Grounding Legal AI through Statute Identification: Towards Comprehensive Modeling of Indian Law

Abstract

The Indian judiciary faces an unprecedented crisis of pendency, with more than 53 million cases unresolved, a judge-to-population ratio far below global standards, and systemic delays that undermine timely justice. Artificial Intelligence, and in particular advances in Legal Natural Language Processing (NLP), offer the potential to alleviate this burden by improving access for practitioners and lay people, and enabling reproducible and transparent research practices. However, Legal NLP poses challenges distinct from general-domain NLP: legal texts are unusually long and rhetorically complex, densely interlinked through citations, and there is a mismatch in granularity between different kinds of legal texts. In India, these difficulties are compounded by code-switching, stylistic inconsistencies, and the scarcity of standardized datasets.

This thesis addresses these challenges through the unifying thread of Legal Statute Identification (LSI), the task of retrieving relevant statutory provisions given the facts of a situation. The first contribution develops the ILSI dataset and introduces LeSICiN, a graph-based model that exploits citation networks and statutory hierarchies to mitigate granularity mismatch. The second contribution creates domain-adapted pre-trained language models for Indian law – InLegalBERT, InCaseLawBERT, and CustomInLawBERT. Trained on 5.4 million legal documents, these models demonstrate measurable gains across downstream tasks such as LSI, rhetorical role segmentation, and judgment prediction. The third contribution presents the first systematic benchmarking of LSI methods, comparing encoder-only and structure-aware approaches under controlled protocols, and showing that structural models are especially effective for rare and ambiguous statutes. The final contribution introduces IL-PCSR, the first dataset enabling statute and precedent retrieval in parallel from the same query, and demonstrates that hybrid ensembles with two-stage LLM re-ranking improve both tasks, with cross-task re-ranking yielding further gains.

Together, these contributions establish datasets, models, and evaluation frameworks that advance Legal NLP in India while addressing global challenges, laying a foundation for reliable, jurisdiction-aware, and integrative Legal AI systems.