1 Acknowledgements

I would like to express my gratitude to Professor Hovy for his continued mentorship along this journey and for making this experience enriching and enjoyable. Thank you to all the colleagues who participated in the surveys; your contribution is highly valued and appreciated. Finally, I would like to thank my family for their unwavering support. You are my constant source of inspiration and I could not have done any of this without you.

From Statutes to Summaries: Metrics, Length Constraints, and Model Adaptation in Transformer-Based Legal Text Summarization

Redi Alico

Contents

2 Abstract

This thesis investigates the problem of summarizing long, complex legal texts into concise and accessible summaries using transformer-based models. We provide comprehensive analyses of models' summaries through carefully curated metrics accompanied by human evaluation. We extend previous work, which was focused solely on summarizing text, by fine-tuning models on specific legal text (US congressional bills). Furthermore, we explore a more thorough avenue, looking at longer texts up to 16,000 tokens, and compare results among various input and output token limits to determine its role as a factor in summary quality. Our experiments find that our best legal text summarizer achieves notable average metrics scores such as X ROUGE-1, Y F1 BERTScore, and Z BLANC Help.

However, this improved performance on the metrics comes at a cost in readability, with generated summaries often exceeding recommended grade levels. Although preliminary exploration of simplification techniques suggests that readability can be improved, such approaches fall outside the scope of this thesis and are proposed as avenues for future research.

3 Introduction

- choose a model and evaluate using bleu, rouge and human eval
- Models:
  - nsi319/legal-led-16384, (performance) for english
  - csebuetnlp/mT5-multilingual-XLSum

  Currently using Pegasus xsum since the bigger models are infeasible due to high compute power.

  Could use the multilingual model only if the aim is to make this applicable to a variety of languages.

  Models in mind:

  - A weak but classic baseline (BART-base)
  - A mid-range specialized model (Pegasus)
  - A target strong model (Llama legal instruct/ Legal LED base)

4 Contributions

- Adaptation of transformer-based models for **effective** summarization of long legal texts
- Introduction of a **comprehensive** multi-metric evaluation framework
- **Empirical** analysis of length constraints in summarization
- **Interactive** website incorporating this tool, making it easily accessible to all ??

5 Background ??

6 Related Work

7 Methodology

8 Data

The primary dataset used for this experiment is BillSum[1]. It comprises a series of bills from the US Congress. The dataset is divided into three different splits, train, test, and ca_test. For our purpose, we will disregard ca_test, since it only has a limited number of samples and there is no train counterpart. A fundamental reason why we chose this dataset is the fact that it contains the whole bill and a summary. The latter acts as a reference for our model to produce an independent, well-designed summary for any given legal text.

9 Model

10 Experiments

11 Results

12 Discussion

- acknowledge that FK & Dale-Chall are too high
- propose the idea that a simplifier could decrease score
- from preliminary testing though, the scores barely fall but other metrics like ROUGE tank
- A more thorough analysis is needed but we leave it as future work
- Regarding tests, we expect BART base to be the baseline score to beat, but we acknowledge that for short inputs like 512 tokens, LED models might be lower because they are not tailored to such short texts
- We expect LED models to shine in longer texts, potentially up to 16,000 input tokens, whereas the three other models cannot even run those inputs

13 Conclusion

14 Future Work

15 Appendix

15.1 Model Evaluation Results - Consolidated Table

Model	Config	FK	DC	R1	R2	RL	BERT	BLANC
BART base	512/128	35.83	14.37	0.44	0.25	0.33	0.82	0.15
Pegasus-xsum	512/128	40.97	15.05	0.47	0.29	0.37	0.83	0.17
Legal LED base	512/128	37.13	14.67	0.43	0.24	0.33	0.81	0.15
Legal pegasus	512/128	42.04	15.20	0.48	0.29	0.37	0.83	0.18
Legal LED large	512/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
BART base	512/256	41.92	15.10	0.45	0.25	0.33	0.82	0.15
Pegasus-xsum	512/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	512/256	54.46	16.80	0.43	0.23	0.33	0.81	0.15
Legal pegasus	512/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	512/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
BART base	1024/128	38.56	14.63	0.46	0.28	0.35	0.83	0.17
Pegasus-xsum	1024/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	1024/128	39.72	14.89	0.46	0.28	0.36	0.83	0.17
Legal pegasus	1024/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	1024/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
BART base	1024/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Pegasus-xsum	1024/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	1024/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal pegasus	1024/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	1024/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Pegasus-xsum	2048/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	2048/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal pegasus	2048/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	2048/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Pegasus-xsum	2048/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	2048/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal pegasus	2048/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	2048/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Pegasus-xsum	4096/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	4096/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal pegasus	4096/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	4096/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Pegasus-xsum	4096/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	4096/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal pegasus	4096/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA

Continued on next page

Model	Config	FK	DC	R1	R2	RL	BERT	BLANC
Legal LED large	4096/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	8192/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	8192/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	8192/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	8192/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	16384/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	16384/128	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED base	16384/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA
Legal LED large	16384/256	TBA	TBA	TBA	TBA	TBA	TBA	TBA

15.2 Summary Notes

- **Configuration Format:** All models use 3 epochs, shown as input_length/output_length
- **Column Abbreviations:** FK=Flesch-Kincaid Grade, DC=Dale-Chall Score, R1/R2/RL=ROUGE-1/2/L, BERT=BERTScore F1, BLANC=BLANC Help Score
- **TBA:** Results to be added
- **Readability:** Higher FK and DC scores indicate more complex text
- **ROUGE/BERT/BLANC:** Range 0-1, higher is better

References

[1] Anastassia Kornilova and Vladimir Eidelman. “BillSum: A Corpus for Automatic Summarization of US Legislation”. In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Ed. by Lu Wang et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 48–56. DOI: 10.18653/v1/D19-5406. arXiv: 1910.00523 [cs.CL]. URL: https://aclanthology.org/D19-5406.