DL4DS Legal Contract Dataset

Ethan Chang, Heng Chang, Josh Yip

April, 2024

Abstract

This project aims to develop a high-quality, labeled dataset of legal contract clauses to support AI-driven contract analysis. We introduce a hybrid clause segmentation pipeline that merges regex-based heuristics, semantic similarity using sentence embeddings, and supervised boundary classification to accurately detect clause boundaries in unstructured legal text. Building on this foundation, we assign contract-type-specific clause labels and explore intra-contract relationships by analyzing semantic and structural connections between clauses within the same document. Our work enhances the granularity and contextual understanding of contract language, enabling more robust clause classification, summarization, and fairness analysis.

Introduction

Legal contracts are complex documents filled with dense, technical language that can be difficult to analyze and understand without expert knowledge. With the growing demand for AI-assisted legal tools, there is a critical need for structured, high-quality datasets that support tasks like clause segmentation, classification, and summarization. Existing legal NLP datasets, such as CUAD and ContractNLI, provide useful foundations, but they often lack fine-grained clause-level annotations and domain-specific context.

Our project addresses this gap by building a dataset and system capable of accurately identifying and labeling contract clauses. We implement a hybrid clause segmentation pipeline that combines regex-based patterns, semantic similarity modeling, and supervised classification to detect clause boundaries with high precision. After segmentation, we apply contract-type-specific clause categorization and explore intra-contract clause relationships to model how clauses interrelate within the same legal context.

This work contributes a reproducible framework for fine-grained contract analysis, with applications in contract review, compliance checking, and legal summarization. Our code and documentation are available at: https://github.com/joshyipp/542-LegalContract-AI.git.

Related Work

This is where you give a brief overview of any prior work by others (or yourself) that is relevant to the problem and solution you are proposing. Cite any papers using the citation and bibliography syntax illustrated below.

A comment on related work. You may find a paper or project that directly solves the problem you are proposing. Did they also release code and models? If not, is there value in reproducing their results and releasing code and the model? If they did release the code and the model, is it possible to build on their work directly and improve it?

Related work text here. This also shows how put a single citation [1] or also multiple citations [1, 2]. The bibliography is embedded in this LaTeX file.

Approach (or Methodology)

This is where you describe your solution to the problem. Elaborate on the benefits of your approach.

Proposed work text here.

Here is also an example of how to render math inline such as $f(x) = \phi_0 + \phi_1 \cdot x$. Or you can render it as a block

$$f(x) = \phi_0 + \phi_1 \cdot x$$

or numbered

$$f(x) = \phi_0 + \phi_1 \cdot x. \tag{1}$$

Datasets

Describe the datasets that you used to train and evaluate your models. Or if you are doing a dataset project, describe the dataset you created. If it is a theoretical or algorithmic project, describe any datasets that your theory or algorithm may be applicable to.

Dataset text here.

Evaluation Results

Describe your evaluation results. What metrics will did you use? What baseline from an existing solutions can you compare to?

Evalutation text here.

Conclusion

Summarize your project, results and contributions. Describe any future work that you or someone else would do if they continued the project.

Conclusion text here.

References

[1] Author1. *Title1*. Publisher, Year.

[2] Author2. *Title2*. Publisher, Year.