Accelerating Patent Research Using Advanced Text Summarization Techniques

Saathyak Rao Kasuganti

Lokesh Katuri

Satish Reddy Bathula

Aditya Sirangi

MPS in Data Science, University of Maryland, Baltimore County

DATA 690: Special Topics in Data Science: Introduction to NLP

Dr. Antonio Diana

May 12, 2023

Abstract

Legal documents are difficult for most people to grasp due to their complexity and technical language. The text summarization methods can help in overcoming this issue by generating concise and summaries of legal documents. In this study, we investigate the effectiveness of transformer-based models for summarizing legal documents using the Big Patents dataset provided by Hugging Face. We evaluate a pre-trained transformer model, Facebook's BART, and compare its performance with traditional summarization techniques including GPT-2, BERT along with extractive methods like Sumy and TextRank. However, despite the accuracy of the generated summaries, our evaluation does not yield high ROUGE scores. The performance of Bert is similar, as a result, we also analyze the reasons behind this and explore an alternative approach to evaluating technical documents for abstractive summarization. Our findings emphasize the potential of transformer-based models in accelerating patent research by generating more effective summaries of legal documents.

Introduction:

Patent research is a critical task for any patent law attorneys and professionals. Efficient techniques are needed to improve accuracy, reduce time, and minimize the effort required for summarizing Patent documents. Analyzing technical and legal documentation to extract relevant information is a resource-intensive and time-consuming process. Therefore, the need for the development of more efficient methods is crucial to streamline and expedite the summarization of Patent documents. Some machine learning techniques have shown potential in accelerating patent research by extracting essential information from patent documents automatically. In this study,

we aim to investigate the effectiveness of the Facebook BART model for text summarization in patent research, using the Big Patents dataset.

The research questions that this study will address are:

1. Can the Facebook BART model effectively summarize patent documents and improve the accuracy of patent research when using the Big Patents dataset?
2. Can the Facebook BART model reduce the time and effort required for patent research compared to traditional text summarization techniques when using the Big Patents dataset?
3. How does the performance of the Facebook BART model compare to other techniques such as BERT and GPT-2 in patent research when using the Big Patents dataset?

The hypothesis of our study is that the use of the Facebook BART model for text summarization in patent data will lead to a significant improvement in accuracy and a reduction in the time and effort required by patent law attorneys and professionals when using the Big Patent dataset. By answering these research questions, we seek for gaining insights into the robustness of the Facebook BART model for patent research and its implications for patent law professionals and organizations.

Literature Review:

Several studies have explored the application of transformer models for text summarization in patent data. For instance, a study by (Moreno, 2023), proposed a transformers-based abstractive

summarization approach for generating patent claims. The authors achieved promising results, with their proposed model outperforming several baseline models. Similarly, (Pilault, Li, Subramanian, & Pal, 2020), investigated the performance of extractive and abstractive neural document summarization with transformer language models on the patent domain. The authors showed that their proposed model outperformed state-of-the-art models in terms of both ROUGE scores and human evaluation.

Another relevant study is by (Gustafsson, 2020), who proposed a method for automatic text summarization of patent documents using a deep learning model. The authors demonstrated that their proposed model achieved high performance on a benchmark dataset, outperforming several other models. Additionally, a study by (Sharma, Li, & Lu, 2019), introduced a large-scale dataset for abstractive and coherent summarization of patent documents, named Big Patents. The authors showed that their proposed model has the potential for future research for constructing robust systems capable of generating abstractive text summaries that are coherent.

Lastly, a study by (Furniturewala, Jain, Kumari, & Sharma, 2021) proposed the use of joint text features and transformer models for legal text classification and summarization. The authors used LEGAL_BERT and BERT and demonstrated that the model that was trained on a specific domain performed better than the generalized BERT model. This indicates that domain-specific training may greatly improve the performance of the summarization models.

The literature reviewed indicates that transformer models hold great potential for text summarization in the field of patent analysis. Several studies cited in this review demonstrate that transformer models have outperformed several baseline models and achieved state-of-the-art results on benchmark datasets. These findings support the need to investigate the use of

transformer models for text summarization in patent data and test the hypothesis that their use can significantly improve accuracy. However, it should be noted that the complexity and domain-specific language used in patents pose challenges for effective summarization, suggesting that further research is necessary. One potential avenue for future investigation is the use of hybrid transformer-based models that can leverage both extractive and abstractive summarization techniques.

