dylan's “ like a rolling stone, ” which sold for $ 2 million in june. opinion: what 's so great about ‘ american pie ’?"



Generated (ROUGE-1 F_1 26.9): the unidentified man was admitted to the hospital in may complaining of nausea, weakness, fatigue and body aches. **the 16-daily-cups** habit was striking on its own , since federal studies have found that the average american drinks about 10 to 11 cups of liquid every day , including water, coffee and other beverages. the man was consuming anything from 3 to 10 times **the amount an average american takes in** tea.

Target: doctors at the university of arkansas for medical sciences found a 56-year-old man's kidney problems stemmed from drinking too much iced tea. black tea contains oxalate, a chemical known to produce kidney stones and sometimes lead to kidney failure. the unidentified man will likely spend the rest of his life in dialysis.

Article: (...) the man was admitted to the hospital in may complaining of nausea, weakness, fatigue and body aches. the chemical oxalate was found to have clogged the man's kidneys, to the point that they had become inflamed. with a habit of drinking about 16 8-ounce cups of iced tea every day, the unidentified man was consuming anywhere from 3 to 10 times **the amount an average american takes in. the 16-daily-cups** habit was striking on its own, since federal studies have found that the average us adult drinks about 10 to 11 cups of liquid every day, including water, coffee and other beverages. (...)

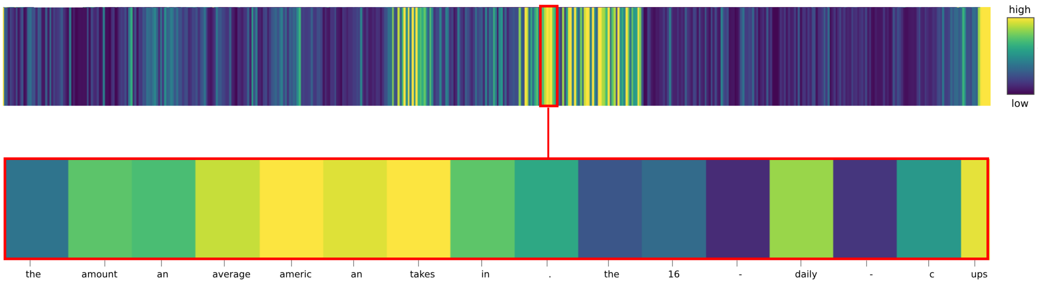


Figure 4.2: Article/summary pairs from the CNN/DailyMail test set with a generated summary. Red text corresponds to where our model paid much attention. Each sample is followed by attention visualization for the article. Top part of an attention figure visualizes the sum of all attention heads from the generated steps for the whole article. The bottom part shows a zoomed in version of the top part marked in red, where there is high attention. Darker colors indicate low attention while brighter colors indicate high attention.

Table 4.6 shows an example of a generated summary that does not act as a summary, but as an extension to the original article. This sample comes from the unprocessed version[1] of the CNN/DailyMail dataset and is one of the side effects which could occur from fine tuning a pre-trained language model on text summarization. The model also produced text summaries which included foreign languages, e.g., Hindi, Chinese, Ukrainian, Arabic etc, see Figure 4.7. Liu et al. [17] did notice the same side effect when training their model. More examples are reported in their paper. We hypothesize that this occurs when a model is trained on data where English translations are accompanied in its original language. This is most likely present in the training data used for pre-training GPT-2 and is often found in Wikipedia articles. Beware that it is just hypothesis and we could not find evidence to confirm this side effect.

Generated (ROUGE-1 F_1 0.0): © 2013 Financial Times Ltd. All rights reserved. This material may not be published, broadcast, rewritten or redistributed
Target: Myanmar cleared bulk of \$11.3B in outstanding foreign debts. Myanmar reached agreements with Paris Club creditors, World Bank, ADB. Paris Club agreement cancels at least 50% of further \$4.4B in bilater al debt.
Article: Yangon (Financial Times) – Myanmar passed a milestone in its efforts to clear its foreign debts after announcing agreements with the Paris Club of creditors, the World Bank, and the Asian Development Bank. (...) © The Financial Times Limited 2013

Table 4.6: Sample from the unprocessed CNN/DailyMail dataset which demonstrates a generated summary that is a continuation of the article, where the model produces an all rights reserved clause.

