Machine Learning for Topic News Title Generation

Fengyu Zhang SI630 WN2024

1 Project Goal

In the era of digital information, the volume of news content available to readers has grown exponentially, making it increasingly challenging for individuals to stay informed without becoming overwhelmed. The project's primary goal is to leverage machine learning (ML) techniques for the effective summarization of news articles and generate ideal news title, aiming to improve the efficiency, accuracy, and readability of these summaries, allowing readers to grasp the essence of news stories without dedicating extensive time to reading full articles.

The stakeholders of this project can be individual readers, news organizations, educational sectors, and potentially government bodies reliant on swift and accurate information dissemination. Improved news title generation models can transform media consumption by providing accessible, succinct title generation method of complex news stories, thereby enhancing public knowledge and engagement. Additionally, in broader view, enhanced news title generation techniques could pave the way for similar advancements in summarizing other forms of text, such as academic literature, legal documents, and social media feeds

2 NLP Task Definition

The system of this project takes text of news article as input, and the output aims to produce a brief summary that captures the main essence of the article. However, it should be noticed that this news title generation system will not play the role as chatbot like chat-GPT. Our main concern is still how to generate the exact content rather than the form of the conversation.

3 Data

The dataset: A Multilabeled News Dataset for News Articles Hierarchical Classification (Petukhova and Fachada, 2023), is a collection of 10,917 news articles with hierarchical news categories collected between January 1st 2019, and December 31st 2019 classified by using NewsCodes Media Topic taxonomy. The author has manually labelled the articles based on a hierarchical taxonomy with 17 first-level and 109 second-level categories. This dataset was initially created for topic classification, but its detailed content and title, accompanied by specific topic can also be applied for extracting information from news articles.

For example, column values for one instance of news are stored as:

- **Title**: Virginia mom charged with murder in 2-year-old son's death
- **Content**: The Virginia woman whose 2-year-old son was found in a trash incinerator has been charged with murder in his death. Hampton Commonwealth's Attorney Anton Bell told a news conference on Thursday that a warrant had been issued for 34-year-old Julia Leanna Tomlin, who also will be charged with unlawfully disposing of Noah Tomlin's body. Bell said skull fractures found on the toddler's body indicates a level of force so severe it was as if the child had fallen several stories from a building. Julia Tomlin reported her son missing in June, and searchers sifted through a landfill and steam plant over 10 days before the body was found on July 3. Tomlin is already jailed and charged with three counts of felony child neglect prior to reporting Noah missing.
- **Category Level1**: crime, law and justice
- **Category Level2**: crime

The distribution of news length in the dataset is given as:

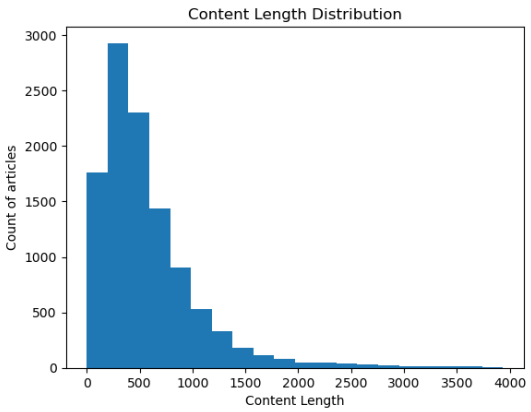

Figure 1: Corpus Length Distribution

4 Related Work

Text summarization is one of the most important applications in the field of nlp. Many studies have proposed different models and algorithms based on it, which helps me better understand potential challenges and existing methods for my current work.

In Performance of Optimizers in Text Summarization for News Articles (Kumari et al., 2023), the article mentions that extractive-based text summarization has reached its peak and as a result, its scope has shrunk. As a result, the paper compares and assesses the effectiveness of two optimizers: adam and rmsprop. The paper compares their performance on various datasets as they are widely employed in text summarization.

The second article DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization (Joshi et al., 2023), they propose DeepSumm, a novel method based on topic modeling and word embeddings for the extractive summarization of single documents. It is mentioned that recent summarization methods based on sequence networks fail to capture the long range semantics of the document which are encapsulated in the topic vectors of the document. In DeepSumm, authors' aim is to utilize the latent information in the document estimated via topic vectors and sequence networks to improve the quality and accuracy of the summarized text.

The third article Applying Transformer-Based Text Summarization for Keyphrase Generation (Glazkova, 2023) propose several evaluation methods for generative text, including ROUGE-1, F1-score and BERTScore. In information extraction, the article also emphasize the significance of keyphrase generation. Keyphrases contain a brief representation of the contents of a text. They help search engines find and systematize papers. A qualitative selection of keyphrases positively affects a paper's visibility and its number of citations.

Based on the methods above, my current work will try to explore topic and transformer-based network for text-generation. Additionally, I will also compare the machine generated title or summary with manually generated text to

5 Methodology

5.1 Bert Summarizer Baseline:

First, as naive baseline, we leverage bert extractive summarizer (Devlin et al., 2018) library on hugging face to directly summarize the content of news, and generate one sentence output as news title without fine tuning on the train set.

