,,

Universiteit Leiden
The Netherlands

Opleiding Informatica

Evaluating the zero-shot performance of three transformer models in Reddit Posts

Jingwen Liu

Supervisors:
Suzan Verberne
Zhaochun Ren

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 23/08/2023

Abstract

Nowadays, we are surrounded by large amounts of information. It takes a lot of time to read and to get the main information from plenty of contents. In this situation, automatic summarization methods are useful. They could help people saving time and getting the key points quickly. Extractive summarization and abstractive summarization are the two ways to generate summaries. In this thesis, two representative models of the two ways are selected, which are the BART-large-cnn and PEGASUS-xsum models. For the model BART-large-cnn, it is designed as an abstractive model, but the result obtained by the model has an extractive form. Also, ChatGPT, being a recent hot product, plays well with summarization and it will be used for the same summarization tasks in this thesis. In this thesis, I investigate which of the three models is better for making summarization of reddit posts. I use the ROUGE metrics to measure the quality of the summarization model. However, the limit of ROUGE metric is that the score depends only on the common words, it may cause the content of the summary to be less in line with human's ideological cognition, though this summary can be more representative of the original content. For future research, more other metrics should be used to evaluate the quality of the models.

Contents

1 Introduction **4**
1.1 Thesis overview 5

2 Background **5**
2.1 Summarization techniques 5
2.1.1 Extractive summarization 5
2.1.2 Abstractive summarization 6
2.2 Transformer 6
2.3 Transformer models for summarization 7
2.3.1 BART 7
2.3.2 PEGASUS 8
2.4 ChatGPT 8
2.5 ROUGE metrics 9

3 Methods **9**
3.1 Dataset 10
3.2 Dataset preparation 11
3.3 Zero-shot use of BART and PEGASUS 11
3.4 Prompting ChatGPT 11
3.5 Evaluation 12

4 Results **12**
4.1 BART-large-cnn model 12
4.2 PEGASUS-xsum model 13
4.3 ChatGPT 14

5 Discussions **15**

6 Conclusion **16**

References **17**

A Some examples of the summarization **18**

1 Introduction

Now that we are in the information age, lots of information is flooded in front of our eyes, including but not limited to news and social media. People usually spend a lot of time to understand plenty of information. In order to save time and help people understand the key information faster, it is particularly important to make a concise summary. Correspondingly, natural language processing is also developing rapidly in recent years. Automatic summarization methods are often based on sequence-to-sequence models such as recurrent neural networks and, more recently, transformers. In this thesis, models based on transformer will be discussed.

The idea of the research question comes from the thesis, Automated News Summarization Using Transformers [1]. In the thesis, four different models, including extractive summarization and abstractive summarization, are selected to make summarization tasks for the BBC news dataset. At the end of the thesis, the ROUGE metric is used to calculate the scores of each model, combined with the quality analysis of the summary text to judge which one performs best. The two different summarization ways are not analyzed separately.

With regard to extractive and abstractive summarization, the difference is as follows. Extractive summarization refers to the extraction of the most critical phrases or sentences from the original text to compose a summary. In contrast, abstractive summarization uses phrases that are not the same as those in the original text to form a summary that has the same meaning as the original.

In this thesis we will address the following research question:

Which summarization model gives the best summaries of Reddit posts in a zero-shot setting?

Because the two models chosen come from two different summarination ways, as discussed above, the result of which can give us a glimpse of the differences and advantages and disadvantages of these two methods. The Reddit posts which are used are from the Reddit dataset [2] of the Huggingface platform, and they contain the users' posts and the users' summaries of the posts. Each post and the corresponding summary is treated as an input set for training. Because the content of posts has a strong personal subjectivity, when we use the summarizaton model to generate summaries, usually no matter what type of summarization method it is, there will be a certain degree of difference in content or subjective ideas. Therefore, this thesis uses the two selected models for fine-tuning, and then scores the generated summaries through ROUGE to compare which model is better for summarizing posts to some extent.

First, we select two pre-trained models from Huggingface platform to perform the study of the research problem. We choose this platform because it is a large community for sharing NLP models based on Transformers. Especially the open source natural language processing library on github, the pre-trained model library Transformers, has been downloaded more than one million times and has more than 24,000 stars on github. Two models are also chosen from two summarization ways, BART-large-cnn [3] for extractive summarization (Although BART itself is an abstractive summarization model, when using BART-large-cnn, a pre-trained model for summarization, we find that the results are extractive, so here we treat it as an abstractive summarization model with extractive summarization characteristics.) and PEGASUS-xsum [4] for abstractive summarization.

We first use the two original pre-trained models to get zero-shot summaries of the Reddit dataset [2]. For zero-shot summary, it is just outputed from the original pre-trained model without fine-tuning. Then we use the ChatGPT model to get another zero-shot summaries from the dataset. After also getting the summaries generated by the ChatGPT model, we use the ROUGE metric to evaluate the results. Through the obtained ROUGE scores, we can evaluate the quality of the selected models. Limitations of this method are also discussed at the end, as a basis for possible further research in the future.

1.1 Thesis overview

This chapter contains the introduction; Section 2 includes the background; Section 3 discusses the methods used in this thesis; Section 4 describes the experiments and their outcome; Section 6 concludes the conclusion and discussion for further research.

