Exploring Extractive and Hybrid Summarization on US Legal Documents

Proposal

With our 266 NLP Final Project, we aim to develop a summarization system for U.S. and California legislative bills using the BillSum dataset (~22,000 examples of bills with human-written summaries). Legislative documents are typically very lengthy, formal, and structurally complex documents that are hard to interpret, and creating a summarization for these documents can be helpful in deciphering these documents at face value. An effective summary should be able to improve accessibility for policymakers, legal researchers, and the general public by simplifying complex bills into concise representations of what they stand for. We plan to explore extractive methods like BERT-based sentence scoring models and cluster-based extraction, and hybrid summarization techniques like maximum marginal relevance (MMR) and pointer generator networks. We will look at the available pre-trained models on websites like Hugging Face to fine tune to the BillSum dataset before deciding if we want to attempt training our own model for this task.

We plan on comparing our models against the human-written summaries using ROUGE and BERTScore, balancing faithfulness to the facts and readability. Ultimately, we hope our work will contribute to more interpretable summarization systems for legal frameworks.

References:

- A sentence is known by the company it keeps: Improving Legal Document Summarization Using Deep Clustering (AI & Law, 2023)
- Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions (Zhong & Litman, NLLP 2022)
- Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance (Agarwal et al., Findings 2022)
- BillSum: A Corpus for Automatic Summarization of US Legislation (Kornilova & Eidelman 2019)