Automatic Tweet Summarization using Transformers

CS4099D Project
End Semester Report

Submitted by

Abid Ali Karuvally Pathikkal (B180466CS)

Under the Guidance of

Mr. T A Sumesh
Assistant Professor

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

May, 2022

NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
KERALA, INDIA - 673 601

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that this is a bonafide report of the project work titled

AUTOMATIC TWEET SUMMARIZATION USING TRANSFORMERS

done by

Abid Ali Karuvally Pathikkal

of Eighth Semester B. Tech, during the Winter Semester 2021-'22, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of the National Institute of Technology, Calicut.

(Mr. T A Sumesh)
(Assistant Professor)
Project Guide

04-05-2022
Date

DECLARATION

I hereby declare that the project titled, **Automatic Tweet Summarization using Transformers**, is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or any other institute of higher learning, except where due acknowledgement and reference has been made in the text.

Place : NIT Calicut	Name : Abid Ali Karuvally Pathikkal
Date : 04-05-2022	Roll. No. : B180466CS

Abstract

Social Media has become a major source for gaining information about current events happening around us. Twitter is one of the fastest and most popular online social media where thousands of people tweet everyday. Tweets talking about the same event are called an event cluster. Reading an entire event cluster is time consuming and is not practical these days. The role of summary is to give readers relevant information in the event cluster. Hence Automatic Tweet Summarization is a promising research topic and could be a handy tool in day to day life. Current models result in generating false, repetitive information. The purpose of this project is to understand different techniques in natural language processing and arrive at a better model to generate a summary of an event cluster.

ACKNOWLEDGEMENT

I would like to express my sincere and heartfelt gratitude to my guide Mr. T A Sumesh, who have guided me throughout the course of the final year project. Without his active guidance, help, cooperation and encouragement, I would not have made headway in the project. I would like to thank my parents and the faculty members for motivating me and being supportive throughout my work. I also take this opportunity to thank my friends who have cooperated with me throughout the course of the project.

Contents

1 Introduction **2**

2 Problem Statement **3**

3 Literature Survey **4**

4 Proposed Work **7**
- 4.1 Pre-processing 7
- 4.2 Pre-training 8
- 4.3 Fine-Tuning 9
- 4.4 Generating Summary 9

5 Experimental Results **10**

6 Conclusion **12**

References **13**

List of Figures

4.1 Design 8

List of Tables

5.1 ROUGE Metric Scores using Beam Search 11
5.2 ROUGE Metric Scores using top-p, top-k Sampling 11

Chapter 1

Introduction

Artificial intelligence (AI) is a broad field of computer science concerned with developing intelligent machines that can accomplish activities that would normally need human intelligence. NLP stands for natural language processing.AI discipline involved with providing machines the ability to comprehend text in the same manner as humans can. Automatic Tweet Summarization is one of the field's most difficult and intriguing problems. NLP is a type of natural language processing. It is the process of writing a brief and informative summary from a massive data text. There are two types of summarization. Generating a summary with the important phrases in the source text is called Extractive Summarization. Generating a summary by understanding the concept in the source text in a more natural language is called Abstractive Summarization. Twitter is one of the most widely used social media platforms. An event that is being discussed on Twitter usually has a lot of tweets about it. Reading every tweet will be a difficult task and hence these should be summarized.

Chapter 2

Problem Statement

Given a set of tweets on a particular event, we have to generate a summary for these tweets. That is, create a machine learning model to extract useful information contained in an event cluster(tweets talking about the same event). Deep Learning techniques, which are based on seq-2-seq models, suffer from issues such as hallucination of facts, copying long phrases from source text and repetition of phrases within the summary. State-of-the-art transformer language models like T5, BART can be used to generate coherent summaries.

