LegalText+: Advancing Legal Documents

Elizabeth Kim[1] **Jesse Woo**[1] **Lawrence Leung**[1] **Nikhil Ghosh**[1]
[1]Columbia University
{ek2935,jw4202,ls12162,nrg2156}@columbia.edu

Abstract

We propose to collect a dataset of legal arguments extracted from court briefs filed in U.S. Supreme Court cases. The legal arguments will represent an explicit human chain of thought reasoning for legal NLP, which is an area that is generally lacking in useful training data. We intend to use a model-assisted approach to extract legal arguments from the section headings of court briefs, because by convention these headings are a semi-structured summary of a legal argument. By providing a large, high-quality legal NLP dataset, work could open up entirely new lines of research in the fields of legal NLP and human reasoning.

1 Background

As ChatGPT has shown, large language models (LLMs) have become quite capable of generating text that have both high fluency and faithfulness, but they sometimes fail to operate in a way that aligns with human reasoning or chains of thought. While human reasoning is often implicit in an NLP corpus, this dataset would be a source of explicit human reasoning on a specific task: legal argumentation. The convention of legal brief writing is to have the heading of each section and subsection state a concise version of the argument that would be contained in that section, with each section building toward the ultimate argument. So in a good brief, the table of contents ought to serve as a fairly structured form of human legal reasoning.

There are a few existing datasets related to law and NLP, though none that explicitly capture structured legal reasoning or chain of thought. The CaseHOLD [5] dataset is a short excerpt from a judicial opinion paired with a set of multiple choice answers which purport to represent the holding of the case. A holding is a short summary of the case's rule and facts that forms the core of the precedential value of the case. LexGLUE [1] is a law-themed General Language Understanding Evaluation benchmark. It consists of seven datasets/tasks: the European Court of Human Rights cases tasks A and B, the Supreme Court of the U.S. (SCOTUS) cases, European Union Legislative text, American contract provisions, Terms of Service from several large web platforms, and the CaseHOLD dataset. The Caselaw Access Project offers free, programmatic access to millions of U.S. cases from several federal courts and all 50 states. Zhong et. al published JEC-QA, a question/answering dataset based on the National Judicial Examination of China.

Huggingface has a dataset called pile-of-law [2] that purports to include some briefs filed in federal courts from the RECAP archive of the Free Law Project. While the RECAP archive does not include Supreme Court cases, the convenience of having access to a database of this size and quality may make this worth using. One previous chain-of-thought rationale dataset is the CoT Collection [3]. Building upon 9 previous NLP datasets, this dataset comprises 1060 NLP tasks/prompts with 1.88 million CoT rationales. The format of the CoT collection is primarily three chunks of text starting with a source, a target, and a rationale on how to reason from the source to the target. For example, on a math related task, the source will have a word problem followed by a question, the target would be a numerical number, and the rationale are the steps taken to derive the solution. Another related dataset is LogiCoT [4] , where researchers gave GPT-4 a set of instructions (usually request a derivation or reasoning) and an input (a problem with a solution). From here, GPT-4 returned a reasoning blurb.

Preprint. Under review.

This instruction tuning dataset is similar to this project proposal, but the domains do not overlap as LogiCoT is not based on legal reasoning.

2 Format

The inputs will be legal briefs, which are the written arguments filed by the parties in a lawsuit. We may also use amicus briefs, which are supplemental briefs filed by groups or people who are not themselves litigants, but have some interest in the outcome of the case. Amicus briefs are most common in Supreme Court cases. The federal government maintains a database of all documents filed in federal courts called PACER, but access is behind a paywall. Certain library sources like LexisNexis, Westlaw, Bloomberg Law, and ProQuest also maintain databases of court briefs, but accessing programmatically may violate terms of service. Some law schools and law libraries maintain databases of briefs, especially Supreme Court briefs. Finally, the SCOTUS blog has briefs of all cases decided on the merits as far back as 2007. We also plan to consult the Columbia law librarians.

This task will require more exploration, but we will likely rely on briefs of Supreme Court cases for two main reasons. First, SCOTUS briefs are the most widely available. Second, the legal reasoning in the briefs are likely to be high quality, as only the most accomplished attorneys are admitted to practice in front of the Supreme Court.