Data Description:

The Big Patents dataset is a valuable resource for training and evaluating text summarization models, containing over 1.3 million US patent documents with full patent text and human-written summaries. It will help us assess the effectiveness of Facebook BART for patent research text summarization and compare it with other methods such as GPT-2 and BERT. In Addition to this, it has diverse collection of patent documents covering different technical areas is an excellent resource for NLP model development and evaluation in patent analysis, such as text classification, information extraction, and summarization.

The patent documents in the dataset are organized by their International Patent Classification (IPC) codes, which classify patents into various technical areas. Each patent includes a detailed description of the invention, publication date, inventor names, assignee, and a human-written short abstractive summary that captures the key information and essence of the patent text. This summary serves as a reference that can be used to evaluate NLP models' performance for text summarization. The dataset is a rich and diverse source of data, providing a

wealth of technical information for NLP tasks and a significant contribution to the field of NLP for patent analysis. (huggingface.co, n.d.)

Table – 1:

Description of dataset and data-types.

Column Name	**Data Type**	**Data Class**
Description	String	Text
Abstract	String	Text

Both the columns consist of string data-type and is of text class.

Table – 2:

Description of Subsets of dataset.

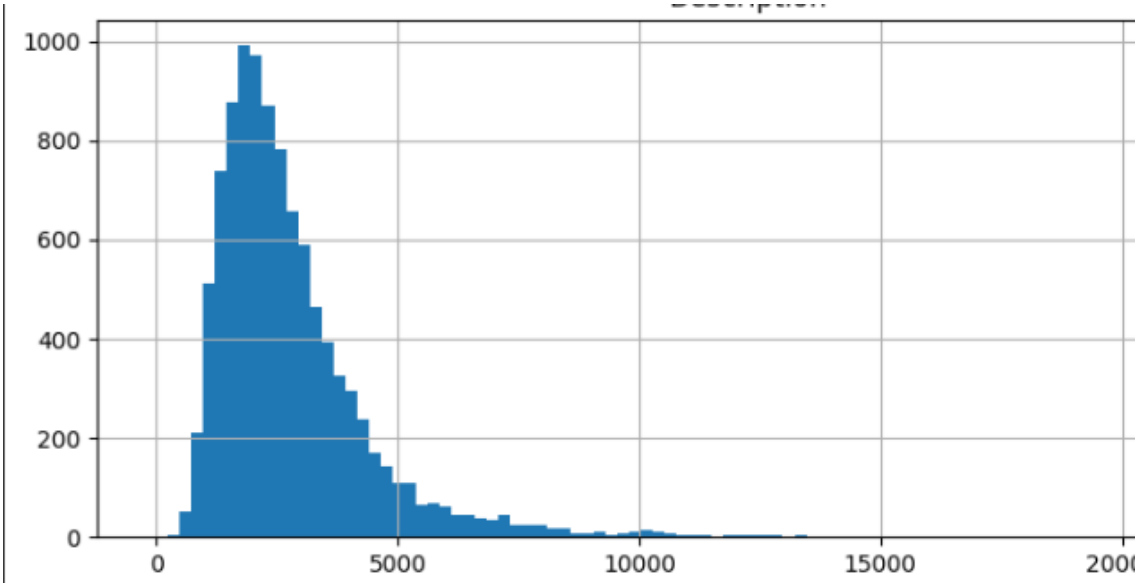
Subset	Domain	Train	Validation	Test
all	Complete Dataset	1207222	67068	67072
a	Human Necessities	174134	9674	9675
b	Operations / Transport	161520	8973	8974
c	Chemistry / Metallurgy	101042	5613	5614
d	Textiles and Paper	10164	565	565
e	Fixed Constructions	34443	1914	1914
f	Mechanical Engineering	85568	4574	4574
g	Physics	258935	14385	14386
h	Electricity	257019	14279	14279
y	General	124397	6911	6911

Exploratory Data Analysis

We performed basic EDA on the data to understand the nature of data, most frequently used words, and their distributions based on size of patent documents and the abstracts.