2 Background

Notably, document summarization has been a longstanding pursuit in Natural Language Processing (NLP) since the 1950s. Pertinent early research on document summarization is featured in a FnTIR paper [5]. Moreover, contemporary surveys focusing on document summarization utilizing deep learning techniques can be found [6]. And then as long as we talk about the summarization techniques, it will be divided into two directions, extractive summarization and abstractive summarization. The BART-large-cnn model itself is designed to generate abstract summaries, but the summaries contained using the model are extractive summaries. Therefore, in this thesis, we use this model as an extractive model to make a comparison. The recent summarization models are all transformer-based; the transformer is still the current go-to architecture for natural language processing. At the same time, ROUGE is a widely used evaluation standard for weighing the summarization model.

2.1 Summarization techniques

2.1.1 Extractive summarization

Extractive summarization [7] is a method for extracting important content from the original text. By this method, the extracted output is achieved as the form of summary. For the extracted parts, they are important words or sentences, which depends on the length of the original subsentences.

Here is an example [1]:

- Content: The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building, and the tallest structure in Paris. Its base is square, measuring 125 metres (410 ft) on each side. During its construction, the Eiffel Tower surpassed the Washington Monument to become the tallest man-made structure in the world, a title it held for 41 years until the Chrysler Building in New York City was finished in 1930. It was the first structure to reach a height of 300 metres. Due to the addition of a broadcasting aerial at the top of the tower in 1957, it is now taller than the Chrysler Building by 5.2 metres (17 ft). Excluding transmitters, the Eiffel Tower is the second tallest free-standing structure in France after the Millau Viaduct.

- Summary: The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building. Its base is square, measuring 125 metres (410 ft) on each side. During its construction,

---

[1] An example of extractive summarization url `https://huggingface.co/facebook/bart-large-cnn`.

the Eiffel Tower surpassed the Washington Monument to become the tallest man-made structure in the world.

2.1.2 Abstractive summarization

Abstractive summarization [8] is a method of interpreting the original text, but not using an extraction method. By this way, the summary obtained by this method will be more coherent, and it is more inclined to be understood by people's thinking logic. On the other hand, because abstractive summaries may differ to a large extent from the original content, abstractive models have the risk of hallucinating untrue content in their output.

Here is the same example as the section 2.1.1, the abstractive summarization is as follows [2]:

- Summary: The Eiffel Tower is a landmark in Paris, France.

2.2 Transformer

Before the advent of transformer model, Recurrent Neural Networks (RNNs) were popular. But slowly the disadvantages of RNNs emerged. RNNs can only take in one input token and the previous hidden state at a time and then get the output. This time series structure enables the model to obtain long-term dependencies, but it also makes it impossible for parallel computing, and the model efficiency is very low. Therefore, the transfomer model was introduced. Google proposed the field of sequence transcription, completely abandoned CNN and RNN, and only relied on a simple network architecture of the attention structure, named transformer [9].

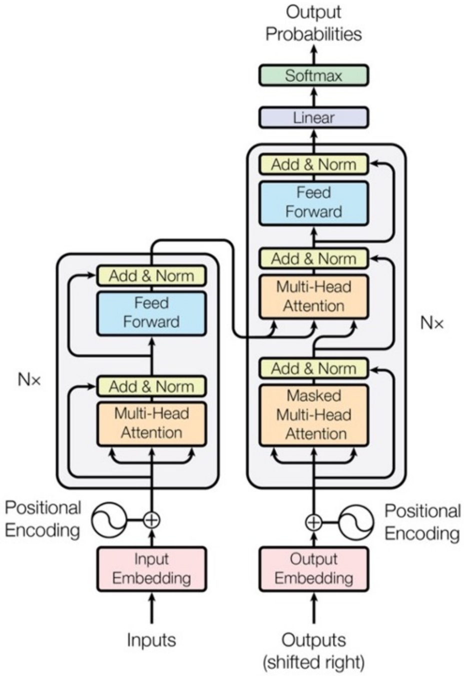


Figure 1: The Transformer - model architecture [9]

---

[2] An example of abstractive summarization url `https://huggingface.co/google/pegasus-xsum`.

Transformer is an encoder-decoder architecture with self-attention between the input tokens. Encoder-decoder models are designed for sequence-to-sequence learning. The Seq2seq task refers to the model used when the length of the output is uncertain for tasks whose input and output are both sequences. Generally speaking, the most commonly used mode of seq2seq is the encoder-decoder mode, which first encodes the sequence into a context matrix, and then uses the decoder to decode it. We just use the context vector as the input from the encoder to the decoder.

The self-attentive mechanism will focus on the relevance of words within the text, by calculating scores to get which words are more relevant to each other. By this it is also possible to get which words contribute more to the summary. The advantages of Attention can be simply analyzed as following:

- The speed is fast and solves the shortcoming that RNNs cannot run in parallel. Each step of the computation does not depend on the result of the previous one, so it can be done in parallel.
- The effect is good, which solves the shortcoming that long distance information will be weakened.

But on the other hand, it also has disadvantages. For example, the memory load could be high because the computational complexity is quadratic to the length of the input. This will require high-performance GPU computing.

