Automated Judicial Case Briefing

NYU Center for Data Science

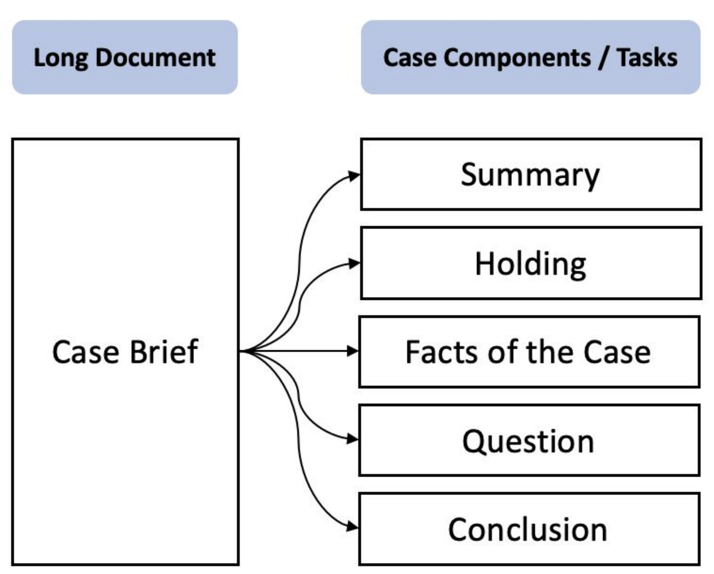
Group Members: Yoobin Cheong, Yeong Koh, Yoon Tae Park

Mentors: Kannan Venkataramanan, Najoung Kim

THE WORLD BANK

Abstract

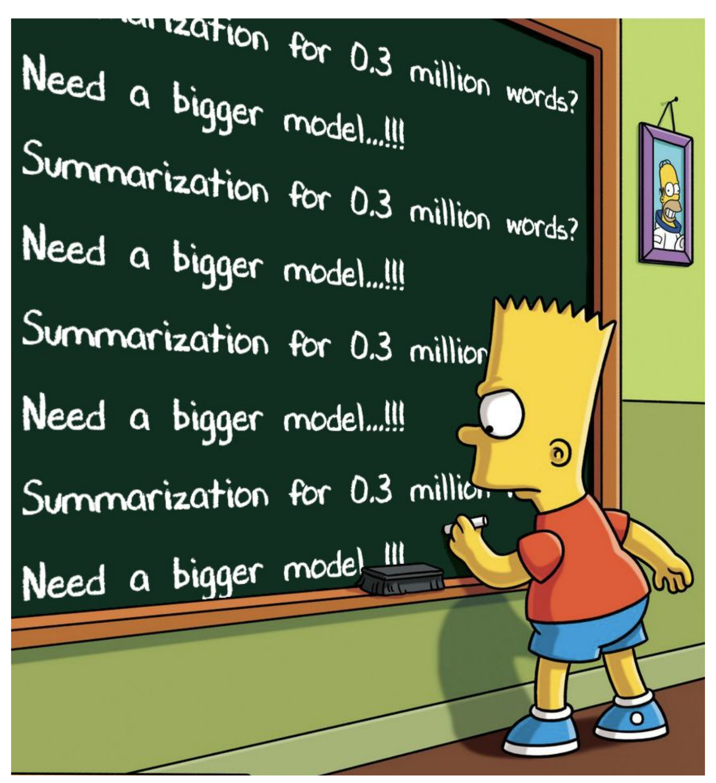
Our project is to build a model that extracts five distinct sections from a judicial case opinion using natural summarization processing techniques such as abstractive summarization methodologies.



We've built a pipeline that uses SOTA models (i.e. BART) that shows Baseline 0.3 and fine-tuning 0.47 on RougeL score, and this will be applied in legal domain of World Bank.

Background

Research Question: Can we create a model that extracts relevant information from extremely long legal documents?



Introduction

Text summarization is the task of generating a smaller coherent version of a document while preserving key information.
To produce a summarization, it requires a lot of reading.

