Artificial Intelligence In Industry

Project Report

Abstractive Summarisation of Long-form Medical Papers

Gee Jun Hui Leonidas Yunani

Academic Year: 2021 / 2022

Contents

1 Introduction

The domain of text summarisation involves the compression of semantic information into a shorter form without a large loss of relevant information. There are two types of summarisation involved, namely: extractive and abstractive. Extractive summarisation takes a text and reduces it by selecting only relevant sentences according to some defined metric. Abstractive summarisation on the other hand takes a text and reduces it by rewriting the text, while preserving the most relevant semantic information. The latter represents the most common method used by humans in solving summarisation tasks and has seen great results achieved lately using transformer models such as BART and PEGASUS.

Transformer models typically impose a restriction on the maximum length of the input due to the needed amount of memory to compute the self-attention over it. The amount of memory needed is quadratic on the length of the input. This means that increasing the maximum length of the input, increases drastically the needed memory for self-attention. For certain language tasks, such a limit is sufficient in practice. However, in the domain of text summarisation, the number of tokens of a document may exceed the input limit especially in fields such as medicine and law.

To overcome this limitation, two different approaches have been proposed. The first involves the application of extractive summarisation to shorten a document by keeping only the most relevant sentences. This shortened text is then passed to a transformer to produce the final abstractive summary. The second approach instead involves the designing of novel transformer architectures whose input limit is increased beyond the typical 1023 tokens by using self-attention layers that scale linearly with input length. The trained model is then used to summarise the document directly without requiring any prior reduction in the document size.

This project proposes a method involving the first approach to summarise long-form documents. In short, an extractive summarization method is proposed that extracts the most relevant sentences based on their sentence embeddings and using nearest neighborhood search. The extracted text is then passed to a pre-trained transformer to produce the final abstractive summary. The method is compared to a baseline approach whereby the abstractive summarization is applied directly to the text and an approach involving an initial extractive summarization using Textrank. The alternative methods are also compared to the state-of-the-art approaches on the Papers With Code website.

2 Motivation

The primary motivation for this project is to design a system that is able to automatically summarise a medical document in an abstractive way. In the medical field, medical documents are often long and complex textual information that require considerable time to parse and understand. An informative summary would allow medical practitioners to quickly parse through a list of

medical documents in order to filter out the most relevant documents according to their needs. Hence, medical practitioners may focus their limited time on reading and understanding the full text from the most relevant medical documents.

3 Dataset

The dataset used consists of documents from PubMed along with their golden summaries. The dataset has already been divided into a training set, a validation set and a test set. For this project, the test set is used to evaluate the performance of the methods, which contains 6658 instances.

To begin, the texts and summaries are cleaned by joining them into single strings and removing the special tokens from the summaries. A fast transformer tokeniser is then used to determine the total number of tokens for each document. Finally, statistical analysis is applied to understand the overall length of the documents.

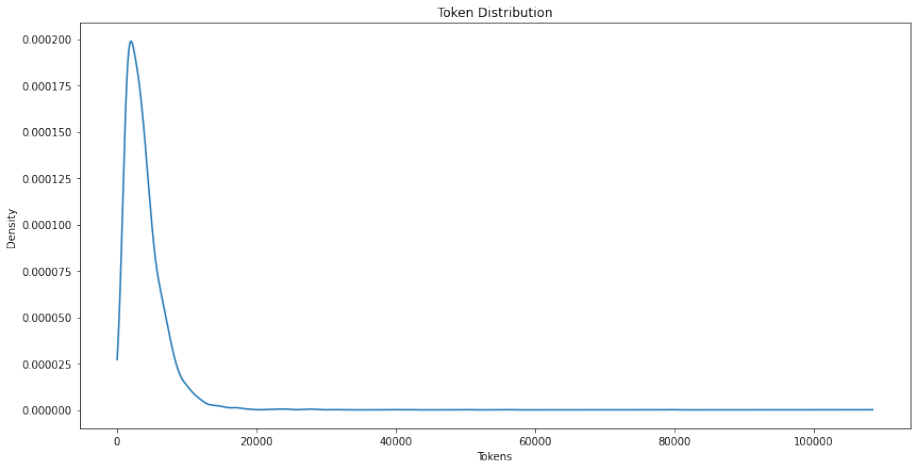

Figure 1: Token distribution of the documents.

- Mean: 3948
- Median: 3249
- Mode: 2003
- Minimum: 25
- Maximum: 108507
- Standard deviation: 3425

Based on the analysis above, it can be shown that the the distribution of documents by their number of tokens is right-skewed. A majority of documents have tokens that exceed 1023. There are also outliers with documents as short as 25 tokens and documents as long as 108507 tokens. A high standard deviation also shows that documents are more spread out in terms of their number of tokens. Hence, the dataset is appropriate in evaluating a methodology for summarising long-form documents.