Generated (ROUGE-1 F_1 0.0): ارحمان القبل امتقبر موشع الموشع القبل امتقبر.
Target: Iran will pursue “new round of diplomatic activity,” Iranian president reportedly says. Ahmadinejad criticizes foreign powers for “meddling” in Iran’s affairs. Obama administration has sought dialogue with Iran while increasing criticism.
Article: TEHRAN, Iran (CNN) – Iranian President Mahmoud Ahmadinejad said he wants to engage President Obama in “negotiations” before international media, a semi-official Iranian news outlet reported on Saturday (...) Meanwhile, the semi-official Fars News Agency reported that 20 people between the ages of 35 to 48 were executed in Iran on Saturday for "buying, selling and holding heroin, cocaine and opium."

Table 4.7: Sample from the unprocessed CNN/DailyMail dataset. Model produces a summary in a foreign language which, according to Google translate, translates to "Arhman al-Qabeel, who was standing in the place of the shrine, kissed him" in Arabic.

Table 4.8 reports four different generated summaries for the same CNN/DailyMail article. The summary generated by our model is somewhat coherent with minimal amount of copying from the original article. However, the generated summary reports fake news as it was the gunman that was shot to death and not the police officer. Additionally, the shooting took place in an African-American neighborhood of Roxbury in Boston, Massachusetts and not in Roxbury, Africa as generated by our model.

[1] `https://github.com/JafferWilson/Process-Data-of-CNN-DailyMail`

Our model (ROUGE-1 F_1 24.7): john moynihan , 34 , was shot to death by officers in roxbury , africa , last month . he was a former u.s. army ranger who was honored at the white house for his heroism in the wake of the boston marathon bombing . " i think people understand that the decisions mr. west made put his life in grave jeopardy , " clergyman mark v. scott says .
See et al., 2017 (ROUGE-1 F_1 34.9) [31]: boston prosecutors released video friday of the shooting of a police officer last month. the gunman shot to death by officers , was black . one said the officers were forced to return fire. he was placed in a medically induced coma at a boston hospital.
Liu et al., 2018 (ROUGE-1 F_1 36.4) [17]: boston prosecutors released video of the shooting of a police officer last month . the shooting occurred in the wake of the boston marathon bombing. the video shows west sprang out and fired a shot with a pistol at officer's face.
Kryściński et al., 2018 (ROUGE-1 F_1 25.4) [14]: new: boston police release video of shooting of officer , john moynihan. new: angelo west had several prior gun convictions , police say. boston police officer john moynihan, 34, survived with a bullet wound . he was in a medically induced coma at a boston hospital , a police officer says.
Target: boston police officer john moynihan is released from the hospital. video shows that the man later shot dead by police in boston opened fire first. moynihan was shot in the face during a traffic stop.
Article: (cnn) to allay possible concerns, boston prosecutors released video friday of the shooting of a police officer last month that resulted in the killing of the gunman. the officer wounded, john moynihan, is white. angelo west, the gunman shot to death by officers, was black. after the shooting, community leaders in the predominantly african-american neighborhood of roxbury, where the shooting occurred (...)

Table 4.8: Sample article/target from the processed CNN/DailyMail test set, together with four different summaries from various approaches.

4.3 Human Evaluation

Table 4.9 shows average score and confidence interval at 95% from the human evaluation performed on the Podcast dataset and the official CNN/DailyMail dataset. The confidence interval was estimated using student's t-distribution, more specifically using scipy stats' implementation[2]. We gathered 110 and 187 samples for our CNN/DailyMail and Podcast survey, respectively. The scale goes from one to ten. Each participant was instructed to drag the sliders, according to their own judgment, based on how well the generated summary performs in readability and relevancy with respect to the news article.

