title

Daniel Gruhn, Long Truong, Deon Seyfi, Kalin Zaluzec

Abstract

For this project, we focused on different methods of title generation and summary generation. There were three main approaches we used for summarization. These approaches are extractive summarization, the T5 transformer, and the BART model. For title generation, we only used the latter two of these approaches. When training and testing our model, we tested different hyperparameters for each of these approaches as well as combinations to see which configuration led to the best results. Our goal was to refine title generation by using different combinations of the above methods along with different sets of hyperparameters.

1 Introduction

This project explores different techniques for title generation. The specific technique that we will focus on is title generation based on summaries of a document rather than generating titles directly from a document. We plan to use varying approaches to summary generation and compare the resulting titles that we generate using these summaries as input. We also plan to use more direct title generation methods that do not involve using summaries as a basis for comparison for our results.

Titles can be seen as the most important part of the document; it is the part seen by the most people.(Tullu, 2019) There have been many studies on how different characteristics in titles cause it to be viewed or cited more often. For example, short titles are more appealing to longer titles which could be seen as complex and boring.(Paiva et al., 2012) Unless the title can grab the reader's attention, it does not matter what the contents of it are. The most important things that titles need to do are convey the main topics, highlight the importance, be concise, and be attractive to readers.

Each of these four aspects contribute to the creation of successful titles in unique ways. Conveying the main topic is the most important aspect for catching the attention of the target audience.(Schmieder, 2018) Highlighting the importance lets the audience know the difference between one article and another on a similar topic. Being concise ensures that every word in the title is doing its job by drawing the reader in rather than pushing them away by being drawn-out and boring. Finally, being attractive to readers is the basic function of the title.

Summarization on its own can fulfill two of these aspects. Conveying main topics is guaranteed by the definition of a summary and highlighting the importance of a document is likely accomplished by adequate summaries.(Sethi et al., 2016) It also directly helps with the creation of a concise title by limiting the amount of text to work with to the most important parts. A standard summary, however, has too many words for a title. Since extractive summarization takes direct quotes from the text, it is very likely to have unnecessary text mixed in with the important ones. Attracting readers can be seen as the greatest variable factor as it can be highly opinionated.

There are also many types of titles that change based on the type of writing. Writing in literature has four variants of titles-clear, distinguished, symbolic and distinguished. Meanwhile, academic writing only has three types- declarative, descriptive and interrogative.(Schmieder, 2018) Declarative titles are good because they get straight to the point and have a very clear final result presented. This makes it very easy to cite by letting the audience know what kind of information it leads to without too much of a time investment. Descriptive titles reference the methods used rather than the results directly. Interrogative titles ask the question that gets answered by the research. These questions can very easily draw attention, but since they have no details about how the research was conducted or the results, it is more difficult to deliberately search for using keywords.

2 Related Works

3 Methods

When we approached the problem of summarization (and by extension, title-generation), we focused on two different methods. These methods are extractive summarization and abstractive summarization. Extractive summarization takes parts of the text that are deemed important word-for-word and uses them to construct a summary.(Sciforce, 2019) What changes in extractive summarization models is deciding which words are important enough to be kept and which are to be thrown out. Abstractive summarization attempts to find the general meaning of a given text and broadly explain the main idea while leaving out less crucial details. Due to the generation of new text as a summary, information may not necessarily be the same as the original document.

3.1 Weight-Based Extractive Summarization

The first thing we did when approaching this project was to code something to apply extractive summarization to text. We kept this to be sentence-based, so our extractive summarization methods were unable to create summaries shorter than a single sentence. With this in mind, we only used this procedure as a method of text truncation to generate our titles rather than as a stand-alone method for generating titles.

Our extractive summarization method involves essentially counting words. For each document, we would count the number of occurrences for each word in the document. This is done to determine which words are the most important. In order to only have the meaningful words being marked as important, stop words (including words such as ’the’, ’and’, ’then’, etc.) were omitted from this process. Then, for each sentence, we could calculate the weight of that sentence based on the total counts for each word in the sentence. The sentences were then sorted by weight, and the sentences with the highest weight were kept while the rest were discarded. This process (in theory) leads to only the most crucial sentences to the text remaining.

3.2 The T5 Transformer

The ‘T5’ in T5 transformer stands for ‘Text-to-Text Transfer Transformer. While the focus of this report is document summarization, the T5 model was trained to do a variety of other tasks including question answering, classification, and translation. The purpose of this is to allow the same model, loss function, and hyperparameters to be used across every task the model is designed for. (Raffel et al., 2020a) To differentiate different tasks from each other, a prefix is required on inputs. (Raffel et al., 2020b) For example, the prefix ‘translate English to German: ’ would be an appropriate prefix. For our purposes, we used the prefix ‘summarize: ‘ to indicate that we desire the T5 transformer to generate a summary for us.