4 Methodology

We establish three approaches to summarising long-form medical documents. Our approaches include the following:

- Initial extractive summarization using Textrank, followed by an abstractive summarization using a pre-trained transformer.
- Initial extractive summarization using BERT + KMeans, followed by an abstractive summarization using a pre-trained transformer.
- Initial extractive summarization using SBERT + Neighbourhood Search (NSearch), followed by an abstractive summarization using a pre-trained transformer.

For all the three approaches above, we utilise the same pre-trained transformer that is a PEGASUS model trained on the abstractive summarization of medical papers from PubMed [4]. PEGASUS is a seq2seq model that adopts self-surpervised learning to produce state of the art abstractive summaries. The model consists of an encoder and a decoder stack typical of transformer models. It differs from other models by incorporating a concept called Gap Sentence Genertion (GSG) whereby a sentence is masked and the model is trained to predict this sentence.

During training, three sentences are passed to the model. One sentence is masked and passed to the decoder for target prediction (GSG). The remaining two sentences are passed to the encoder with words being randomly masked via a process called Masked Language Modelling (MLM). PEGASUS has been shown to produce state of the art summaries using the ROUGE metrics on 12 public datasets for abstractive summarisation.

4.1 Baseline Approach

In the baseline approach, we apply the pre-trained PEGASUS model directly to the text. Due to the input token limit of the model, truncation must be applied for the approach to work. The point of truncation within the text cannot be controlled by the user. The approach establishes a set of baseline scores in order to evaluate if the alternative approaches of applying an initial extractive summarization brings any meaningful benefit to the produced summaries.

4.2 Textrank Approach

The Textrank approach involves the application of a graph-based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keywords [3]. In order to find the most relevant sentences in text, a graph is constructed where the vertices of the graph represent each sentence in a document and the edges between sentences are based on content overlap, namely by calculating the number of words that 2 sentences have in common. Based on this network of sentences, the sentences are fed into the Pagerank algorithm which identifies the most important sentences. The extractive summary is then built by taking only the most important sentences.

4.3 SBERT + NSearch Approach

The method begins by converting sentences into sentence embeddings using a pre-trained Sentence BERT model. The centroid of the semantic space is then calculated by taking the mean of all sentence embeddings. A nearest neighbourhood search is then applied to find the k-nearest sentences based on the cosine distance between the sentence embeddings and the centroid.

$$\texttt{cosine\_similarity} = \frac{A.B}{\|A\|\|B\|} \tag{1}$$

$$\texttt{cosine\_distance} = 1 - \texttt{cosine\_similarity} \tag{2}$$

The intuition is that the centroid represents the global semantic of the document in the semantic space. Hence, extracting only sentences whose embedding lie closest to the centroid would preserve only sentences which best explain the general gist of the document, while eliminating irrelevant ones.

A brute force method is applied to the nearest neighbourhood search to obtain the most accurate selection possible. The selected k-nearest sentences are then ordered according to their original order within the document to preserve positional information. Finally, the selected sentences are joined together to form the extracted text.

The selection of the hyperparameter k determines the final form of the extracted text. For this project, a heuristic is developed to approximate the number of sentences required to from an extracted text whose number of tokens is lesser than or equal to the input limit of the transformer.

The heuristic is formed by first calculating the total tokens and total sentences that make up the document. The total tokens is then divided by the total sentences to form a value which represents the approximate tokens per sentence.

$$\texttt{tokens\_per\_sentence} = \frac{\texttt{n\_tokens}}{\texttt{n\_sentences}} \tag{3}$$

Where `tokens_per_sentence` is the approximate tokens per sentence, `n_tokens` is the total number of tokens and `n_sentences` is the total number of sentences.

The final k is found by taking the max input tokens of the transformer and dividing it by the approximate tokens per sentence. The k value is floored to produce an integer that underestimates the number of sentences required.

$$k = \frac{\texttt{max\_tokens}}{\texttt{tokens\_per\_sentence}} \tag{4}$$

Where k is the number of neighbours and `max_tokens` is the input limit of the transformer.

This heuristic is only applied to documents with tokens that exceed 1023 tokens. Here, 1023 is used instead of the actual 1024 limit because 1 token is reserved by the transformer for the special token (<s>). Documents with less than 1023 tokens will simply have the original text returned. This heuristic is also applied to the (Textrank) approach in order to tell the model approximately how many sentences should be extracted from the text.