For the working process of transformer, it is like a black box from a macro perspective, output comes by given the inputs. In the black box, it is composed of encoding components, decoding components and the connections between them. The encoder consists of a bunch of encoders, and the decoder is the same. All encoders are structurally the same, but have no shared parameters. Each encoder can be decomposed into two word layers (self-attention and feed-forward neural network). Self-attention helps the encoder focus on other words in the input sentence while encoding each word. The output of self-attention is passed to the feed-forward neural network, and the feed-forward neural network corresponding to the word at each position is the same. The decoder also has these two layers, but there is also an attention layer in the middle to focus on the relevant parts of the input sentence.

2.3 Transformer models for summarization

2.3.1 BART

BART uses the standard sequence-to-sequence Transformer architecture from (Vaswani et al., 2017), except, following GPT, that ReLU activation functions are modified to GeLUs (Hendrycks & Gimpel, 2016) and initialise parameters from N (0, 0.02). [3] For the base model, 6 layers are used in the encoder and decoder, and for the large model, there are 12 layers in each. In this thesis, the BART model we use is pre-trained on English language and fine-tuned on CNN Daily Mail dataset.

BART [3] proposes a pre-training model combining bidirectionally and auto-regression, which merges the two strengths from BERT and GPT models respectively. The model uses arbitrary noise to destroy original text, and then learns model to reconstruct the text. A point worth noting is that the model will use randomly different noise functions to avoid the model over-relying on information related to the structure of the sequence. Its application to downstream tasks is numerous, of which the summarization mentioned in this thesis is one.

In this thesis, we have chosen this model to make a comparison, because it does return extractive summary output. For example, we choose an example:

- Content: To me, it looks like 505 has Overkill by the balls. Trying to just push out as much DLC as possible. Based on everything OK did with PD1, this just isn't their norm. These weapon packs are completely unnecessary just because of the sheer amount of firearms the game has offered since day 1. The Death Wish difficulty itself had potential but a reskinned bulldozer and cops isn't going to be enough to get me to spend any money. Hell that's like me paying to get my ass kicked, when I can just play the normal Overkill difficulty and do that for free.
  It really does just look like 505 says "Hey, I've got a bigger whip this time! So in the next month or so you guys are going to get another full DLC released and that will be that." And the Overkill guys sit down and birth some ideas prematurely and give it a neat little trailer in the hopes that will help it sell better and that's it.
  I played the armored transport mission was not all that impressed. It was fun once but when the mission was over I sat there wondering if there was going to be more but nope.

By using the pre-trained model BART-large-cnn, we can get the following summary as output:

- Model summary: To me, it looks like 505 has Overkill by the balls. Based on everything OK did with PD1, this just isn't their norm. It really does just look like 505 says "Hey, I've got a bigger whip this time!"

From this example, we can clearly see that this summary is very close to the behavior of an extractive summary.

2.3.2 PEGASUS

The base architecture of PEGASUS is a standard Transformer encoder-decoder. Both GSG and MLM are applied simultaneously to this example as pre-training objectives. [4] The pegasus model we use is pre-trained on both C4 and HugeNews and further tested by the XSum dataset.

PEGASUS [4] proposes a new method for generate abstractive summarization, which called Gap Sentences Generation. In this model, deleting or masking "important sentences" from the original text and using the remaining sentences to generate these deleted or masked sentences in the output, just as what is done in summarization tasks in terms of the output.

2.4 ChatGPT

First, we can have a look at the GPT model. GPT stands for the generative pre-trained model, as a Transformer decoder-only model. [10] The input for the model is prompt, which will be sent to the transformer. The output will then be obtained by focusing on the input sequence through the self-attention mechanism.

GPT has been released in several versions, the latest of which is GPT-4. GPT models have shown excellent performance in areas such as natural language processing, text generation, and machine translation. The ChatGPT we use in this thesis is an instance of GPT-3.5 that was fine-tuned for conversations.

ChatGPT is a natural language generation model that automatically generates human-understandable natural language output based on the input prompt. Because ChatGPT has become very popular in recent times and because of the power of its features, we would like to try to use ChatGPT to do zero-shot summaries and compare them with previous results.

2.5 ROUGE metrics

The meaning of ROUGE [11] is ”Recall-Oriented Understudy for Gisting Evaluation”. By doing this, we generally get two scores, a recall value and a precision value. Comparing the summary obtained by the model with the reference summary, we can get the count of the same word parts, divide this count by the total number of words in the summary obtained by the model to get the recall value, and on the other hand divide it by the total number of words in the reference summary to get the precision value. Finally, we can get the F1-score by calculating the following formula:

$$F1 = \frac{2 * (precision * recall)}{precision + recall}$$ [3]

Thus we can use ROUGE to measure automatic summarization and machine translation. In this article, we only refer to the use of automatic summarization. It calculates a score by comparing the summaries generated by the model with the reference summaries, to measure the similarity between the two summaries, or to say the model accuracy.

In this thesis, we will discuss the metrics of ROUGE-N and ROUGE-L, of which we will only use ROUGE-1 and ROUGE-2 for ROUGE-N. And for each method, we can get precision and recall values separately, then get the F1 score from these two.

1) ROUGE-1: The numerator is the number of single words that are identical between the summaries from the model and reference summaries. Then use this to divide by the number of single words from summaries obtained by the model to get the precision, and otherwise divide by the reference summaries to get the recall.

2) ROUGE-2:The numerator is the number of bigrams that are identical between the summaries from the model and reference summaries. Then use this to divide by the number of bigrams from summaries obtained by the model to get the precision, and otherwise divide by the reference summaries to get the recall.