The BERT model is modified to generate sentence embeddings for multiple sentences. This is done by inserting [CLS] token before the start of the first sentence. The output is then a sentence vector for each sentence. The sentence vectors are then passed through multiple layers that make it easy to capture document level features. The final summary prediction is compared to ground truth and the loss is used to train both the summarization layers and the BERT model.

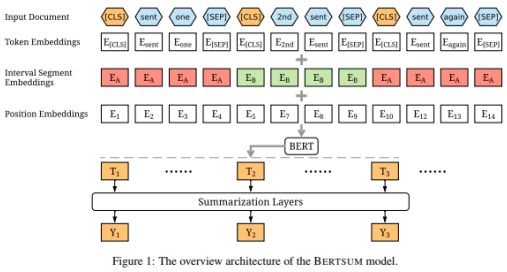

Figure 1: The overview architecture of the BERTSUM model.

Figure 2: Bert Summarizer

5.2 Fine Tune on T5-Base Model

As exploration, we plan to introduce the T5 base model (Raffel et al., 2020) and fine tune the model based on the train set and dev set with size of (8000, 1000), and finally apply our fine tuned model to generate the title for news in test set.

The T5 model, short for "Text-to-Text Transfer Transformer," is a versatile and powerful machine learning model developed by Google Research. Introduced in a paper titled "Exploring the Limits of Transfer Learning with a Unified Text-to-Text

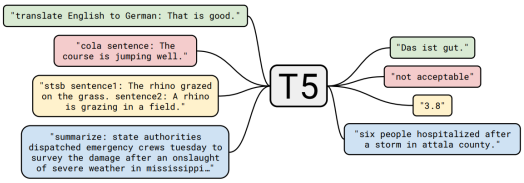

Figure 3: T5 Model

Transformer," T5 is designed to handle a wide range of natural language processing (NLP) tasks using a unified approach. Unlike traditional models that are specifically designed for tasks like translation, summarization, or question answering, T5 treats every NLP problem as a "text-to-text" problem. This means that for any task, the input is text and the model is trained to generate text as output.

Main library:

- AutoTokenizer, T5ForConditionalGeneration, Seq2SeqTrainingArguments, Seq2SeqTrainer

Key parameters should be mentioned:

- content: tokenizer (max length=512)
- title: tokenizer (max length = 32)
- number of training epochs = 3
- learning rate = 2e-5
- per device train batch size = 8
- per device eval batch size = 8
- predict with generate=True
- metric for best model = rouge2
- EarlyStoppingCallback(5)

5.3 GPT3.5 Turbo

Finally, we would like to compare our fine tuned t5 news title generation model with GPT3.5 Turbo. GPT-3.5 Turbo is a language model developed by OpenAI that is part of the GPT (Generative Pre-trained Transformer) family. This model, released after GPT-3, is designed to offer similar capabilities with improvements in speed, efficiency, and sometimes accuracy for specific types of tasks.

In order to generate news title, we will utilize openai api key to get access to gpt3.5 turbo. Then, we will use zero-shot prompting to generate the result:

- Task: Create a title for the following content
- News: Content

6 Evaluation and Results

- **Rouge: Recall Oriented Understudy for Gisting Evaluation**

  Rouge, as mentioned earlier, is another widely reported metric. It is a very common practice to report Rouge along with BLEU scores for standard tasks. It is very similar to the BLEU definition, the difference being that Rouge is recall focused whereas BLEU was precision focused. Standard implementations of these can be found in most ML libraries, n-rouge is most commonly used.

- **Bert-Score**

  BERTScore is an automatic evaluation metric used for testing the goodness of text generation systems. Unlike existing popular methods that compute token level syntactical similarity, BERTScore focuses on computing semantic similarity between tokens of reference and hypothesis.

6.1 Baseline Results: bert extractive summarizer

Metric	F1-score
ROUGE-1	0.2044
ROUGE-2	0.0732
ROUGE-L	0.1794

Table 1: ROUGE Scores

Metric	Score
Precision	0.8582
Recall	0.8844
F1 Score	0.8708

Table 2: BERT Scores

In nutshell, the ROUGE scores suggest that there's significant room for improvement in capturing the exact wording and detailed structure of the reference summaries. In contrast, the high BERTScore indicates that the summary does well in capturing the relevant semantic meaning in terms of referential titles, even if the exact words or phrases aren't always used. Thus, this means the summary is generally effective in conveying the gist of the reference material but doesn't fit our demand for news title.

6.2 T5 Based Fine Tuned Model

Metric	F1-score
ROUGE-1	0.3812
ROUGE-2	0.1745
ROUGE-L	0.3327

Table 3: ROUGE Scores

Metric	Score
Precision	0.8933
Recall	0.8835
F1 Score	0.8882

Table 4: BERT Scores

The T5-based model demonstrates a good capability in generating coherent and contextually relevant news titles as evidenced by the high BERT scores, which suggest a deep semantic understanding and effective synthesis of titles. Moreover, it evidently improves almost two times of rouge1 score and rouge-L socres, as well as two point five times of rouge 2 scores compared to bert summarizer, which demonstrate stronger ability to generate more coherent and structured news title.