Chapter 3

Literature Survey

[1] is the first paper on abstractive summarization using deep learning. They used a fully data-driven approach, local-attention for encoding, beam-search for decoding. Applied their model to the DUC-2004 Gigaword datasets and evaluated using ROUGE Perplexity as evaluation metrics. [2] used a hybrid pointer-generator network to generate novel words and copy words from source text in case of out-of-vocabulary. They also used coverage mechanisms to reduce repetition. Applied their model to the CNN/Daily Mail dataset and evaluated using ROUGE. [3] produced a novel neural intra-attention architecture in which they added an intra-temporal attention to bidirectional LSTM encoder and an intra-decoder attention to single LSTM decoder. A new training approach has been added that combines standard supervised word prediction with reinforcement learning to produce more readable summaries. They tested their model utilising ROUGE human evaluation on CNN/Daily Mail New York Times datasets.

In [4,] a single layer bidirectional encoder was used, as well as two single layer LSTM decoders - forward and backward. The attention mechanism was used in both the encoder backward decoder and the pointer mechanism in both decoders to address the out-of-vocabulary problem. Applied their model to the CNN/Daily Mail TTNews datasets and evaluated using ROUGE. [5] used bidirectional LSTM Encoder-Decoder architecture and bidirectional beam search to generate balanced summaries. Applied their model to the CNN/Daily Mail dataset and evaluated using ROUGE. [6] introduced a novel network architecture, Transformer, which is completely based on attention mechanisms. A multihead self-attention network and a position wise completely connected feed-forward network make up each layer of their encoder. The decoder uses the same architecture as the encoder, with the addition of a multi-head attention network for the output of the encoder stack. A positional encoding was introduced to extract the relative/absolute position of the tokens in the sequence. They used their model to evaluate the WMT 2014 English-to-German test using the BLEU score.

Introduced encode-encode-decode architecture in [7], in which they first encoded with a transformer then with the seq2seq model. They used GRU-RNN seq2seq model decoder. Applied their model to the CNN/Daily Mail Newsroom datasets and evaluated using ROUGE. [8] employed a decoder-only network with pre-trained decoders, in which the same Transformer LM encodes the source and outputs the summary. They used ROUGE to evaluate their model using the CNN/Daily Mail dataset. [9] presented BERT (Bidirectional Encoder Representations from Transformers), a novel language representation approach. They employed a bidirectional self-attention system and a multi-layered bidirectional transformer encoder. 'Masked language model' and 'next sentence prediction' were also utilised. On eleven natural language processing tasks, it achieves new state-of-the-art outcomes. [10] used Abstractive and Extractive summarization using pretrained BERT encoder and a 6-layered transformer decoder. Encoder and decoder optimizers are different to reduce mismatch between the two. They tested their model using ROUGE human evaluation on the CNN/Daily Mail and New York Times XSUM datasets. [11] proposed a new method for summarising a social media event automatically. The encoder was the BERT model, while the decoder was the Transformer architecture. Their model includes 'tweet selection model' then 'event topic predcition' and then encoding with 'pretrained BERT model'. Preprocessed tweets are fine-tuned with pre-trained GPT-2 in [12]. They used a combination of top-k and top-p sampling as a decoding strategy.