Dataset	Readability	Relevancy
Podcast	5.89 ± 0.40	5.42 ± 0.37
CNN/DailyMail	5.43 ± 0.44	5.92 ± 0.41

Table 4.9: Average score and confidence interval at 95% from the human evaluation performed on the Podcast dataset and the official CNN/DailyMail dataset.

We compare the results of CNN/DailyMail dataset in Table 4.9 with results compiled by Kryściński et al. [14], see Table 4.10. Since Kryściński et al. did not specify exactly which data they used for their human evaluation, the data points are not identical. Neither are the participants of the survey. Therefore comparing the results should be done with some hesitation. However, it does however provide a reference point to previous work. These results show that our model is not far away to be comparable to current models which are on par with state-of-the-art.

[2] `https://docs.scipy.org/doc/scipy/reference/stats.html`

Model	Readability	Relevancy
See et al., 2017 [31]	6.76 ± 0.17	6.73 ± 0.17
Liu et al., 2018 [17]	6.79 ± 0.16	6.74 ± 0.17
Kryściński et al., 2018 [14]	6.35 ± 0.19	6.63 ± 0.18

Table 4.10: Average readability and relevancy and confidence interval at 95% presented on the official CNN/DailyMail dataset. These results were compiled by Kryściński et al.

In our human evaluation the inter annotator agreement on a single data point in the CNN/DailyMail dataset was low. However, we can not conclude anything since there are not enough samples on a single datapoint. In Figure 4.3 we can observe that each document, uniquely color coded, does not necessarily form a cluster. These documents, together with its generated summary, can be found in Appendix A. This shows that opinions and biases make human evaluation difficult. Although we limited our study group to students at Chalmers University of Technology, participants without knowledge in NLP might criticize generated samples more often than its counter part, students with NLP knowledge. This phenomenon could be a result of knowing about the complexity of working with NLP, which in turn might lead to participants being more forgiving when grading presented examples.

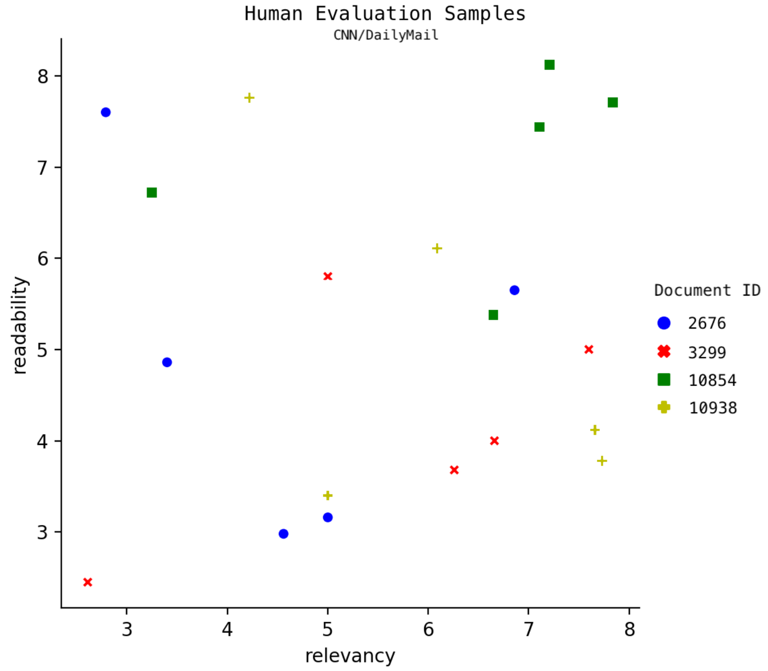


Figure 4.3: A scatter plot showing the results of our human evaluation on four different data points. Each data point is represented by a distinct color and shape, forming a point on the graph, where readability is presented on the y-axis and relevancy on the x-axis.