6.3 GPT3.5 Turbo

Metric	F1-score
ROUGE-1	0.3274
ROUGE-2	0.1248
ROUGE-L	0.2732

Table 5: ROUGE Scores

Metric	Score
Precision	0.8777
Recall	0.8820
F1 Score	0.8796

Table 6: BERT Scores

It is interesting to find that our fine tuned T5 based model has surpass GPT 3.5 turbo model on the task of generating news tiltle on the corpus of MN-DS-news-classification.

7 Discussion and Conclusion

In order to more intuitively access the differences in the results produced by different models, we provide the following example

- **News Content ID98978**: "Capitalism vs. socialism? For Melinda Gates, the choice is simple. ""What I know to be true is I would far rather live in a capitalistic society than a socialist society,"" Gates said in an interview with CNBC's Becky Quick that aired on ""Squawk Box"" on Wednesday. ""I think when we stop and think of what we have from a capitalistic society, we have to remember what we actually have."" Gates' comments come as the American political system is embroiled in a debate about socialism and capitalism. Several Democratic lawmakers and presidential candidates have called for sweeping changes to a system they say is responsible for growing inequality and division within the country. President Donald Trump has accused them of embracing socialism, which he says would lead to economic ruin in the U.S. Without mentioning the political debate, Gates, who co-chairs the Bill and Melinda Gates Foundation along with her husband, Microsoft co-founder Bill Gates, defended the U.S. system.....

- **Original Title**: Melinda Gates: Capitalism needs work, but it beats socialism and the US is 'lucky' to have it.

- **Bert Generated Title**: For Melinda Gates, the choice is simple.

- **T5 Generated Title**: Melinda Gates: Capitalism vs. Socialism?

- **GPT3.5 Turbo Generated Title**: Melinda Gates Makes a Case for Capitalism Over Socialism: Addressing Inequality and Embracing Change.

According to our comparison, T5 based tuned model and GPT3.5 turbo both achieve the ability to generate the news title based on the given content, and T5 slightly surpass gpt3.5 turbo with respect to the rouge score. However, for bert summarizer, the result accords with our expectation. It is because that summarization and title generaton, as a matter of fact, is a little bit different. Bert summarizer is extractive, but title generation often requires the model to have abstractive ability to perform the tasks, which explains why the bert summarizer oftern chooses the first sentence of content as the title.

Moreover, compared to long content summarization or text generation, the length of title is much shorter, which could explain why the overall rouge scores are all below 0.4 while the bert score achieves much higher. It is easier for LLM to capture and reflect the semantic meaning for title generation than structure and tokens.

8 Other Things We Tried

- Fine tune on mamba: We initially want to fine tune one mamba instead of using GPT turbo as final competitve model. However, after hours of trying, I find it is hard to understand the detailed structure of mamba, and fine tune on it requires extra knowledge such as peft_configureation and sft trainer.

- Beside t5 base, the project also trained and fine tuned on t5 small, and result shows that, with the increase of parameters, the rouge scores rise up round 0.02 but requires 2 times of training.

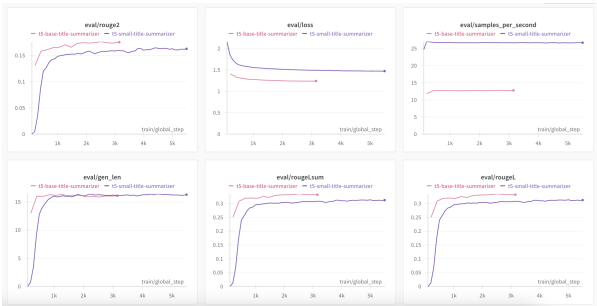

Figure 4: T5 small vs. T5 base

9 What You Would Have Done Differently or Next

- Since we have already have a taste of fine tuning and prompting on popular LLM, it is worthwhile to approach the project from another aspects, i.e., build from scratch of language model like bert and gpt to help us achieve the goal title generation and summarization.

- Moreover, if I have enough time in the future, it is also worthwhile to explore other LLm such as Llama, mamba and learn techniques to fine tune on personal PC or GPU.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Morozov D.A Glazkova, A.V. 2023. Applying transformer-based text summarization for keyphrase generation. *Lobachevskii J Math*, 44:123–136.

Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, and Laura Fern°ndez-Robles. 2023. Deepsumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. *Expert Systems with Applications*, 211:118442.

Namrata Kumari, Nikhil Sharma, and Pradeep Singh. 2023. Performance of optimizers in text summarization for news articles. *Procedia Computer Science*, 218:2430–2437. International Conference on Machine Learning and Data Engineering.

Alina Petukhova and Nuno Fachada. 2023. Mn-ds: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.