The output will be a set of legal arguments that are represented by the brief's section headings. At a minimum we would want to annotate the body of the argument (i.e. the individual aspects that build toward the conclusion), and the conclusion itself. If time and capacity allows, we might mask the conclusions or individual aspects in the argument body to make the dataset more appealing to researchers, who are likely to use masking when training models.

3 Data Collection

Although we intend to use a model in the loop to scale extraction and annotation of the legal arguments, we will also need to use at least some human annotators. Human annotated data could bootstrap a model to collect data at scale.

Annotators would be instructed to record and annotate the section headings as well as the ultimate argument in the conclusion. Conclusions would be labeled as one of a few different options: affirm, reverse, remand, affirm in part, reverse in part. Identifying conclusions may require additional practice or instruction but should be within the capabilities of a Columbia student.

Ideally we will use a model-assisted approach, where a model can automatically extract arguments from the table of contents or section headers. A ML model will be necessary because while there are some conventions as to the structure of legal briefs, there is no strictly enforced format that would allow for simpler approaches. For example, some briefs do not state the conclusion explicitly in the header, but simply label it "conclusion." A sufficiently advanced model such as GPT 3 or 4 may be able to handle this task on its own, but we will need to test this to be sure. The context window of the model could also present a challenge, as these can be very long documents.

OpenAI's API and larger open source model such as Llama2 will be used as a zero-shot generation of these legal arguments. Fine-tuning an open source LLM on legal data might be another option for improved performance. However, since we have access to a single 3090 Ti GPU with 24 GB of VRAM, fine-tuning might prove too resource intensive for a single GPU. Thus, fine-tuning might be limited to smaller models (30b parameters and below). Nevertheless, this project does not absolutely require the fine-tuning of any models. Fine-tuning and frameworks would just be one possible avenue for better legal argument generation. At some point in the pipeline we may also need to use human annotators to evaluate the model's performance.

4 Analysis

When evaluating the extracted headings from the sections or the tables of contents, we simply need to see that the two match. This may be done automatically. The legal conclusions will be chosen from a small, fixed set of options which should be found in the text of the brief as well.

5 Impact

As discussed above, this dataset will be novel in the legal NLP and chain-of-thought reasoning literature. Our hope is that this dataset will allow researchers to train models with better performance on reasoning tasks, as well as improve post-hoc explainability. There is relatively little work on legal NLP in general and even less so on legal reasoning, in part because there are few good datasets. This work could open up entirely new lines of research in the fields of legal NLP and human reasoning.

6 Limitations

While sourcing only or primarily from SCOTUS cases has certain benefits as discussed above, it will also have some limitations that are worth considering. SCOTUS cases represent only a small subset of the overall cases that are litigated in the United States. They also tend to skew toward certain topics in the law, such as administrative or criminal law. Further, some areas of law or legal reasoning will not be captured by litigation, e.g. transactional law. There are existing cultural and demographic biases in the legal field that may be reflected in the dataset. Lawyers at the highest level skew affluent, white, and male, and their reasoning may reflect the biases of these demographic classes.

There may also be limitations with our approach of using a model in the loop to scale data extraction. We will need a robust process to detect errors and address the fact that briefs are only semi-structured.

In addition, because we are limiting our labels of the conclusions to a limited set of outcomes (affirm, reverse, remand, affirm-in-part/reverse-in-part), we will lose some nuance of the legal reasoning. However we think labeling like this will improve our annotation process significantly enough to make it worth it. Requiring more sophisticated extraction or labeling from the document would be beyond the capabilities of annotators without legal training.

References

[1] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras. Lexglue: A benchmark dataset for legal language understanding in english, 2022. URL https://arxiv.org/abs/2110.00976.

[2] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset, 2022. URL https://openreview.net/forum?id=3HCT3xfNm9r.

[3] S. Kim, S. J. Joo, D. Kim, J. Jang, S. Ye, J. Shin, and M. Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning, 2023. URL https://arxiv.org/abs/2305.14045.

[4] H. Liu, Z. Teng, L. Cui, C. Zhang, Q. Zhou, and Y. Zhang. Logicot: Logical chain-of-thought instruction-tuning data collection with gpt-4, 2023. URL https://arxiv.org/abs/2305.12147.

[5] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *CoRR*, abs/2104.08671, 2021. URL https://arxiv.org/abs/2104.08671.