3) ROUGE-L:The numerator is the length of longest common subsequence that are identical between the summaries from the model and reference summaries. Then use this to divide by the number of words from summaries obtained by the model to get the precision, and otherwise divide by the reference summaries to get the recall.

3 Methods

All the code are runned in the Google colab. We chose to use the services of colab pro because it contains more usable computing units and more memory.

---

[3] The F1 score url `https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9`.

3.1 Dataset

The dataset used for evaluation is the Reddit dataset on the Huggingface platform [4]. This is a dataset which contains posts contents and reference summaries. It is interesting because the reference summaries are written by users, which also means that they are subjective.

The dataset consists of 3,848,330 posts with an average length of 270 words for content, and 28 words for the summary [2]. All the summaries are the TLDRs are written by the authors of the posts. An example from the dataset will be given below, we can also see in the example that body is more of a summary than content:

```
{
    "author": "raysofdarkmatter",
    "body": "I think it should be fixed on either UTC standard or UTC+1 year
            around, with the current zone offsets. Moving timescales add a lot of
            complexity to the implementation of timekeeping systems and have
            [dubious value]( I think seasonal shifting time made sense in the
            pre-electric past, when timekeeping was more flexible and artificial
            light was inefficient and often dangerous. Now we have machines that
            work easily with simple timekeeping rules, and it’s more beneficial
            to spend a small amount on energy for lighting, and save the larger
            cost of engineering things to work with the complex timekeeping rules,
            as well as saving the irritation to humans. Lighting has gotten much
            more efficient over time; we can squeeze out a lot more photons per
            unit of energy from a 2012 CFL or LED than a candle could in 1780, or
            a lightbulb could in 1950. There’s a lot of room for improvement in
            how we use lights as well; as lighting control gets more intelligent,
            there will be a lot of savings from not illuminating inactive spaces
            constantly. tl;dr: Shifting seasonal time is no longer worth it.",
    "content": "I think it should be fixed on either UTC standard or UTC+1 year
            around, with the current zone offsets. Moving timescales add a lot of
            complexity to the implementation of timekeeping systems and have
            [dubious value]( I think seasonal shifting time made sense in the
            pre-electric past, when timekeeping was more flexible and artificial
            light was inefficient and often dangerous. Now we have machines that
            work easily with simple timekeeping rules, and it’s more beneficial
            to spend a small amount on energy for lighting, and save the larger
            cost of engineering things to work with the complex timekeeping rules,
            as well as saving the irritation to humans. Lighting has gotten much
            more efficient over time; we can squeeze out a lot more photons per
            unit of energy from a 2012 CFL or LED than a candle could in 1780, or
            a lightbulb could in 1950. There’s a lot of room for improvement in
            how we use lights as well; as lighting control gets more intelligent,
            there will be a lot of savings from not illuminating inactive spaces
            constantly.",
```

---

[4] The Reddit dataset url `https://huggingface.co/datasets/webis/tldr-17`.

```
    "id": "c69al3r",
    "normalizedBody": "I think it should be fixed on either UTC standard or UTC+1
            year around, with the current zone offsets. Moving timescales add a
            lot of complexity to the implementation of timekeeping systems and
            have [dubious value]( I think seasonal shifting time made sense in the
            pre-electric past, when timekeeping was more flexible and artificial
            light was inefficient and often dangerous. Now we have machines that
            work easily with simple timekeeping rules, and it’s more beneficial
            to spend a small amount on energy for lighting, and save the larger
            cost of engineering things to work with the complex timekeeping rules,
            as well as saving the irritation to humans. Lighting has gotten much
            more efficient over time; we can squeeze out a lot more photons per
            unit of energy from a 2012 CFL or LED than a candle could in 1780, or
            a lightbulb could in 1950. There’s a lot of room for improvement in
            how we use lights as well; as lighting control gets more intelligent,
            there will be a lot of savings from not illuminating inactive spaces
            constantly. tl;dr: Shifting seasonal time is no longer worth it.",
    "subreddit": "math",
    "subreddit_id": "t5_2qh0n",
    "summary": "Shifting seasonal time is no longer worth it."
}
```

3.2 Dataset preparation

In this step, the dataset and the model are loaded. Then the irrelevant columns are removed and we only keep the “content” and “summary” columns. Then because running the entire data is impossible, we took the first 2,000 rows as a sample for subsequent use. The selected 2000 rows are applied equally to BART, PEGASUS and ChatGPT, in other words all three models deal with the same batch of data.

3.3 Zero-shot use of BART and PEGASUS

We compare two encoder-decoder models in a zero-shot setting to get zero-shot summaries, which are the summaries obtained not using fine-tuning. We use the 2000 rows data mentioned above to test the zero-shot summaries. In this process we used the datasets package to process the imported Reddit dataset. Then we used the package transformers to import and run the model to get the summary content we wanted. At last we compute the ROUGE scores by comparing the model summary with the reference summary.

For the code that will be used, we chose to reference a similar approach in this article. [12]

3.4 Prompting ChatGPT

We use ChatGPT in a zero-shot way with minimal instructions. We call the ChatGPT API to get access to the generative model GPT-3.5 (text-davinci-003). In this step, to ensure that we are not affected by the uncertainty introduced by the probabilistic summary of generative LLMs, we set the temperature parameter to 0. We also kept the output summary length similar to the previous two models by declaring in the prompt that we want the summaries of the maximum number of words, in order to have some comparability of the results obtained, and also it is a good way to check whether the model can summarize the post in more concise terms and also not loose the original meaning.