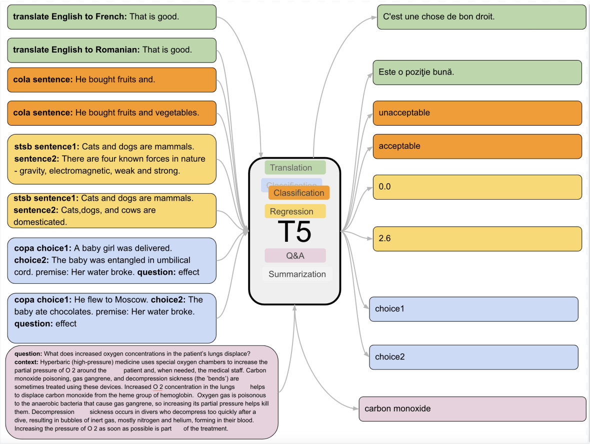
translate English to French: That is good.
translate English to Romanian: That is good.
cola sentence: He bought fruits and.
cola sentence: He bought fruits and vegetables.
stsb sentence1: Cats and dogs are mammals. sentence2: There are four known forces in nature - gravity, electromagnetic, weak and strong.
stsb sentence1: Cats and dogs are mammals. sentence2: Cats,dogs, and cows are domesticated.
copa choice1: A baby girl was delivered. choice2: The baby was entangled in umbilical cord. premise: Her water broke. question: effect
copa choice1: He flew to Moscow. choice2: The baby ate chocolates. premise: Her water broke. question: effect
question: What does increased oxygen concentrations in the patient’s lungs displace? context: Hyperbaric (high-pressure) medicine uses special oxygen chambers to increase the partial pressure of O 2 around the patient and, when needed, the medical staff. Carbon monoxide poisoning, gas gangrene, and decompression sickness (the ’bends’) are sometimes treated using these devices. Increased O 2 concentration in the lungs helps to displace carbon monoxide from the heme group of hemoglobin. Oxygen gas is poisonous to the anaerobic bacteria that cause gas gangrene, so increasing its partial pressure helps kill them. Decompression sickness occurs in divers who decompress too quickly after a dive, resulting in bubbles of inert gas, mostly nitrogen and helium, forming in their blood. Increasing the pressure of O 2 as soon as possible is part of the treatment.
Translation
Classification
Regression
T5
Q&A
Summarization
C'est une chose de bon droit.
Este o poziţie bună.
unacceptable
acceptable
0.0
2.6
choice1
choice2
carbon monoxide

Figure 1: Example inputs and outputs of the T5 transformer (Rajasekharan, 2019)

The T5 model was trained and tested on the ‘Colossal Clean Crawled Corpus’ (C4). The cleaned (duplication removal, bad word removal, etc.) C4 dataset consists of about 365 million English text entries scrapped from around the internet. (Raffel et al., 2020a) One interesting fact to note is that the common crawl regularly produces about 20 terabytes of scraped text per month. The majority of this scraped text cannot be used for the dataset; it consists mostly of non-useable gibberish, text from menus, or error messages. Other forms of text that are not useful are also found in this such as offensive text or source code text. The C4 dataset was cleaned according to the following criteria: (Raffel et al., 2020a)

1. Lines without a terminal punctuation mark were discarded
2. Pages with fewer than 5 sentences were discarded
3. Lines with fewer than 3 words were discarded

4. Any page that contains a word from the ‘List of Dirty, Naughty, Obscene, or Otherwise Bad Words’ was discarded (Emeric, 2020)

5. Any line that contains the word ‘javascript’ was discarded (to remove javascript errors)

6. Any page with the placeholder phrase ‘lorem ipsum’ was discarded

7. Any pages with a certain configuration of curly braces (‘{‘ or ‘}’) were discarded to remove code from the dataset

8. Any span of three sentences that occurred more than once were discarded

The T5 pretrained model comes with 5 different variants: t5-small, t5-base, t5-large, t5-3b, t5-11b. The t5-small model has 6 attention modules and the t5-base has 12 attention modules. All other models have 24 attention modules. Additionally, these models have 60 million, 220 million, 770 million, 3 billion, and 11 billion parameters respectively. (Raffel et al., 2020a) Each t5 model was evaluated using 34 different metrics (such as GLUE average). In multiple categories, the t5-11b model outperformed the previous best prior to the creation of the t5 model. In all cases, the t5-11b model outperformed the other variants of the t5 model.

3.3 The BART model

3.4 Training and Evaluation

When training our model, we decided to use a set of 10,000 documents and to conduct a grid search to find out which set of hyperparameters performed the best. Our grid search initially consisted of the following different hyperparameters:

Model Type	’t5-small’, ’BART’
Batch Size	100, 500, 1000
Truncation	100, 500, no truncation
Learning Rate	0.01, 0.001, 0.0001

This would have resulted in 54 different models being trained, which we felt would have been too time-consuming for us to do considering the impending deadline. With that in mind, we removed the learning rates from the set of hyperparameters. In the end, we trained 18 different models in our grid search.