4.4 Quality of the Datasets

As stated in Section 1.3 there exist several differences between the CNN/DailyMail dataset and the Podcast dataset. The former contains article/summary pairs written in a formal language by humans, often leading the article with a short summary, which is a reason why the lead-3 method was such a strong baseline [31]. The latter dataset has several drawbacks. The first is that it is composed of many episodes that originate from different shows, some of which include *My Brother, My Brother and Me* a comedy advice podcast hosted by three brothers and *The Daily* which is a news podcast hosted by a single person. This means that the dataset contains episodes with different structures, since each show may have a specific structure that they follow. The second drawback is that podcasts often have more than one hosts, this causes a problem when using automatic speech recognition since such systems do not distinguish which person is speaking. Thirdly, podcasts often tend to have a more common everyday language in contrast to news articles. Lastly, many people who upload podcasts to the streaming platform do not include meaningful episode descriptions. They choose to promote their social media sites instead of describing what the episode is actually about.

GPT-2 achieved competitive results on the official version of the CNN/DailyMail dataset. However, the dataset lacks sentence case which results in the model producing summaries that omit the use of capital letters. This favors older methods when working with abstractive summarization but is not necessary when using modern language models which use byte pair encoding. Because of this, our results are less readable since it is not the conventional way of writing. When we trained on the same data but with no preprocessing we still achieved summaries which are competitive enough to be compared to state-of-the-art and also more readable, see Table 4.11.

Generated (ROUGE-1 F_1 41.5): Author J.K. Rowling reveals she is writing a new movie based on the "Harry Potter" series. It's being released in theaters and is based on the "Fantastic Beasts and Where To Find Them". Rowling says she was "excited" at the idea and was "worried" by studio. Warner Bros. will be the first studio to enter into an "expanded creative partnership" with Rowling.
Target: J.K. Rowling will make her screenwriting debut with a wizarding movie. It'll be based on one of Harry Potter's textbooks. Studio Warner Bros. says it's the first in a film series.
Article: (CNN) – J.K. Rowling still has some tricks up her sleeve.The "Harry Potter" author is returning to the world of wizards that made her famous, but it's not going to be with a new book – and it's not going to feature The Boy Who Lived.J.K. Rowling revealed as secret author of crime novel instead, Rowling is going to make her screenwriting debut with a movie centered around one of Harry Potter's textbooks, "Fantastic Beasts and Where To Find Them," and the sojourns of its author, Newt Scamander.According to Rowling, she became intrigued with the idea after a proposal from studio Warner Bros. (...) The laws and customs of the hidden magical society will be familiar to anyone who has read the 'Harry Potter' books or seen the films, but Newt's story will start in New York, 70 years before Harry's gets under way."J.K. Rowling penning 'Harry Potter' encyclopedia According to Warner Bros., "Fantastic Beast s" will be the first in a planned series of films. In addition to the new franchise, Warner Bros. Entertainment has also entered into an "expanded creative partnership" with Rowling, which will cover projects lik e video games, digital initiatives and tourist attractions.For Warner Bros., this has to be a day of celebration. All told, the eight films produced based on Rowling's best—selling books have earned more than $7. 7 billion worldwide, which by 2011 helped it edge out "Star Wars" as the top—grossing film series ever.

Table 4.11: Sample article/summary pair together with a generated summary from the unprocessed CNN/DailyMail dataset.

4.5 Ethics

When Radford et al. [27] published their paper in which they presented the GPT-2 they only released the smallest of their four models. They claimed that their biggest model, the GPT-2, was too powerful to release as it could be used for malicious purposes such as the following:

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

While the list is not exhaustive, it is still enough to raise awareness. Despite the danger the model implies, OpenAI received a huge backlash from the deep learning community for withholding three of their four models. A few months after the release of the original paper the authors released the second to smallest model with 345M parameters to the public. In the release they claim that the misuse risk of the 345M model is higher than that of the 117M model, but still substantially lower than that of 1.5B. The two bigger models of 762M and 1.5B parameters are still being withhold from the public, but shared with partners in academia and non-profits working on counter-measures against the risks of large language models.

In this thesis we got to experiment with the smallest model, GPT-2 with 117M parameters. Neither of our proposed methods are on par with a human when it comes to producing a readable and relevant summary. However, even though our model does not surpass human ability, some results are surprisingly good. Thus, researchers should be transparent and let people know what your model can be used for and encourage skepticism while reading content published online.

