Congressional Bill Summarization with Transformers

Abstract

Document summarization specific to congressional bills presents a set of unique challenges and opportunities. The BillSum project catalyzed efforts to introduce SOTA methods to this emerging field. Utilizing a combination of transformer models (HuggingFace T5), this paper summarizes efforts to improve upon the results observed by BillSum. The objective is to develop methods that can effectively summarize regulation to support understanding and compliance. Applying T5 models specifically to this problem is a different approach from those known to exist at the time of this writing.

1 Introduction & Background

This paper attempts to advance the research on document summarization specific to congressional legislative documents. The goal of summarization specific to regulatory and legislative documents is in part to support compliance. The sheer volume of regulations is difficult if not impossible for individuals to maintain comprehensive awareness. Thousands of pages of dense documents are produced each year by Congress, presenting a seemingly overwhelming corpus. We can therefore improve regulatory understanding, compliance, and ultimately safety by providing a simple tool for distilling the key elements of relevant legislation.

Research on document summarization has been performed across a variety of datasets, including news articles and scientific papers. However, there is limited extant research on legislative document summarization. The BillSum project (Kornilova & Eidelman, 2019) was the first to examine summarization in the context of US government legislation. This project also produced the only text corpus specific to this area. By design, the BillSum paper appeals to further investigation and motivates research.

This paper will apply SOTA pre-trained and custom-trained models to the BillSum text corpus, attempting to build on the progress made thus far in this area. Models will draw from the HuggingFace project's API ("HF", Wolf et al, 2019). This paper will also focus on examining the appropriate scoring methodologies.

2 Dataset Exploration

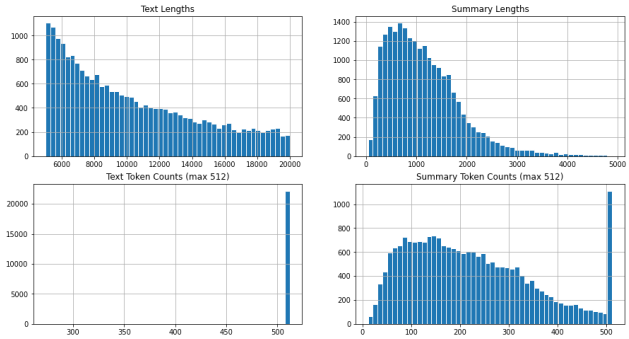
The text corpus was compiled initially by the BillSum project. As detailed in the paper: *"The BillSum dataset consists of three parts: US training bills, US test bills and California test bills. The US bills were collected from the Govinfo service provided by the United States Government Publishing Office. Our corpus consists of bills from the 103rd-115th (1993-2018) sessions of Congress. The data was split into 18,949 train bills and 3,269 test bills."* The data is publicly available on the BillSum GitHub page *here*, Kaggle, and as a TensorFlow dataset.

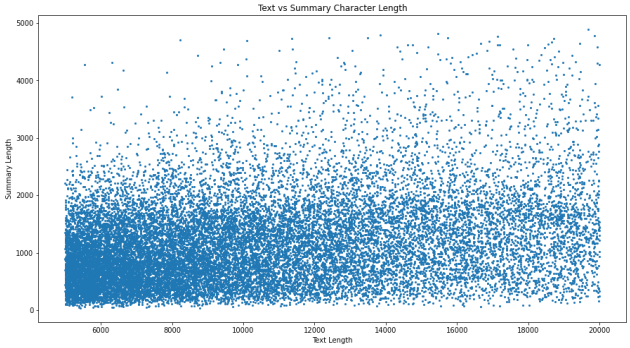
The corpus primarily consists of mid-range legislative bills ranging from 5,000 to 20,000 characters. For purposes of implementing the HF transformer models, word count is more relevant due to tokenization. Figure 1 illustrates corpus size in terms of both character and word (token) count. This paper focuses on the BillSum train and test datasets for US Congressional bills and excludes the California bill dataset, which was used as a test set by the original BillSum paper.