Figure-1:

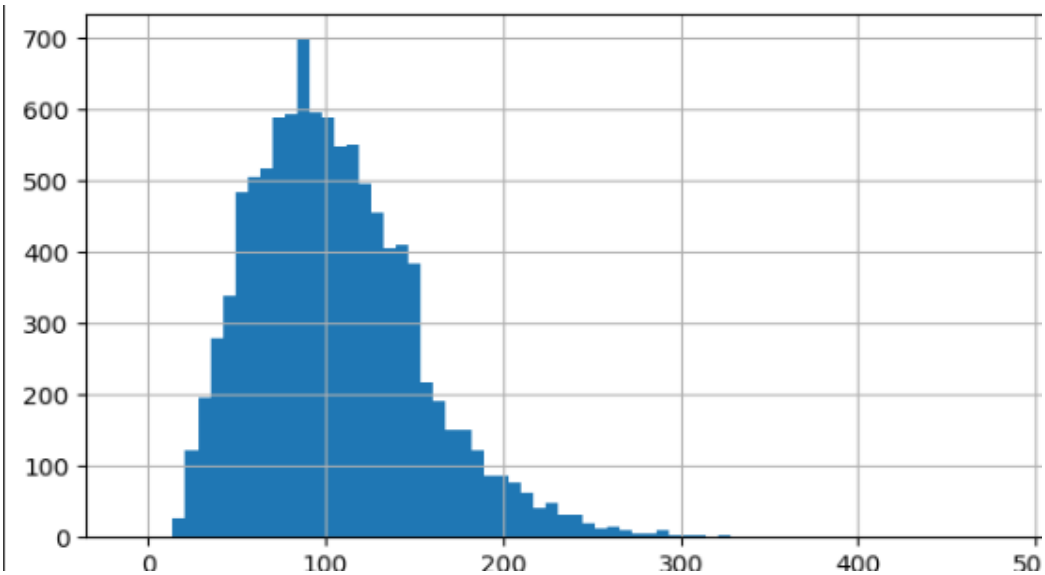
Distribution of lengths of patent descriptions



Note: The distributions of patents are right-skewed indicating that there are more patents with shorter descriptions.

Figure-2:

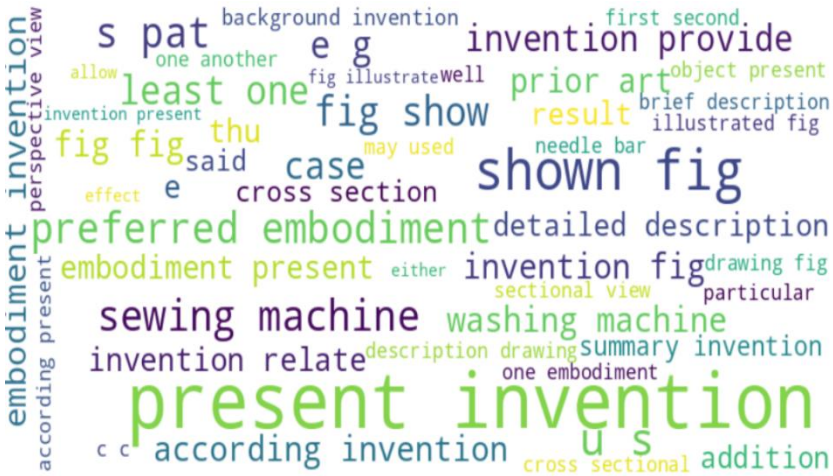
Distribution of lengths of patent summaries:



Note: The distributions of summaries are normally distributed, From the distributions above, we can infer that the summaries have a gaussian distribution with respect to summary length Next, we generated the word cloud of the most frequent words in the patent documents,

Figure -3:

Word Cloud of most frequent words in Patent documents



The most frequent words identified include and are not limited to ‘present invention’, ’shown fig’. In addition to that some technical terminology such as ‘cross sectional’ and ‘perspective view’ are frequently used.

Methodology

BART model

BART is a Denoising Sequence to Sequence pretraining for Natural Language Generation, Translation and Comprehension by Mike Lewis, Yahin Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer on 29 Oct 2019 by Facebook AI Research.

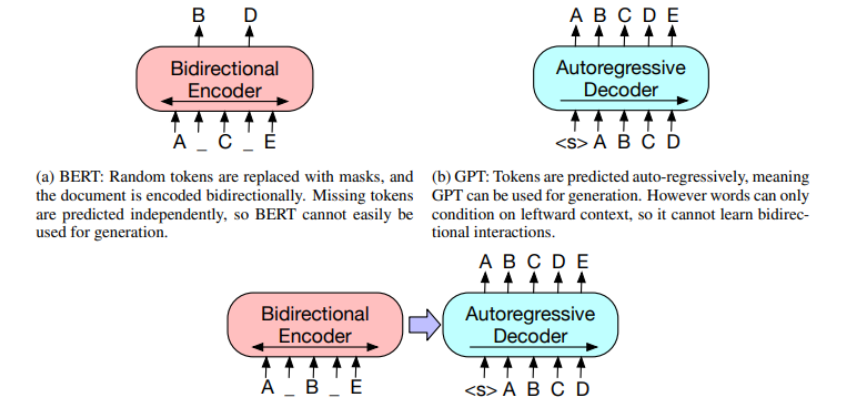
According to the abstract, (Lewis, Liu, Goyal, & Ghazvininejad, BART: Denoising Sequence-to-Sequence Pre-training for Natural, 2020), Bart model is a standard sequence to sequence machine translation architecture with a bidirectional encoder (similar to BERT) and left to right decoder (similar to GPT). The process of pretraining BART involves a couple of key steps. Firstly, it randomly shuffles the order of the original sentences, and secondly, it uses a novel in-filling scheme where certain spans of text are replaced with a special mask token.