In Table 4.4 forest ecologist Suzan Simard was given a new title by our model, namely an evolutionary biologist. This might not be a harmful example, however the model did misinterpret and we had more severe misinterpretations by our model. One example of a generated summary which was not correct follows: "Join Robert as he explains the man who would become the greatest American hero of all time: Osama Bin Laden". This generated summary is close to the opposite of the actual podcast topic, the episode did include conversations regarding what an awful person Osama Bin Laden was. Therefore, it is important to let a human quality control generated summaries to guarantee that the model do not promote hate speech or fake news.

5
Conclusion

The work presented in this thesis indicates that a Transformer model that was pre-trained on large data as a language model can easily be fine-tuned to achieve good results for summarization. Our approach leverages the transfer learning power that we have seen on many other tasks with the Transformer architecture.

In the case where we summarize news articles we note that the summaries produced by the proposed system is both consistent with the input documents, and have a high fluency, as expected by a system based on the Transformer architecture. Once applying our method on podcast data the quality of the generated summaries varies quite a lot. Once in a while the model really gets it right but sometimes it only gets the readability or relevancy part. There also exist cases where neither of these criteria are met. We could have skipped preprocessing podcast episodes with extractive summarization to fit within GPT-2's context size. It would have been more interesting to see if a single abstractive Transformer model could have improved the results if more context was allowed.

For future work in the area of text summarization, we would like to investigate models that accepts larger context size and branch away even more from standard datasets. Once we realized that our extractive preprocessing method did not achieve good results we looked at other ways of working around the context limit in language models. We did encounter the use of recurrent Transformer architectures such as Transformer-XL [4] which is something we would have liked to investigate further. Branching away from standard datasets is something we did explore in this thesis. However, we would like to see models trained exclusively on human written podcast transcripts with episode summaries that are only relevant to the episode.

Bibliography

[1] Dario Amodei Daniela Amodei Jack Clark Miles Brundage Alec Radford, Jeffrey Wu and Ilya Sutskever. Better language models and their implications. `https://openai.com/blog/better-language-models/`, 2019.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[6] 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.

[7] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics, 2010.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[9] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. corr, vol. abs/1512.03385, 2015.

[10] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[11] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339, 2018.

[12] Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. Extractive summarization using continuous vector space models. In *Proceedings*

of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), pages 31–39, 2014.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. Improving abstraction in text summarization. *arXiv preprint arXiv:1808.07913*, 2018.

[15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[16] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics, 2011.

[17] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] Olof Mogren, Mikael Kågebäck, and Devdatt Dubhashi. Extractive summarization by aggregating multiple similarities. *Recent advances in natural language processing 2015*, page 451.

[20] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[21] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[23] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper. pdf*, 2018.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[28] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[29] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

[30] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[31] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[32] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[33] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[34] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

[35] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*, 2016.

[36] 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.

[37] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[38] Haoyu Zhang, Yeyun Gong, Yu Yan, Nan Duan, Jianjun Xu, Ji Wang, Ming Gong, and Ming Zhou. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*, 2019.

[39] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*, 2018.