The large size of both documents and summaries present the biggest challenge to building an adequate model. In some summarization research (e.g., Li et al, 2018), the text being summarized is often shorter than the *summaries* included in the BillSum corpus. Additionally, the

documents often include technical language which likely is not recognized by pre-trained tokenizers but relevant to extractive summarization. Overall, harnessing adequate computing power is a key challenge.

Figure 1: Document Lengths





3 Current SOTA Methods

To make evaluations, ROUGE-1, ROUGE-2, and ROUGE-L are calculated (Chin-Yew, Lin, 2004). Each is informative as to the relative closeness of the candidate summary to the reference. The Rouge F-score is most holistic in illustrating the combination of relative completeness and brevity.

The BillSum paper, which at the time of this writing likely remains the SOTA on congressional bill summarization, focuses on extractive methods, leaving abstractive to future research. Ultimately, the SOTA method utilized a combination of Bert-Large Uncased model on the "next sentence prediction" and document context models. Ensemble selection was conducted using MMR (Goldstein et al., 2000). The upper bound (oracle) model applies MMR selection using the Rouge-2 (precision) scores. These results are compared to the baseline and experimental results in the next section.

4 Results

This section summarizes the results of both the baseline models and experimentation independently performed. Ultimately, the goal was to improve on the results of the BillSum project - I was unable to successfully match or improve upon these results. The "Conclusions" section provides explanations and thoughts for next steps.

4.1 Baseline Model

To establish a strong baseline model, I utilized the HuggingFace pretrained sequence-to-sequence transformer models (*TFT5ForConditionalGeneration*). Specifically, I implemented the "t5-small" and "t5-base" with different parameters like beams, temperature, and minimum and maximum length. The optimal baseline used the pretrained "t5-small" with 5 beams and temperature of 1.

Figure 2: ROUGE F-scores

	Rouge-1	Rouge-2	Rouge-L
Oracle	45.11	28.74	37.38
DOC	38.51	21.38	31.49
SUM	40.69	23.88	33.65
DOC + SUM	40.80	23.83	33.73
Baseline t5**	34.10	15.51	31.05

**Congressional Bills, test dataset

Since this was a baseline, I used a relatively simple methodology to decompose text entries into shortened components that were then passed into the pretrained model with summary outputs combined at the backend. Modest

cleanup was performed on the output summaries to improve grammatical correctness and readability. The results of this model are provided below, with comparison to the BillSum results for reference to the SOTA.

4.2 Experimentation

The baseline model's summaries, even with subsequent clean-up of punctuation and grammatical errors, are not generally presented in understandable for a human reader. Figure 3 illustrates this issue.

Figure 3: Prediction Example from Baseline

Reference Summary

Defense of Environment and Property Act of 2015 This bill amends the Federal Water Pollution Control Act to specify the types of water bodies that are quot, navigable watersquot. And therefore fall under the scope of the Act. Groundwater is considered to be state water. Activities carried out by the Environmental Protection Agency (EPA) or the US Army Corps of Engineers may not impinge upon states' power over land and water use. States or individual property owners may obtain judicial review of jurisdictional determinations by the EPA or the Army Corps that would affect their ability to plan the development and use of land and water resources within 30 days after a determination. The EPA may not use a significant nexus test to determine federal jurisdiction over navigable waters and waters of the United States. The following rules are nullified: (1) the Army Corps' rule entitled quot, Final Rule for Regulatory Programs of the Corps of Engineers,quot, (2) the EPA's proposed rule entitled quot. Advance Notice of Proposed Rulemaking on the Clean Water Act Regulatory Definition of 'Waters of the United States,'quot, (3) the guidance document entitled quot. Clean Water Act Jurisdiction Following the US Supreme Court's Decision in Rapanos v. United States amp, Carabell v. United Statesquot. and (4) any subsequent regulation or guidance issued by federal agencies that defines or interprets the term quot, navigable waters. quot. The Army Corps and the EPA may not promulgate rules or issue guidance that expands or interprets the definition of navigable waters unless expressly authorized by Congress. Federal agencies must obtain the consent of private property owners prior to entering their land to collect information about navigable waters. Federal agencies that issue regulations that relate to the definition of navigable waters or waters of the United States and diminish the fair market value or economic viability of a property must pay the affected property owner an amount equal to twice the value of the loss. Until the landowners have been compensated, such a regulation will have no force or effect.