Most of our hyperparameters are fairly straightforward, but our truncation size hyperparameter is perhaps the most important one. The truncation size will apply extractive text summarization to reduce the size of the text to either under 100 words, under 500 words, or it won’t reduce the size at all. In the case where we do no apply extractive summarization, this doesn’t mean that our text won’t be truncated. It means that the tensors will be truncated according to the max input size (512 for T5, 1024 for BART) that can be accepted by the model.

In order to evaluate our models, we had no objective way to do this. Instead, we had all four of our team members evaluate titles. For each given document, we chose the hyperparameter set that resulted in the title that we considered the best. The following are the top three performing sets of hyperparameters pertaining to the t5-small model. We normalized our scores to be between 0 and 100 so that they are easier to interpret.

Batch Size: 100 Truncation: 100	100
Batch Size: 1000 Truncation: 100	29.6
Batch Size: 1000 Truncation: 500	24.1

4 Results

For our results, we decided to evaluate our results in two different sets. These two sets include all results produced by T5 and all results produced by BART.

4.1 T5 results

Using the T5 transformer, we had a degree of success in our title generation, but there is still room for improvement. Our fine-tuned T5-small model generated quite a few good titles. One such example is ’How the Field of Infectious Diseases Can Leverage Digital Strategy and Social Media Use During a Pandemic.’ While we all agreed that this is a good title and that it matched the content of the paper, it is somewhat of an interesting result. The original title was ’Journal Pre-proof COVID-19 anti-vaccine movement and mental health: Challenges and A way forward Ramdas Ransing Title: COVID-19 anti-vaccine movement and mental health: Challenges and A way forward.’ The original title doesn’t talk about social media, yet our model shifted the focus from the anti-vaccine movement to talk about social media.

We also had a set of bizarre results from the T5-transformer. Two titles generated from separate documents were 'Innovative liver research continues during the current pandemic' and 'Innovative liver research continues during the early COVID19 pandemic.' Given how similar these titles are, we might expect the input documents to be nearly the same. This was not the case, however. The original titles for these documents were 'Multi-Task Driven Explainable Diagnosis of COVID-19 using Chest X-ray Images' and 'Supporting Information Quantification of mRNA Expression Using Single-Molecule Nanopore Sensing' respectively. We weren't able to precisely determine why this happened, but we have had a few theories. The most likely things that we feel are to be the causes are using the t5-small model, having a data set that is too small, model overfitting, and having a data set that is not diverse enough.

4.2 BART results

5 Conclusion and Further Research

Our work on this project has produced mixed results. We are satisfied with what titles we were able to generate using the methods we have chosen for this project, but there is much room for improvement and we are confident that if given more time, the quality of the titles that we generate could be improved significantly.

If we had more time to train our models, the most immediate improvement would be to increase the size of the data set that we used. The decision to use a data set of size 10,000 was only made to allow training to finish on time; if we had unlimited time, we would have used 100,000 documents or more. We possibly would even use the entire data set, which consisted of over 300,000 documents. Secondly, we would have liked to test more hyper-parameters in our grid search. We wanted to test more different batch sizes as well as text truncation lengths. Additionally, we could have tested more variants of the T5 transformer other than T5-small. We also would have liked to test different learning rates, different optimizers, and different loss functions. Finally, we would have liked to train on other data sets to compare results to our covid-19 data set to determine if the (lack of) diversity within our data set affected our results.

The second thing that we would have liked to improve is our means of evaluating our results. For this project we only had our four team members evaluate titles. As much as we tried to be unbiased, it would be dishonest to say that we were completely unbiased when evaluating our results. If we wanted to get a better evaluation result, we would like to have people unrelated to this project evaluate titles, and have far more than only four people evaluating titles.

Overall we feel that we produced satisfactory results for this project. If we were to continue working on the project and the above changes were made, we are confident that we could significantly improve the quality of the titles that we are able to generate.

References

Jacob Emeric. 2020. List-of-dirty-naughty-obscene-and-otherwise-bad-words.

Carlos Eduardo Paiva, JoÃPaulo da Silveira Nogueira Lima, and Bianca Sakamoto Ribeiro Paiva. 2012. Articles with short titles describing the results are cited more often. *Clinics*, 67:509 – 513.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. T5.

Ajit Rajasekharan. 2019. T5 — a model that explores the limits of transfer learning.

Eric Schmieder. 2018. How to write an engaging title for your academic journal article.

Sciforce. 2019. Towards automatic text summarization: Extractive methods.

Nandini Sethi, Prateek Agrawal, Vishu Madaan, Sanjay Singh, and A. Kumar. 2016. Automated title generation in english language using nlp. 9:5159–5168.

Milind S. Tullu. 2019. Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key. *Saudi journal of anaesthesia*.