A

Appendix 1

Generated (ROUGE-1 F_1 43.8): pekingese marley and mitzy were snatched from front garden in doncaster , south yorkshire . neighbours believe they have caught the thief on cctv . pair were left to play in front garden of their home in doncaster , south yorkshire . ” none of us have been able to sleep or eat since . none of us have been able to sleep or eat since ’ "’ none of us have been able to sleep or eat
Target: pekingese dogs marley and mitzy were taken from garden in doncaster . suspected pet-napper caught on camera after the animals were snatched . theft reported to the police and owners offer # 1,000 reward for their return . ’
Article: with the helpless animals clutched tightly under his arms , this is the moment a suspected dog thief made off with two beloved family pets . pekingese marley and mitzy had been left to play in the front garden by their owners adele and steven worgan . but they were snatched when the couple were n’t looking – and neighbours believe they have caught the culprit on cctv . caught on camera : this is the suspected dog-napper seen on cctv footage with what appears to be two animals under his arms . ‘we just want them back , ’ said mrs worgan , 49 .‘ it ’s devastating . none of us have been able to sleep or eat since . ’ mr worgan was inside the house in doncaster at around midday when he heard a yelp from the garden . he rushed outside in time to see a man get into a van with marley and mitzy and drive off . springing into action , the 56-year-old gave chase around midday on saturday – but was unable to keep up as the van sped away . now , the couple ’s neighbours believe that they have captured the alleged thief on cctv . missing : pekingese dogs marley and mitzy disappeared from the front garden of their home in doncaster , south yorkshire . the theft has been reported to the police and the pairs owners are offering a # 1,000 reward for information leading to their safe return . the couple are now offering a # 1,000 reward to anyone who can return their pets , who have been missing for a week , and have reported the theft to police . marley , the younger of the pair , should be easily identifiable because he only has one eye . a police spokesman said inquiries were continuing . any one with information about the dog snatching was asked to call south yorkshire police on 101 quoting incident number 428 of 4 april 2015 ."

Table A.1: Document ID 2676 ●.

Generated (ROUGE-1 F_1 55.0): benaud 's funeral was attended by close family and friends on wednesday . benaud 's memorial booklet , titled ' remembering richie ' , was attended by family and friends . benaud died last week of complications from skin cancer . australian prime minister tony abbott extended the offer to his wife daphne . """""""""
Target: legendary former australian cricketer and commentator richie benaud died on friday at the age of 84 of complications from skin cancer . his wife , daphne , declined a government offer for a state funeral . instead , there was a smaller service attended by family and close friends . in the memorial booklet at the funeral , benaud 's family described him as ' a special person who means so much to each of us in many different ways ' click here to watch 10 of richie benaud 's finest moments . "
Article: a private funeral service attended by ex-players shane warne and ian chappell was held on wednesday for former australia cricket captain and commentator richie benaud . benaud died last friday at the age of 84 of complications from skin cancer . australian prime minister tony abbott had extended the offer of a state funeral to benaut 's wife , daphne , which she declined . mr abbott said the offer was a mark of respect for a man who meant so much to millions of people in australia and around the world . richie benaud 's small private funeral was attended by close family and friends in sydney on wednesday . benaud 's memorial booklet , titled 'remembering richie ' , included a few words thanking the guests . daphne benaud -lrb- centre -rrb- , with family and friends , listens to speeches about her late husband richie . former australia captain and now cricket commentator ian chappell makes a speech during benaud 's wake . january 1952 : test debut against west indies at sydney cricket ground . january 1952 : first of 248 test wickets and 2,201 test runs . december 1958 : first test as australia captain , v england at brisbane . summer 1960 : first radio commentary for bbc . december 1963 : in his 60th test , the first to 2,000 test runs & 200 wickets . summer 1963 : first television commentary for bbc . february 1964 : final test against south africa at sydney cricket ground . september 2005 : final commentary in england after 42 years . but after he received a phone call from mrs benaud , he remarked' my understanding is that richie 's own wishes was for something very , very quiet and very , very private ' . instead , benaut 's family opted for a smaller service at the eastern suburbs memorial park attended only by family and close friends . benaud 's former teammate , brian booth , led the service . in the memorial booklet , his family described benaud as ' a special person who means so much to each of us in many different ways . ' benaud , considered one of the most influential cricket identities of the past century , played 63 tests for australia but was more well-known for his career in the commentating booth . following the funeral service , a memorial was held at the australian golf club attended by richie 's former cricket commentating colleagues , channel nine network executives and test cricketers past and present . benaud was loved by many cricket fans around the world for his distinctive voice and commentary style . benaud , pictured on his wedding day to wife daphne , and during his time as australia captain . tributes for benaud were placed at a statue outside the sydney cricket ground in australia . tributes were laid for the former broadcaster in sydney as the cricket world mourned his loss . members of the england and west indies cricket team observe a minute 's silence for benaud on monday ."

Table A.2: Document ID 3299 x.