The following is the function we use for the prompt:

```
def summary(content):

    txt_sum = "Please summarize the following reddit post in max 40 words:\n\n"+content
    completion = openai.Completion.create(
    model="text-davinci-003",
    prompt= txt_sum,
    temperature=0,
    max_tokens=64,
    frequency_penalty=0.0,
    presence_penalty=0.0
    )
    response = completion.choices[0].text

    return response
```

After that we use a simple for loop to bring all the original posts into the function and get the result, storing it in a new array. The array is then compared to the reference summaries to get the ROUGE values we need for subsequent discussion.

3.5 Evaluation

We use the ROUGE metrics for evaluation (see section 2.5). In this thesis we are using the `rouge` package, through the rouge package to import `Rouge`, after that you can use `Rouge()` to create a `rouge` object. Using this function, `rouge.get_scores(model summary, reference summary, avg=True)`, we can get the scores we need for ROUGE.

4 Results

In this section we use the same example content as section 2.3.1 to show the results of the model, the reference summary is shown below:

- Reference summary: 505 is beating Overkill shitless for more DLC and it really isn't pretty.

4.1 BART-large-cnn model

For this model, because it basically extracts the important words or sentences from the original text, the score for recall should be relatively high, even though now we can see that the score for recall in the Table 1 is quite low but is higher compared the result by PEGASUS-xsum model in Table 2 afterwards. And also for precision, because the reference summaries are written by people,

	recall	precision	F1-score
ROUGE-1	0.22	0.10	0.12
ROUGE-2	0.03	0.01	0.02
ROUGE-L	0.20	0.09	0.11

Table 1: BART-large-cnn zero-shot ROUGE score

the degree of abstraction is a bit higher, so inevitably we will get fewer matching text words for the summaries obtained by the model.

The model summary is as follows:

- Model summary: To me, it looks like 505 has Overkill by the balls. Based on everything OK did with PD1, this just isn't their norm. It really does just look like 505 says "Hey, I've got a bigger whip this time!"

4.2 PEGASUS-xsum model

	recall	precision	F1-score
ROUGE-1	0.10	0.12	0.10
ROUGE-2	0.01	0.01	0.01
ROUGE-L	0.08	0.10	0.08

Table 2: PEGASUS-xsum zero-shot ROUGE score

For this model, it makes more abstract summaries compared to the BART-large-cnn model. Based on it, the recall value of this model is quite low. And also the precision is not so good.

The model summary is as following:

- Model summary: Call of Duty: Black Ops II's first Death Wish DLC has been released and it's pretty disappointing.

Discussion about the zero-shot use for the two models

From the example given above, we can see that BART-large-cnn model is behaving like an extractive model. Although this model was not originally dedicated to generating extractive summaries, based on previous research [13] we can find that the result obtained from this model is an extractive summary. As we have mentioned, the extractive model extracts key words and phrases from the original text to compose the summary. The summaries thus obtained will have more overlap with the reference summaries in our dataset. And because the text length of the extractive summary is usually longer than that of the reference summary, the higher value of recall over precision can be reflected in the table above.

For the PEGASUS-xsum model, it generates abstractive summaries. Corresponding to the data in the table we can see that the recall value is very low because the abstractive summary is obtained. Also, a lower precision value is returned. Although this abstractive summary is more fluent, it is possible that the abstractive summary of this model is a little different in direction, and therefore the precision is lower.

According to the comparison of these examples, we can clearly see the advantage of BART-large-cnn model for the recall value. The final result of F1-score is also slightly better for BART-large-cnn.

4.3 ChatGPT

	recall	precision	F1-score
ROUGE-1	0.18	0.11	0.13
ROUGE-2	0.03	0.02	0.02
ROUGE-L	0.16	0.10	0.11