5 Experimentation

In the initial phase of the project, an approach involving K-Means clustering was tried to divide a document into k-number of disjoint paragraphs based on the cosine similarity of the sentence embeddings. This approach focuses on creating paragraphs which have similar local semantic value. However, setting the k-value to determine the number of paragraphs required is difficult. Setting a value that is too high may result in sentences with mathematical formulas being clustered together into a single paragraph, thus removing their original meaning within the context of the document. Moreover, paragraphs will not necessarily contain 1023 tokens or lesser. As such, the SBERT + NSearch approach was adopted instead to create paragraphs based on the global semantic value of the document.

In the SBERT + NSearch approach, a hyperparameter tuning was done on the k-number of sentences to search for. Multiple integer values and percentages were tried to improve the model's performance. It was determined that selecting as many sentences as possible, such that they formed a paragraph with 1023 tokens or less is optimal. As such, a heuristic was developed to approximate this value.

An additional extension to method was tested by creating an n-number of disjoint paragraphs instead of only one paragraph. This is because constraining the extractive text to a single 1023 tokens or less text may cause a great loss of information in larger documents. Attempts on a small subset of the data showed two issues with this approach.

The first is that setting the max length for the output of the abstractive summarisation model is important. Since the paragraphs are disjoint, the model may produce summaries that when joined together, exceed the length of the reference summary by two or more times. However, setting this max length optimally is difficult.

Secondly, the additional number of paragraphs that must be processed per document increases the querying time significantly. One method to solve this will be to parallelise the querying process. However, this requires additional computational cost that may not be available to most companies or individuals.

As such, creating more disjoint paragraphs could possibly capture more semantically relevant sentences from larger documents for abstractive summarisation, however this comes with certain drawbacks that must be accounted for, as demonstrated in the DANCER approach to long document summarisation [2].

6 Evaluation

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics used to evaluate the goodness of a summary. Typically, ROUGE-1, ROUGE-2 and ROUGE-L metrics are used for summary evaluation. The ROUGE-1 and ROUGE-2 metrics represent the number of matching unigrams and bigrams respectively. The ROUGE-L metric is the longest common subsequence (LCS) between the produced summary and the golden summary. It counts the longest sequence of tokens that is shared between both summaries.

The ROUGE metric values are calculated in terms of the recall, precision and F1-score. The F1-score is used as it provides a more reliable measure of the model's performance by relying not only on the model capturing as many words as possible (recall) but also on it not outputting irrelevant words (precision). The final model's performance is the average of ROUGE-1, ROUGE-2 and ROUGE-L scores.

7 Results

We evaluate the three approaches using the stated ROUGE-1, ROUGE-2, and ROUGE-L metrics [1].

Methodology	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	38.75	17.27	35.60
Textrank	39.10	17.15	35.67
SBERT + NSearch	38.92	17.12	35.59

Table 1: Quantitative analysis of the three approaches using the ROUGE metrics.

Based on the results in Table 1, we can see that the benefits in terms of applying an initial extractive summarization are not immensely substantive. A probable hypothesis may be that most of the relevant details in a medical paper are located in the initial parts of the text, hence a simple truncation after reaching the maximum input token limit may be sufficient to generate a good abstractive summary as done by the baseline approach. We can also observe from Table 3, that the abstractive summaries generated by all three

approaches are relatively similar from a qualitative standpoint. However, other initial extractive approaches such as the DANCER method, whereby disjoint extractive summaries are abstractively summarised and merged have shown to produce state of the art performance. Hence, applying an alternative initial extractive summarisation to abstractively summarise long-form text may still be promising.

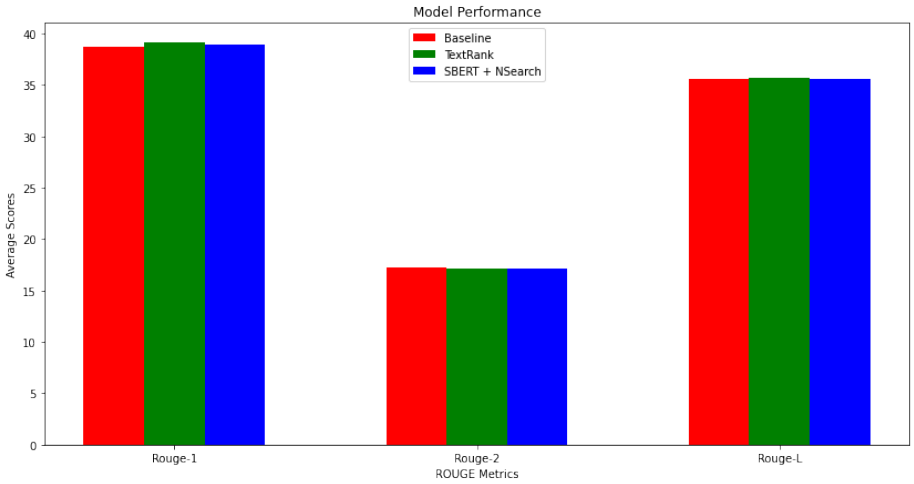

Figure 2: Model performance of the approaches.