Generated (ROUGE-1 F_1 36.8): a tornado touched down near belle glade in palm beach county , florida , on thursday afternoon . the weather may see storms in a number of states on saturday . the weather channel reported severe thunderstorms in several states . the weather may also affect parts of south dakota , nebraska , kansas , and texas . the weather service said that the detroit tigers and the new york yankees played in the snow
Target: bad weather may destroy weekend plans in multiple states , with thunderstorms and tornadoes forecast to strike . there was one tornado in florida and two tornadoes hit colorado thursday . severe thunderstorms may take place in several states friday and affect metropolitan areas , including dallas and houston . storms could also happen in the south and along the gulf coast saturday . winds , tornadoes , and hail were forecast to be potential issues from friday through sunday . the forecasts come after some parts of the country - including ohio , new york , minnesota , wisconsin and michigan - saw snow earlier this week . ’
Article: bad weather may destroy weekend plans in multiple states , with both thunderstorms and tornadoes forecast to strike . the news comes after a tornado touched down in florida on thursday afternoon . the tornado took place at about 2:45 pm near belle glade in palm beach county , but did not cause any damage , the weather channel reported . scroll down for video . rainy : thunderstorms were predicted to take place in many parts of the country on thursday night . storms : friday ’s weather may see storms in a number of major cities , including dallas and houston . inclement weather : severe storms may strike several southern states and along the gulf coast on saturday . havoc : a car was seen destroyed and tree branches were knocked down in miami after a thursday storm . let ’s play ball ! new york yankees ’ chase headley hits a two-run single as snow falls during the first inning of a baseball game against the detroit tigers on wednesday ... during the snow . snow pitch : david price of the detroit tigers played in the snow on wednesday . bundle up ! ian kinsler of the detroit tigers is seen during a wednesday game . colorado also saw two tornadoes on thursday , both of which took place shortly before 6pm , according to nbc news . on thursday evening , thunderstorms were predicted to take place in a number of states , including in parts of south dakota , nebraska , kansas , and texas , the weather channel reported . according to the news outlet , severe thunderstorms may take place in several states friday and affect metropolitan areas , namely wichita , kansas city , missouri , oklahoma city , along with dallas , houston , and san antonio . for saturday , the weather channel predicted severe thunderstorms would strike nashville , memphis , atlanta , tallahassee , jacksonville , new orleans , and jackson . for sunday , thunderstorms may occur in kansas , texas , and oklahoma , the channel said . winds , tornadoes , and hail were forecast to be potential issues from friday through sunday by the news outlet - with flash floods as an potential issue on monday . chilly weather : janice kennedy walks through a snow shower thursday , april 23 , 2015 , as she headed to work at the albion pharmacy in downtown albion , pennsylvania . springtime snow : don spann holds an umbrella over mary lou willis as she photographs the snowfall on the flower-covered trees in rolland moore park last thursday , april 16 , 2015 , in fort collins , colorado . do n’t freeze up ! snow falls on flowers shortly after the arrival of a spring storm , in boulder , colorado , thursday , april 16 , 2015 . dan kottlowski , a meteorologist for accuweather , told the website ‘it really will not be until the end of the week that severe weather really ramps up and a significant outbreak could occur as the main storm system moves out from the southwest . ’ he also said‘ the most likely area for the storms to fire first on friday is west of the i-35 corridor in texas , oklahoma and kansas . ’ the forecasts come after snow hit several states - including ohio , new york , minnesota , wisconsin and michigan - earlier this week , the washington post reported . according to wmtw , there snow was observed in maine and new hampshire on thursday . snowy weather was also photographed hitting pennsylvania this week , as well as wyoming and colorado last week , leaving residents bundled up . the washington post pointed out that the detroit tigers and the new york yankees played in the motor city during snowfall wednesday evening . the newspaper noted that more snow could strike cleveland , minneapolis , salt lake city , milwaukee , denver and detroit later this month . according to the national weather service , those five cities may experience showers this weekend . covered up : pedestrians cross a lower downtown street during a slow-moving spring storm thursday , april 16 , 2015 , in denver , colorado last week

Table A.3: Document ID 10854 🟩.