Dataset

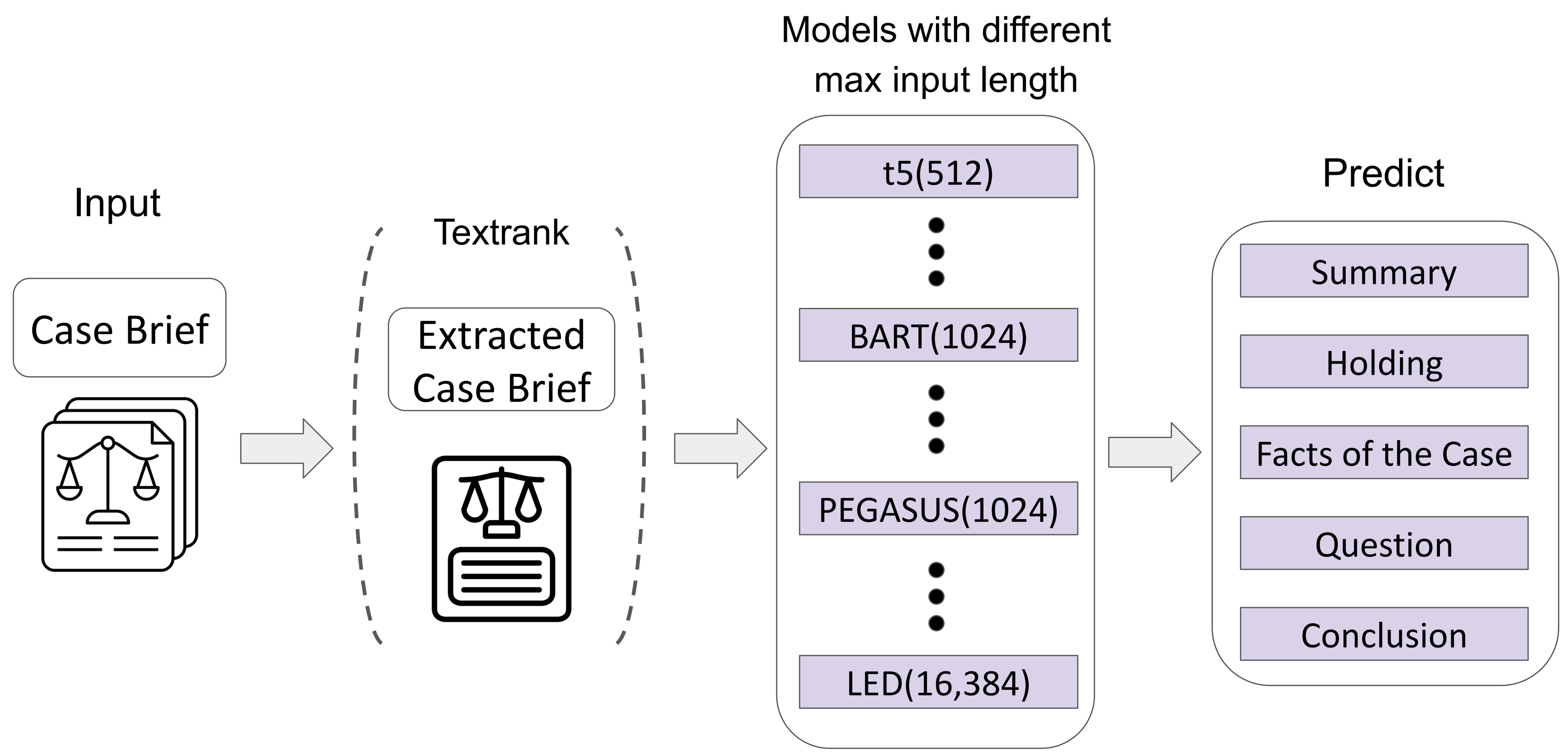
~10k case opinion texts from US Supreme Court

Challenges

1) Extremely long judicial case opinions (0.36+ million words) Current SOTA transformer models cannot cover all sentences.
2) From the same input, each task looks for a summary from different parts or contents of the input.

- Save a lot of human effort
- Help law practitioners refer to multiple cases faster at ease

Methods and Models



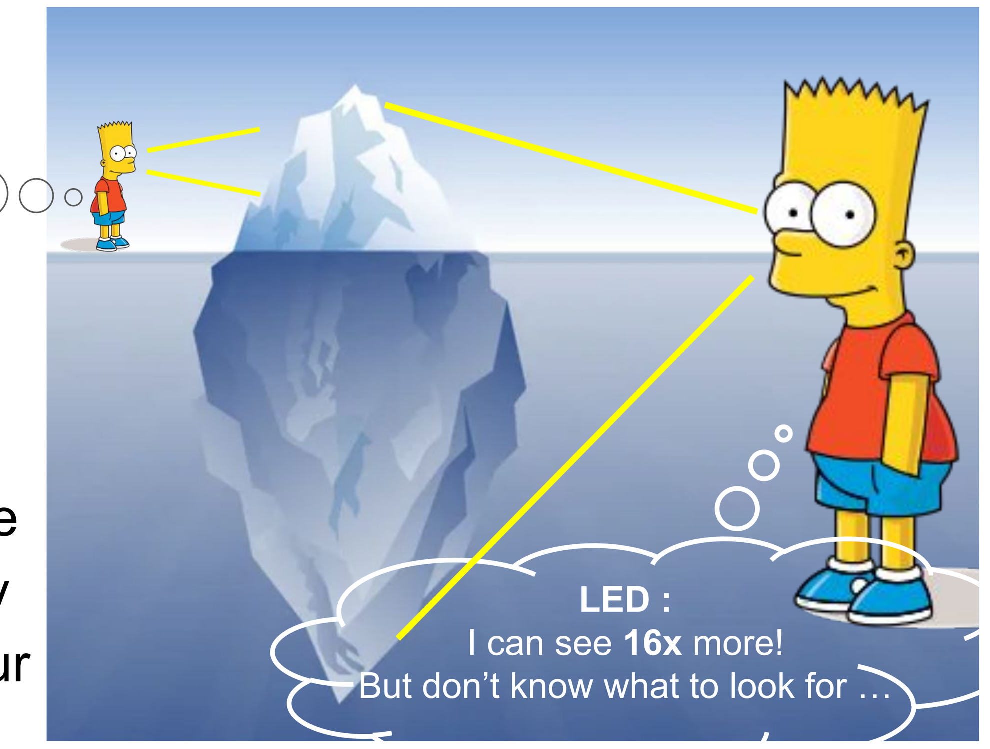
We've built a pipeline that generates 5 different summarizations. Then we experimented with various SOTA models by increasing input length capacities of the models. Also, we applied textrank methodology to extract the document within the model's input length limits.