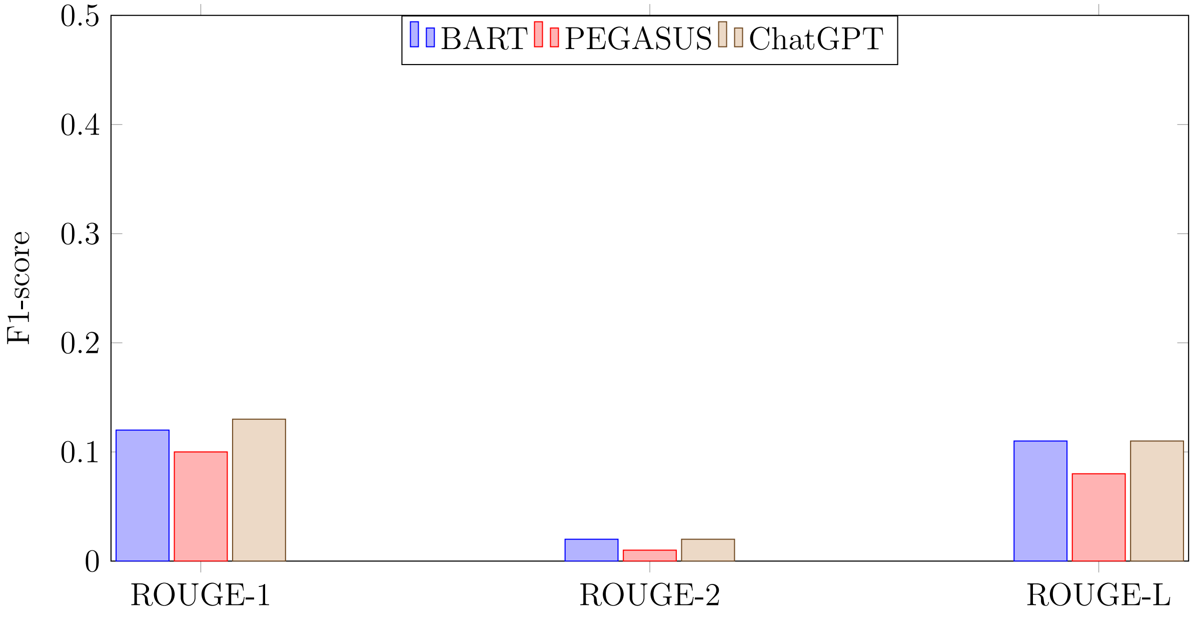
Table 3: ChatGPT zero-shot ROUGE score

With this summary in the next paragraph we can see that ChatGPT's approach to summarization is also close to extractive summarization, but not exactly the same. Most of the words and phrases used in it are also key information in the original content, but instead of just extracting these key words and phrases and simply splicing them together, some processing has been done to make the reader understand the key meaning of the original content better. So we can see in the Table 3 that the recall value of the model is slightly lower than BART, while its precision is slightly better. Finally, in the comparison of F-1 values, we can see that ROUGE-1 score is slightly better than BART, while the other two are relatively equal.

The model summary is as following:

- Model summary: 505 is pushing out DLCs for Overkill, but the Death Wish difficulty and Armored Transport mission don't seem to be enough to get people to spend money.

Bar chart for the F1-score obtained from the three models



Discussion about the comparison of ChatGPT with the other two models

In order to discuss this, we first need to take a deeper look at ChatGPT, whose biggest improvement over previous chatbots is the use of reinforcement learning from human feedback (RLHF), which can be interpreted in such a way that the results it obtains will rely heavily on human feedback. Based on this, we can assume that ChatGPT is more likely to produce summaries that are easier for humans to understand. And the number of parameters of the model is also much larger, so it will be more fluent than the other two models.

And in the results in the tables above, we can see the scores for ChatGPT obtained a little better than BART and PEGASUS. Since the scores of BART are better than PEGASUS, we now compare ChatGPT to BART.

In the previous section we have already mentioned that ChatGPT's summarization is very similar to BART's extractive summarization results but these two models are still very different. Because BART only presents a few important sentences in a complete way, while ChatGPT is not like this, it extracts important words and sentences from the original content, and then integrates them into a suitable summary in a way that readers can understand, although it may use words that do not appear in the original content, it conveys the meaning of the original content well and is more convenient for people to understand. This, on the other hand, confirms that ChatGPT's summaries are closer to the reference summaries than BART's summaries, which is also reflected in the comparison of the precision scores in the table. But it's true that it is interesting that ChatGPT can better summarize than BART while BART was trained for summarization and ChatGPT is a generative model that was finetuned for conversations.

In addition, for PEGASUS, although as an abstract type of summary, it may be more suitable for the expression of human emotions, for example, in some posts, people will not just use the content of the original content to summarise, they may use some of their own feelings or some internet slang to express their summary opinions. But to a certain extent, it also loses its accuracy in summarising the content of the original posts and sometimes confuses people who just read the summary about the content of the original posts instead. So in this respect, ChatGPT performs better.

5 Discussions

In this thesis we compare the quality of a number of summarization models in a zero-shot setting: BART, PEGASUS, and ChatGPT. We use a sample of reddit posts from the Reddit dataset on Huggingface as the dataset to evaluate the models. Because the original dataset is so huge, we use only 2000 rows to do the experiment, trying to balance between the number of samples and the running time of the experiment. So although we did a balance, the partial use of the sample was really a limitation for us as well, which may have some impact on the results to some extent.

On the other hand, however, although we are using partial data, it is still a time-consuming experiment for natural language processing, which makes it challenging to obtain the results. By way of the code used in this thesis, using either BART or PEGASUS, to get a summary of 2000 rows of data both take about 6 hours.

Another limitation is the models we chose. As we mentioned earlier BART itself is an abstractive summarization model, but because the results it obtains are extractive summaries, we treat it as an extractive model in the thesis for comparison. However, this is not essentially an extractive model after all, and there will certainly be some bias in the results.

It is also worth thinking about the choice of ChatGPT's prompt. In this thesis, we used the prompt because we wanted to rely on a shorter summary to get closer to the reference summary, and after several attempts, we felt that this yielded a relatively cleaner and more aesthetically pleasing result that did not lose the original intent.

Furthermore, because this thesis uses the ROUGE metrics, the criterion has its own flaws. Because this method counts the number of overlaps of words or fragments to calculate the recall and precision values, sometimes the resulting summaries may differ in the form of expressions resulting in a lower score, but it can still be a good summary.

This thesis in our opinion will help people to have a deeper understanding of the different summaries brought by different models, so that people can choose a more suitable model to use based on their current situation.