Generated (ROUGE-1 F_1 37.1): the eight-week-old puppies were discovered by a woman who opened the box . the box contained a cardboard box labeled ‘ stuff animals ’ the animals are suffering from heat exhaustion and have been neglected over an extended period of time . the good samaritan was driving by and saw the box tearing a hole into the top . she said she could feel the heat come out and the dogs had been there for a while . the box was sealed and
Target: the three eight-week old puppies were found inside the box on april 18 . a woman saw the box , left outside of a tennessee goodwill donation site , moving and approached it finding the dogs . the labradors were taken to mckamey animal center for treatment which said the animals have worms and are malnourished . it also said it is likely dogs have been neglected over an extended period of time due to their poor condition . since beginning treatment , they have been flourishing and will be put up for adoption once treatment is complete . ’
Article: a woman made a horrifying discovery when she opened a taped cardboard box labeled ‘stuff animals ’ only to find three puppies covered in urine and suffering from heat exhaustion . the eight-week-old labradors were found near a tennessee goodwill donation site on april 18 without any food or water inside of the box where temperatures were reportedly over 100 degrees . the puppies have worms and are malnourished , and it is likely they have lived their entire lives in a cage , according to mckamey animal center where the dogs are being treated . scroll down for video . three eight-week old puppies -lrb- above -rrb- were found inside of a sealed cardboard box on april 18 . the labradors were found by a woman who said the dogs were covered in urine and suffering from heat exhaustion when they were discovered in the box labeled‘ stuff animals ’ the animals have been named greta garbo , bette davis and marilyn monroe , and have been flourishing since they began treatment , according to abc . the good samaritan , who has not been identified , was driving by and saw the box moving when she approached it tearing a hole into the top finding the dogs . ‘she said when she opened the box , she could feel the heat come out . they had been there for a while , ’ chelsea fogal , veterinarian at mckamey animal center told wtvc .‘ it is frustrating when you have someone do something so neglectful and inhumane , ’ she added . ‘but i think these guys really lucked out that someone , a good samaritan , came along and found them . ’ i think the outcome could have been a lot worse . ’ the woman then took the puppies to the animal center for treatment . on facebook , mckamey wrote that the box was securely taped to prevent their escape . mckamey animal center in tennessee , where the dogs have been receiving treatment , said they have worms and are malnourished , and it is likely they have lived their entire lives in a cage . on facebook , the animal center said the puppies appear to have been neglected over an extended period of time due to their poor condition .‘ there was no air ventilation for the puppies and they were covered in urine and suffered from heat exhaustion – the temperature in the box was estimated to be well over 100 degrees , ’ the center wrote . ‘the box had been left at a goodwill drop off site near hwy 58 . the dogs were over heated and covered in urine from a long period of time being inside the box .‘ all three dogs are being treated at mckamey animal center . ‘it would appear they have been neglected over an extended period of time due to their poor condition . ’ once the puppies complete their treatment at the center , they will be put up for adoption . an animal service officer investigating the case said they have dozens of leads regarding the incident . the puppies will be put up for adoption once they finish being treated at the center . the mckamey is offering a $ 200 reward for information leading to the arrest and conviction of the person or people responsible for the incident . chattanooga goodwill industries is grateful for the quick thinking and reaction of the woman who noticed and rescued the three puppies that were left on the sidewalk near the vicinity of one of its donation center trailers , a representative of the company said in a statement to abc . a representative from chattanooga goodwill industries -lrb- file photo above -rrb- said they are grateful for the quick thinking and reaction of the good samaritan . the representative also said the attendant on duty was unaware of the situation and had not been approached or contacted by the woman discovered the box . an animal service officer investigating the case said she has a dozen of leads coming in from tennessee and georgia following the incident .‘ it happened in the middle of the day , so someone had to see something , ’ leslie stokes told abc . ‘ we are trying to get surveillance video from someone at goodwill to see if there were any cars that pulled in . ’"

Table A.4: Document ID 10938 ♦.