Baseline Prediction

this Act may be cited as the "Defense of Environment and Property Act of 2015" the term 'navigable waters' does not include waters that do not physically abut waters described in subparagraph (A).a jurisdictional determination by the Administrator or the Secretary of the Army would affect the ability of a State or individual property owner to plan the development and use of land and water resources. the state or individual property owner may obtain expedited judicial review not later than 30 days after the date on which the determination is made. the Administrator may not use a significant nexus test to determine federal jurisdiction over navigable waters and waters of the united states. the final rule of the Corps of Engineers entitled "Final Rule for Regulatory Program" shall have no force or effect. any subsequent regulation of or guidance issued by any federal agency defines or interprets the term "navigable waters" or "waters of the united states" the secretary of the army, acting through the Chief of Engineers, and the Administrator of the Environmental Protection Agency shall not promulgate any rules or guidance that expands or interprets the definition of navigable waters unless expressly authorized by Congress.</s>landowner is given access to any data collected from the entry. if a federal regulation relating to the definition of navigable waters or waters of the united states diminishes the fair market value or economic viability of a property. the federal agency issuing the regulation shall pay the affected property owner an amount equal to twice the value of the loss.

Experimentation was focused on improving both the Rouge scores and human readability of model outputs. I ran several experiments, unfortunately all with limited success. The remainder of this section outlines these experiments.

Customizing the pretrained HuggingFace models was a focus. I tried using the API to custom-train and effectively place an architecture on top of the baseline t5 model. Since the text entries in the corpus are all significantly longer than the 512 tokens permissible for the pretrained "t5-small" model, I used model outputs as inputs to the customized model, effectively reflecting a layered transformer architecture. I was unable to employ this model effectively. In addition to significant compute time, the results were materially worse than the baseline.

I also tried using the ROUGE metric as a mechanism to reduce the length of individual text elements of the corpus prior to application of transformer models. Specifically, I would apply the ROUGE metric to each sentence in a text element (i.e., to identify those sentences most important to a text). Ranking according to ROUGE, I compiled sentences into a text subset up to the maximum 512 token length. This was remarkably unsuccessful, primarily due to the unique nature of bill texts. Since most bills have sentence lengths which vary significantly, longer sentences will naturally have higher ROUGE scores (levels of importance). This does not necessarily translate into an effective abbreviation mechanism for the text. Running the results of this mechanism produced results which were much worse than the baseline.

Lastly, I tried running the outputs of the baseline model through another transformer model to improve readability. This also did not produce favorable results relative to the baseline.

5 Conclusions

Larger models that can consume larger text datasets like Big Bird (Zaheer et al, 2020) would likely improve results. A more effective

methodology to layer transformer layers with the necessary compute capacity would also likely enhance results. The method of evaluation should also be reexamined.

The scoring mechanism should be key input and based on objective of person utilizing the summary. A person may want a more abstractive summary that's shorter and less technical versus someone looking to identify key technical elements of the rule which would imply a longer more extractive summary. Quite simply, judgement is required to develop the targets. Refining the summaries according to target audience would lead to more usable models and outputs.

References

Anastassia Kornilova, Vlad Eidelman. 2019. *BillSum: A Corpus for Automatic Summarization of US Legislation.*

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. 2019. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*.

Wei Li, Xinyan Xiao, Yajuan Lyu, Yuanzhuo Wang. 2018. *Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out.*

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. *In Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLPAutoSum '00, pages 40–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed. 2020. *Big Bird: Transformers for Longer Sequences*