BART is a highly effective language model that excels at text generation tasks when fine-tuned, but it also performs well on comprehension tasks. In fact, it has been found to match the performance of RoBERTa on the GLUE and SQuAD datasets, even when trained with comparable resources. Furthermore, BART has been shown to achieve state-of-the-art results on various summarization, dialogue generation, and question answering tasks, with improvements of up to 6 ROUGE scores.

In this paper, (Lewis, Liu, Goyal, & Ghazvininejad, BART: Denoising Sequence-to-Sequence Pre-training for Natural, 2020), the author has presented the model which was pretrained on a model combining Bidirectional and Auto-Regressive Transformers. It uses a denoising autoencoder that is applicable to a wide range of natural language processing tasks. The pre-training process involves two stages: corrupting the text with an arbitrary noising function and then training a sequence-to-sequence model to reconstruct the original text. Despite its simplicity, BART uses a standard transformer-based neural machine translation architecture that generalizes BERT and GPT, among other pre-training schemes (see Figure 1)

Figure 4:

BART Architecture



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

(Devlin, Chang, Lee, & Toutanova, 2019)

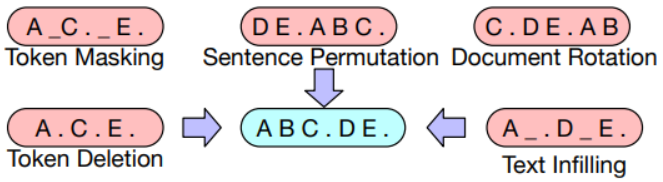
BART adopts the standard sequence-to-sequence Transformer architecture introduced in (Vaswani, et al., 2017), with modifications inspired by GPT. Specifically, the ReLU activation functions in the original architecture are replaced with GeLUs (Hendrycks & Gimpel, 2020) and the parameters are initialized from N (0, 0.02). The basic model has 6 layers in both the encoder and decoder, while the large model has 12 layers in each. Compared to BERT, the decoder layers in BART perform cross-attention over the final hidden layer of the encoder, similar to the transformer sequence-to-sequence model. However, BART does not use an additional feed-forward network before word prediction. As a result, BART has approximately 10% more parameters than a similarly sized BERT model.

Pretraining of BART

BART is trained using a denoising autoencoder approach where the input text is corrupted with an arbitrary noising function and the model is trained to reconstruct the original text. The reconstruction loss is optimized by minimizing the cross-entropy between the decoder's output and the original document. One advantage of BART is that it allows for flexibility in the type of document corruption that can be applied during training, which can help improve its performance on downstream tasks. The paper (Lewis, Liu, Goyal, & Ghazvininejad, BART: Denoising Sequence-to-Sequence Pre-training for Natural, 2020) outlines several previously proposed and novel transformations that were experimented with during training. Examples of these transformations can be found in Figure 5 of the paper.

Figure 5:

Transformations for noising the input that we experiment with.



(Lewis, Liu, Goyal, & Ghazvininejad, BART: Denoising Sequence-to-Sequence Pre-training for Natural, 2020)

Token masking and deletion

The technique was used in pre-training of transformer-based language models, in BERT (Devlin, Chang, Lee, & Toutanova, 2019). The basic idea is to randomly mask (replace) some of

the tokens in the input sequence with a special [MASK] token, and model is trained to predict the original tokens depending on the context. The same method for BART model. Random tokens are deleted. The model must decide the deleted token and generate the text.

Steps followed to build the model

1. Converting entire text to lower case the
2. Removing extra-spaces and concatenating the text.
3. Use BartTokenizer: uses a byte-level byte-pair-encoding.
4. Next, the tokenized inputs are returned as PyTorch tensors.
5. The corresponding input-ids and attention-mask tensors are retrieved from the tokenized inputs.
6. These are truncated to ensure that they fit the size limits to 512 tokens.
7. The tensors are unqueened to add batch-dimension.
8. Finally, the generate () function is used for generating the abstracts.