Considering the limitations and future Work

As mentioned above, there are many limitations to this thesis. One obvious point that could be optimized for future exploration of this thesis is the limitations of the ROUGE criterion mentioned at the end above. Because other metrics can be taken to measure whether these models are still as such as this thesis suggests. Of course, apart from this, there are still a lot of limitations that can be improved.

6 Conclusion

The release of ChatGPT was undoubtedly a big breakthrough, both in terms of the model size, the amount of training data, and the directions in which it can be applied. In this thesis, we use ChatGPT for summarising Reddit posts and compare it with two models, BART and PEGASUS. In the end, we believe that ChatGPT performs better and it not only inherits the important information (words) in the original text well but also conveys the meaning of the original content to the reader well.

References

[1] Anushka Gupta, Diksha Chugh, Anjum, and Rahul Katarya. Automated news summarization using transformers. *CoRR*, abs/2108.01064, 2021. URL `https://arxiv.org/abs/2108.01064`.

[2] Reddit dataset. URL `https://huggingface.co/datasets/webis/tldr-17`.

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL `http://arxiv.org/abs/1910.13461`.

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

[5] Pertinent early research. . URL `https://www.cis.upenn.edu/~nenkova/1500000015-Nenkova.pdf`.

[6] Document summarization. . URL `https://dl.acm.org/doi/full/10.1145/3529754`.

[7] N. Moratanch and Chitrakala Gopalan. A survey on extractive text summarization. pages 1–6, 01 2017. doi: 10.1109/ICCCSP.2017.7944061.

[8] N. Moratanch and Chitrakala Gopalan. A survey on abstractive text summarization. pages 1–7, 03 2016. doi: 10.1109/ICCPCT.2016.7530193.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL `https://arxiv.org/abs/1706.03762`.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

[12] Usage of text summarization code. URL `https://towardsdatascience.com/the-secret-guide-to-human-like-text-summarization-fcea0bfbe801`.

[13] Gabriel Orlanski and Alex Gittens. Reading stackoverflow encourages cheating: Adding question text improves extractive code generation. *CoRR*, abs/2106.04447, 2021. URL `https://arxiv.org/abs/2106.04447`.

A Some examples of the summarization

Reddit post 1

- Content: The missions do seem repetitive at times, however, I notice that playing on a different map/level each time makes it feel a little more diverse. I usually end up making a stupid decision, going back a save, and changing my outcome.

  I bought the game when it wasn't on sale for about \$30 and so far I've invested 38 hours of gameplay into it. I'll be honest, I don't feel the best at purchasing it for 30 bucks but I do feel that it was worth it. (\$30/38hours = about \$0.79/h)

  As for the multiplayer, I found it highly repetitive, and wound up changing my team formation a lot to find out what I felt best suits me. Of coarse, I've never tried custom matches either :/.

  The enemies always have different tactics in place for you if load back a save, this may seem inconsistent but leaves it to be a constant challenge.

  I can't recall ever losing a character due to capturing an alien myself, but I have yet to play iron man mode.

  I've had a lot of fun screwing with my character's info. [example](

  I'd say you have to have a specific taste for this game, as it's turn based. Summed up, it's like Destroy All Humans mixed with civ 5, and gears of war.

- BART summarization: The missions do seem repetitive at times, however, I notice that playing on a different map/level each time makes it feel a little more diverse. The enemies always have different tactics in place for you if load back a save, this may seem inconsistent but leaves it to be a constant challenge. I'd say you have to have a specific taste for this game, as it's turn based.

- PEGASUS summarization: What's your opinion on the game?

- ChatGPT summarization: Game has repetitive missions, but playing on different maps/levels makes it diverse. Worth \$30 for 38 hours of gameplay. Multiplayer is repetitive, but enemies have different tactics. Fun screwing with character info. Specific taste for turn-based game like Destroy All Humans, Civ 5, and Gears of War.

- Reference summary: If you feel that you'd like this kind of game, purchasing it for \$10 is great. If you aren't too sure, there's always the winter sale :)

Reddit post 2

- Content: A number of years ago while I was still in high school, I had a few friends over for some drinks and stuff while we went swimming in the pool (my parents were out of town). We had all taken about 2.5 grams, and were enjoying the dip in the pool, when all of a sudden the pool light turned off randomly. It kind of spooked us all, but I just got up and walked over to the switch inside to check it out, it had been pushed down, however no one was inside at all. When I turned around to go back outside, my dog was growling at the corner of the room that I had been in, barking and snarling, with the hair on his back raised. My dog

NEVER barks, and in the 5 years I had had him I had no idea he could even growl, so this fucking freaked me out. I grabbed my dog, brought him back outside, and jumped back into the now lit pool. I explained to my friends what had happened, to which one of them said, ”Well at least the pool light is back on no-..” then the pool light shut off again, immediately barking ensues. I to this day have no idea what was going on in that house, but it really led to a bad trip for all of us.

- BART summarization: A number of years ago while I was still in high school, I had a few friends over for some drinks and stuff while we went swimming in the pool. All of a sudden the pool light turned off randomly. It kind of spooked us all, but I just got up and walked over to the switch inside to check it out. No one was inside at all. When I turned around to go back outside, my dog was growling at the corner of the room that I had been in, barking and snarling. My dog NEVER barks, and in the 5 years
- PEGASUS summarization: In our series of letters from African-American journalists, film-maker, and columnist Orson Scott Card looks back at some of his most memorable stories.
- ChatGPT summarization: Friends had a pool party while parents were away, took 2.5 grams and were spooked when the pool light randomly turned off. Dog started growling and barking, then the light shut off again. Led to a bad trip for all.
- Reference summary: Mushrooms, and ghosts.