Chapter 4

Proposed Work

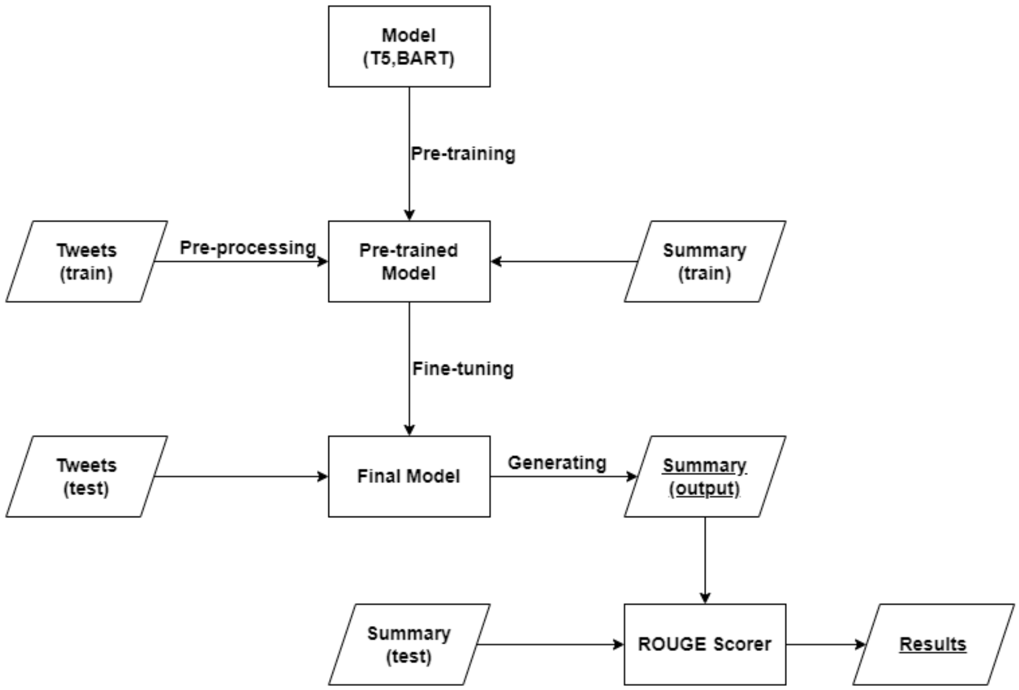
There are four steps involved in the tweet summarization task which include preprocessing of the given tweets, pre-training of a language model, fine-tuning of preproccessed tweets using the pre-trained language model, generating summary using the language model.

4.1 Pre-processing

A tweet may contain noise such as URLs, Mentions, Hashtags, Emojis, Smileys etc. URLs, Emojis do not convey any meaning to the sentence. Hence they are completely removed. Mentions, Hashtags may contain some important information. Hence they are kept by just removing ‘’ and ‘@’ symbols.

Tweets contain informal text and slang words. We could add a dictionary to replace the words with the actual words. But the dataset we are dealing with is of extractive type. Hence we need the sentence as it is for the model to learn the slangs and other informal texts.

Tokenization is the process of converting the sentences to a list of tokens of words. Every transformer architecture has its own tokenizer. Since we are using the T5 and BART Models, we will be using T5TokenizerFast and BARTTokenizer, which is basically a Sentencepiece tokenizer.



1

Figure 4.1: Design

4.2 Pre-training

Pre-training is when a model is trained with one task in order to assist it create parameters that can be utilised in other tasks. It is a model that learns every task such as question-answering, summarization, sentiment-analysis, etc. It is exposed to big datasets like wikipedia and mostly everything on the internet. These models are trained in such a way that they “learn” the grammatical structures and semantics of a language. The main advantage of pre-trained models is that they can be used for various tasks without training from scratch. It will reduce time for learning. It will also reduce the computational usage.

This step does not include our participation. We used pre-trained language model T5: Text-To-Text Transfer Transformer, which is pretrained on common crawl. The model is essentially an Encoder-Decoder Transformer. T5 tries to put all of the jobs that come after it into a text-to-text format. We also used the pre-trained language model BART: which is a combination of BERT(Bi-directional encoder) and GPT(left-to-right decoder).

4.3 Fine-Tuning

It is the process of re-training a pre-trained language model using our own custom data. As a result of the fine-tuning, the weights of the original model are updated to account for the characteristics of the domain data and the task we are interested in. We only need a small dataset while fine-tuning since the pre-trained model would already know syntax and semantics of the sentence.

This is one of the most major steps involved in understanding our downstream task of tweet summarization. In this process, our T5 and BART model will understand the source text and target text, that is how tweets are summarised.