Results:

We compared the results of the various text summarization models based on the average ROUGE scores of the descriptions generated. As mentioned above, some of these models are traditionally used for text summarization and remaining are newer transformer-based models.

Table 3:

ROUGE-1 metrics of various models considered.

Model Name	**ROUGE-1 Metric**
Facebook BART	0.122
BERT	0.13
GPT-2	0.2
TextRank	0.36
Sumy	0.25

Based on the comparison of the ROUGE scores, of the 3 primary models, BART, BERT and GPT-2, GPT-2 has the highest ROUGE scores around 0.2. It is followed by, BERT and BART. However, considering the extractive summarization models, the overall highest score is achieved by TextRank with 0.36.

Evaluation through manual annotation:

A sample summary that is has been generated by the BART model, is shared below. The corresponding patent document consists of over 28,000 characters. The BART model we used was able to generate a highly accurate summary-snippet of just 383 characters length. The generated summary is able to capture what the patent document is about. Even though BERT has similar ROUGE scores, more technical keywords are captured by the BART model.

Sample Summary: The present invention relates to weft insertion in a multiple-color air jet loom. It provides a method and an apparatus for inserting different weft threads in and through warp sheds. The weft carrying force remains constant at all times for different wefts threads, and that the speed at which the weft threads are taken through the warp sheds varies from one weft type to another.

Conclusion:

The Facebook BART model was evaluated on the Big Patents dataset for patent document summarization. Despite the much lower ROUGE scores, manual annotation of the model's performance, gives a different result about the performance. The human generates sample summaries included lesser technical keywords, while the opposite holds to for the summaries generated by the BART model. Which to an extent explains the cause for the lower ROUGE metric scores of the model. Another important factor to consider is that some of the other traditional models considered are extractive summarizers that may tend to use similar keywords.

Based on the manual annotation results of the summaries generated by the BART model, we can answer the research questions. Firstly, the results suggest that the model can indeed capture the essence of the input patent-text and generate summaries that capture important technical details.

For the second research question regarding reduction in time and resources, as the BART model is pre-trained on the CNN dataset and general English language, there is no extra computational and time requirement for additional training steps. The nature of the model is

indicative of the potential reduction of time and effort required by patent law attorneys and professionals, thereby improving the efficiency of the patent research process.

The third research question seeks to compare the performance of BART with other models. If the sole metric considered is ROUGE scores, then it may seem like the performance is below-par. However, based on the manual annotation, BART was successful in generating the gist or essence of the entire document.

Finally, future research can explore the use of transfer learning to fine-tune the BART model on a smaller dataset of patent documents specific to a particular industry. This could potentially improve the model's performance for that domain, provided a high-performance distributed GPU is available.

Overall, the results indicate that the BART model has the potential to improve the efficiency and accuracy of patent research. With continued research and development, it may become a valuable tool for patent law attorneys and professionals.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). (p. 4717). Association for Computational Linguistics.

Furniturewala, S., Jain, R., Kumari, V., & Sharma, Y. (2021). Legal Text Classification and Summarization using Transformers and Joint Text Features. *CEUR Workshop Proceedings*, 5.

Gustafsson, E. (2020). Automatic Text Summarization of Patent . 49.

Hendryck, D., & Gimpel, K. (n.d.). GAUSSIAN ERROR LINEAR UNITS (GELUS). 2.

Hendrycks, D., & Gimpel, K. (2020). GAUSSIAN ERROR LINEAR UNITS (GELUS)., (p. 1).

huggingface.co. (n.d.). Retrieved from big_patent: https://huggingface.co/datasets/big_patent

Lewis, M., Liu, Y., Goyal, N., & Ghazvininejad, M. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural. *Association for Computational Linguistics*, 7871.

Lewis, M., Liu, Y., Goyal, N., & Ghazvininejad, M. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural. *58th Annual Meeting of the Association for Computational Linguistics* (p. 7871). Association for Computational Linguistics.

Moreno, S. (2023). Transformers-based Abstractive Summarization for the Generation of Patent Claims. 107-108.

Pilault, J., Li, R., Subramanian, S., & Pal, C. (2020). On Extractive and Abstractive Neural Document Summarization with. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9315.

Sharma, E., Li, C., & Lu, W. (2019). BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *Conference on Neural Information Processing Systems*, (p. 3). Long Beach, CA.