Reddit post 3

- Content: How about increasing the wingers clip size. A point blank pistol shot with the stock pistol deals 22 damage while a point blank winger deals 26 damage. How ever, the winger has a -60% clip size, making it’s clip deal less damage then the stock pistol. Stock: 22 damage x 12 clip = 264 damage total Winger: 26 damage x 5 clip = 130 damage total Now that we have that, we see that the winger does roughly half the damage of the pistol if all shots are unloaded. This may be balanced depending on how high you feel that a extra high first jump is worth. If you are willing to sacrifice 2% damage for 1% jump boost, then the winger is in a balanced spot right now. If you are only willing to sacrifice 1% damage for 1% Jump height, then not quite. In order for that to be balanced, the winger would basically need a 50% clip increase. (50% of 50% is 25%, making it so that the full clip would deal 75% percent of the standard pistol damage.) Using that thought process, the clip size must be 7.5 shots, which is not entirely possible. Round it to 8 clip size and let’s see how much damage a point blank winger clip will deal. Improved winger: 26 damage x 8 clip = 208 total damage 208 Improved winger damage / 264 Stock pistol damage = 0.78787878787 or 79% of the damage as the stock pistol. This buff makes it slightly more powerful than the stock pistol in theory, but not all maps have special winger jumps for you. The Improved winger would be better than stock on maps with many winger jumps, but not as good on one’s without them. Theory aside, the improved winger is really only 4% more powerful than stock in the end, making it not ”Game breaking” per say. All in all, the effectiveness of either weapon would all depend on how the user is able to utilize it, be it through finding and executing tricky jumps with the winger, or by utilizing the full combat potential of the stock pistol.

- BART summarization: A point blank pistol shot with the stock pistol deals 22 damage while a point blank winger deals 26 damage. How ever, the winger has a -60% clip size, making it's clip deal less damage then theStock pistol. The effectiveness of either weapon would all depend on how the user could utilize it.

- PEGASUS summarization: Let's take a look at the winger and the stock pistol.

- ChatGPT summarization: Increase winger clip size to 8 shots for more damage than stock pistol, but effectiveness depends on user's ability to utilize it.

- Reference summary: The winger is either balanced or not depending on your personal value of the jump boost. Math proves that if you value Damage% the same as Jump boost% , the winger must have a 8 clip size to be AROUND equal to the stock pistol.

Reddit post 4

- Content: oh. i see. i will edit my post to correct for this error. thanks man. i'm still new to reddit so stuff like OP and throwaway account are still new concepts. im learning. still dont know what

- BART summarization: i'm still new to reddit so stuff like OP and throwaway account are still new concepts. i will edit my post to correct this error.

- PEGASUS summarization: What is throwaway account and how do you use it?

- ChatGPT summarization: OP means

  Person is new to Reddit and is learning the concepts. They will edit their post to correct an error and thanked the person who pointed it out.

- Reference summary: means.

Reddit post 5

- Content: It happens when you help run a pet care business. Do you think there might be other circumstances besides size that contribute to this anecdotal causality? Perhaps, people who buy bigger dogs recognize they are a bigger investment and harder to take care of, so only more experienced or capable people buy them? Maybe, people with violent, disturbed big dogs cant handle wrestling them into a car and trucking them to your establishment? Maybe, smaller dogs are MUCH more common, so you see a wider variety of personalities? Maybe, the people who buy small dogs tend to not want a dog as a friend and just and accessory, or dont know how to raise a pet (maybe first time owner, got it as a gift etc, no one gives huge dogs as gifts). Maybe some BREEDS of dogs tend to be more violent, but also tend to be small (chihuahua or Pomeranian), maybe the part of town your business is in has a particular demographic of people who buy certain breeds and raise them particular ways to get similar personailities? Maybe your a dick to the dogs and scare the shit out of the dogs too small to understand youre not a threat? Maybe your predisposed opinion of small dogs makes you treat them differently and you use confirmation bias to judge them unfairly? Maybe you just dont have a fucking clue what youre talking about?

- BART summarization: Do you think there might be other circumstances besides size that contribute to this anecdotal causality? Perhaps, people who buy bigger dogs recognize they are a bigger investment and harder to take care of, so only more experienced or capable people buy them. Maybe, people with violent, disturbed big dogs cant handle wrestling them into a car and trucking them to your establishment? Maybe, smaller dogs are MUCH more common, so you see a wider variety of personalities.
- PEGASUS summarization: Have you ever had a customer come into your establishment and say "I want a big dog but I don't want any big dogs" or "I want a small dog but I don't want any small dogs" or "I want a big dog but I don't want any small dogs"?
- ChatGPT summarization: People speculate that size may not be the only factor in why small dogs are more aggressive. Possible explanations include breed, owner experience, and environment.
- Reference summary: Theres a million other reasons that contribute to a dogs personality than size, and theres a myriad of variables that can contribute to why you only see gentle giants and tiny tyrants. Saying all small dogs are bad is a idiotic and ignorant statement made by someone who has no idea what theyre talking about.