4.4 Generating Summary

This is the final step in tweet summarization. That is generating summaries with our model. The main advantage here is that the T5 model is a text-to-text model that is its input and output is text. BART also has a generation model which helps in generating sentences. Beam search is usually used for summarization tasks. Top-p, Top-k sampling is also used. The output from the model is a set of tokens. We just need to decode with the same tokenizer and join the words to get the summary.

Chapter 5

Experimental Results

ROUGE is the evaluation metric we're utilising. Recall-Oriented Understudy for Gisting Evaluation is what it's called. It is used to assess automatic text summarization. It analyses generated and reference summaries to see how similar they are. The overlap of words is used to calculate it. ROUGE-N compares the overlap of unigrams, bigrams, trigrams, and higher order n-grams. Using LCS(Longest Common Subsequence), ROUGE-L determines the longest matching sequence of words .

Each ROUGE metric has recall, precision, and f-measure scores. The percentage of the total number of n-grams in the reference summary that coincide with the model generated summary is expressed as a percentage of the total number of n-grams in the reference summary. The precision score, on the other hand, represents the proportion of overlapping n-grams in the reference and model generated summaries compared to the total number of n-grams in the model generated summaries. The recall and precision scores are used to construct the F-measure.

Table 5.1: ROUGE Metric Scores using Beam Search

Transformer	**Metric**	***Precision***	***Recall***	***F-Measure***
T5	Rouge - 1	0.561	0.700	0.598
	Rouge - 2	0.485	0.593	0.514
	Rouge - L	0.518	0.639	0.550
BART	Rouge - 1	0.589	0.696	0.618
	Rouge - 2	0.516	0.605	0.540
	Rouge - L	0.544	0.639	0.570

Table 5.2: ROUGE Metric Scores using top-p, top-k Sampling

Transformer	**Metric**	***Precision***	***Recall***	***F-Measure***
T5	Rouge - 1	0.564	0.709	0.604
	Rouge - 2	0.488	0.603	0.520
	Rouge - L	0.515	0.639	0.548
BART	Rouge - 1	0.601	0.710	0.630
	Rouge - 2	0.528	0.620	0.552
	Rouge - L	0.558	0.654	0.584

Chapter 6

Conclusion

Deep learning techniques can be used to handle a range of challenges in natural language processing. Summarization is one such task. Transformers are the future in the field of AI. T5 and BART are two of the best transformer models which achieves state of the art results for various tasks. We were able to implement T5 and BART models which can generate tweet summaries. Both were able to generate coherent summaries. While BART showed better results with the existing dataset. In literature, state-of-the-art results were obtained around 0.70 for ROUGE-1, 0.51 for ROUGE-2 and 0.66 for ROUGE-L. Our model were able to achieve comparable results using BART with top-p, top-k sampling.

References

[1] Sumit Chopra Alexander M. Rush and Jason Weston. A neural attention model for abstractive sentence summarization.

[2] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks.

[3] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization.

[4] Ruijia Wang Ding Xiao Xin Wan, Chen Li and Chuan Shi. Abstractive document summarization via bidirectional decoder.

[5] Kamal Al-Sabahi, Zhang Zuping, and Yang Kang. Bidirectional attentional encoder-decoder model and bidirectional beam search for abstractive summarization.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. Attention is all you need.

[7] Elozino Egonmwan and Yllias Chali. Transformer-based model for single documents neural summarization.

[8] Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Łukasz Kaiser. Sample efficient text summarization using a single pre-trained transformer.

[9] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding.

[10] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders.

[11] Quanzhi Li and Qiong Zhang. Abstractive event summarization on twitter.

[12] Shubham Keshao Bhokhade. Tweet Summarization Using Generative Pretrained Transformer 2.

[13] Huy-Tien Nguyen Minh-Tien Nguyen, Dac Viet Lai and Minh-Le Nguyen. Tsix: A human-involved-creation dataset for tweet summarization. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France, may 2018.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.