

## ABSTRACT

With the advancement of the Web and an increasing number of legal documents being made available digitally, legal practitioners are now facing new challenges. From our discussions with law experts from India (faculty members from the Rajiv Gandhi School of Intellectual Property Law, India) as well as other countries like UK (Swansea University) and USA (Thomson Reuters), we understand that it is important to develop automated AI techniques for several tasks in the legal domain, e.g., identifying relevant documents, summarizing legal text, and so on. This thesis addresses certain research challenges towards developing such AI-based techniques for assisting Law practitioners.

Law practitioners have to study a large number of prior cases that match a particular case. This calls for developing legal recommendation and prior-case search systems, the key step for which is to *estimate the similarity between two legal case documents*. Existing methods of legal document similarity (which rely on document text and/or citation network structure) do *not* exploit domain knowledge inherent of a particular jurisdiction. Based on our discussions with Law practitioners, we understand that *legal statutes* (laws of a particular jurisdiction) are a valuable source of legal knowledge that can be used in several tasks including measuring case document similarity. We construct a heterogeneous network that encompasses legal knowledge in the form of statutes and case documents of a particular jurisdiction (India), and utilize this network for legal document similarity. We develop several methods that combine the textual and network information for better estimation of case document similarity.

Legal case documents are lengthy and unstructured making it difficult to understand which parts of the document mention the facts of the case, which parts mention the arguments given by the contending parties, and so on. While case documents

from some countries (Australia, USA, etc.) contain semantically structured units in the form of paragraph headings and section titles, documents of many countries (e.g., India) are devoid of such thematic segments. Hence, an important challenge is to understand the structure of these case documents. The task of *rhetorical role labelling* is to assign semantic labels (e.g., Facts, Arguments, Precedent, Final judgement) to sentences in a given case document. We perform an extensive study with multiple Law experts to analyse the subjectivity associated with the rhetorical role labeling task and introduce several neural models for the task. Through extensive experiments on legal case documents from two jurisdictions – Indian Supreme Court and the UK Supreme Court – we derive several interesting insights into the task.

Since reading and comprehending the lengthy case documents is a difficult task even for a legal expert, automatically creating summaries of these documents is an important problem. Existing methods for the task do not incorporate legal domain knowledge systematically when creating a summary. We propose an unsupervised method for legal document summarization that takes into account the guidelines that should be followed while designing case document summarization systems, based on the opinion of law experts. Thorough experiments on three datasets encompassing two jurisdictions – India and UK Supreme Court – we show the effectiveness of our proposed approach.

To summarize, we explore three practical problems in the area of Law-AI – case document similarity estimation, rhetorical role labelling, and summarization of case documents – and develop several AI techniques for assisting legal practitioners.

**Keywords:** Law-AI; Legal Information Retrieval; Semantic Segmentation; Rhetorical Role Labelling; Document summarization; Legal document similarity; Legal citation network