We compare our methods to the state-of-the art approaches with complete ROUGE scores as listed on the Papers With Code website.

Methodology	ROUGE-1	ROUGE-2	ROUGE-L
HAT-BART	48.25	21.35	36.69
DANCER PEGASUS	46.34	19.97	42.42
BigBird-Pegasus	46.32	20.65	42.33
ExtSum-LG+MMR-Select+	45.39	20.37	40.99
ExtSum-LG+RdLoss	45.3	20.42	40.95
DANCER LSTM	44.09	17.69	40.27
DANCER RUM	43.98	17.65	40.25
MatchSum	41.21	14.91	36.75
Textrank	**39.10**	**17.15**	**35.67**
SBERT + NSearch	**38.92**	**17.12**	**35.59**
Fastformer	38.09	15.44	34.81

Table 2: Quantitative comparison with the state-of-the-art approaches using the ROUGE metrics.

Baseline anxiety is the most prominent and prevalent mood disorder in parkinson's disease (pd) ; however, little is known about the relationship between anxiety and cognition in pd. the aim of this study was to examine the influence of anxiety on cognition in pd by directly comparing groups of pd patients with and without anxiety while excluding depression. we hypothesized that pd patients with anxiety would show impairments in attentional set - shifting and working memory compared to pd patients without anxiety. furthermore, since previous work, albeit limited, has focused on the influence of symptom laterality on anxiety and cognition, we also explored this relationship. seventeen pd patients with anxiety and thirty - three pd patients without anxiety were included in this study. results showed that pd patients with anxiety had worse attentional set - shifting and working memory performance than pd patients without anxiety. in addition, pd patients with anxiety showed impairments in attentional set - shifting and working memory compared to pd patients without anxiety.
Textrank little is known about the relationship between anxiety and cognition in parkinson's disease (pd). we hypothesized that pd patients with anxiety would show impairments in attentional set - shifting and working memory compared to pd patients without anxiety. this study is the first to directly compare cognition between pd patients with and without anxiety. we found that pd patients with anxiety were more impaired on the trail making test part b which assessed attentional set - shifting, on both digit span tests which assessed working memory and attention, and to a lesser extent on the logical memory test which assessed memory and new verbal learning compared to pd patients without anxiety. taken together, these findings suggest that anxiety in pd may reduce processing capacity and impair processing efficiency, especially in the central executive and attentional systems of working memory in a similar way as seen in young healthy adults.
SBERT + NSearch little is known about the relationship between anxiety and cognition in parkinson's disease (pd). we hypothesized that pd patients with anxiety would show impairments in attentional set - shifting and working memory compared to pd patients without anxiety. seventeen pd patients with anxiety and thirty - three pd patients without anxiety were included in this study. results confirmed our hypothesis that anxiety negatively influences attentional set - shifting and working memory in pd. more specifically, pd patients with anxiety were more impaired on the trail making test part b which assessed attentional set - shifting, on both digit span tests which assessed working memory and attention, and to a lesser extent on the logical memory test which assessed memory and new verbal learning compared to pd patients without anxiety. taken together, these findings suggest that anxiety in pd may reduce processing capacity and impair processing efficiency, especially in the central executive and attentional systems of working memory in a similar way as seen in young healthy adults.

Table 3: Samples of the abstracted summaries by the three approaches.

8 Conclusion

The project has developed and shown a method of obtaining an extractive summary before a subsequent abstractive summary by a language model. The extractive summary involving SBERT + NSearch is based on the selection of the most semantically relevant sentences using nearest neighbourhood search and sentence embeddings. Results show that the improvements of using an initial Textrank or SBERT + NSearch extractive summarisation are minimal. A possible improvement to the method would be to increase the number of paragraphs to two or more disjoint ones depending on the size of the given text.

9 References

[1] Papers With Code. *Pubmed Benchmark (Text Summarization)*. URL: `https://paperswithcode.com/sota/text-summarization-on-pubmed-1`.

[2] Alexios Gidiotis and Grigorios Tsoumakas. "A Divide-and-Conquer Approach to the Summarization of Academic Articles". In: *CoRR* abs/2004.06190 (2020). arXiv: `2004.06190`. URL: `https://arxiv.org/abs/2004.06190`.

[3] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 404–411. URL: `https://aclanthology.org/W04-3252`.

[4] Jingqing Zhang et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *CoRR* abs/1912.08777 (2019). arXiv: `1912.08777`. URL: `http://arxiv.org/abs/1912.08777`.