Results

The table below shows the Rouge-L score based on different models and approaches that were defined previously.

Model (Input Length)	RougeL									
	without Extraction					with Extraction				
	Summary	Holding	Facts	Question	Conclusion	Summary	Holding	Facts	Question	Conclusion
t5-small(512)	0.3280	0.2527	0.2819	0.2696	0.2558	0.1850	0.1976	0.1977	0.2360	0.2367
t5-base(512)	0.3892	0.2611	0.2731	0.2736	0.2599	0.2207	0.2001	0.1848	0.2507	**0.2385**
t5-large(512)	0.1171	0.1187	0.1196	0.1220	0.1242	0.1364	0.1415	0.1429	0.1468	0.1537
bart-base(1024)	0.4375	**0.2972**	0.2861	0.2988	0.2856	0.2320	0.2005	0.2017	0.2568	0.2351
bart-large(1024)	0.4631	0.2942	0.2987	0.2966	**0.2895**	0.2440	0.2138	0.2155	0.2442	0.2375
bart-large-cnn(1024)	0.4698	0.2559	**0.3054**	0.2406	0.2808	**0.2471**	0.1944	**0.2215**	0.1316	0.2360
Pegasus-xsum(512)	0.4619	0.2909	0.3014	**0.3061**	0.2556	0.2336	0.2138	0.2024	**0.2721**	0.2192
Pegasus-large(1024)	**0.4753**	0.2806	0.3024	0.2996	0.2591	0.2372	0.2162	0.2048	0.2626	0.2204
LED(1024)	0.4232	0.2875	0.2697	0.2768	0.2489	0.2023	0.2007	0.1862	0.2490	0.2023
LED(2048)	0.4360	0.2804	0.2795	0.2835	0.2682	0.2107	0.2072	0.1903	0.1481	0.2205
LED(4096)	0.4273	0.2872	0.2893	0.2951	0.2835	0.2064	0.2099	0.1965	0.2514	0.2208
LED(8192)	0.4125	0.2922	0.2884	0.3002	0.2749	0.2065	0.2191	0.1947	0.2570	0.2272
LED(16384)	0.4269	0.2862	0.2718	0.3023	0.2831	0.2350	**0.2309**	0.2030	0.2665	0.2236

Conclusion and Future Work



Although additional experiments were performed to address the underlying input capacity limit that is inherent in our baseline BART model, we can see that using a larger model with some extraction method did not help achieve a better Rouge score for any of our summarization tasks.

However, we cannot guarantee that a high RougeL score from our models' summarization output is directly related to a high quality/accuracy summarization task, as the Rogue score can only calculate the sequences of the overlapping words between the prediction and actual summary that are truncated by the model's max capacity. Therefore,

Bigger Model ≠ High Quality of Summarization

High Rouge-L ≠ High Quality of Summarization

For future work, it can be further improved by having the human annotators validate the model's actual summarization performance. Also, applying other extracting approaches could also improve its performance.

Acknowledgement

We would like to thank our amazing mentors, Kannan Venkataramanan at World Bank and Najoung Kim at NYU, for all their constructive feedbacks and supports throughout the project.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Moro, Gianluca and Ragazzi, Luca, 2022, Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Prad-hiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterer, Rajarshi Das, and Andrew McCallum. 2021. Long document summarization in a low resource setting using pre-trained language models.

Mohamed Elaraby and Diane Litman. 2022. Arglegal-summ: Improving abstractive summarization of legal documents with argument mining.