JAIST Repository
https://dspace.jaist.ac.jp/

Title	An Effective Framework for Legal Entailment Retrieval with Large Language Models and Optimal Transport
Author(s)	TRAN, Thanh Cong
Citation	
Issue Date	2024-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19360
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Abstract

Legal case entailment is a fundamental principle of the legal system in which the verdict of previous cases serves as a guiding precedent for later cases with similar factual circumstances. By following established rulings, this concept ensures judicial consistency and promotes predictability in legal outcomes. Due to the intricate nature of legal documents, identifying entailment between legal cases requires considerable time and effort, necessitating expertise in legal interpretation and analysis. In the field of legal AI development, a prominent initiative is the Competition on Legal Information Extraction & Entailment (COLIEE), held annually to drive advancements in information retrieval and entailment methods for legal texts. To address the legal case entailment task, early approaches from COLIEE utilized Bag-of-Words text representation and employed traditional machine learning methods for entailment prediction. While these approaches are fast and cost-efficient, they lack sufficient semantic and contextual representation for legal texts. Following the emergence of pre-trained language models such as BERT, subsequent methods have leveraged the language modeling capabilities of these architectures for legal case entailment and yielding promising results. Recent task-winners in the COLIEE competition for this task capitalize on Large Language Models (LLMs), particularly leveraging the pre-trained encoder-decoder MonoT5 architecture for entailment ranking and prediction. Despite their works are detailed in competition reports, there exists a significant gap in the literature dedicated specifically to legal case entailment. Furthermore, the performance benchmark of previous methods indicates opportunities for enhancement, underscoring the requirement for high-performance systems that are applicable in real-world scenarios.

To accelerate the process of legal case entailment through high-performance systems, this thesis proposes a two-stage framework centered on entailment information retrieval. We conceptualize this task as a document retrieval problem and develop a cost-efficient system that leverages advanced language models for legal case entailment. In the first stage, we introduce the ColBERT-UOT document retrieval model, which builds upon the ColBERT architecture by incorporating a sparse keyword alignment strategy utilizing the Unbalanced Optimal Transport framework. Our study demonstrates that

by focusing on the interaction of contextually and semantically similar keyword pairs between the query and the document, the proposed alignment method enhances the retrieval capability of ColBERT in the legal domain. In the second stage, we employ a fine-tuned MonoT5 document ranking model to refine the retrieval results and predict entailment instances. As a supplementary study, we benchmark state-of-the-art open-source LLMs in zero-shot legal case entailment to evaluate their performance and potential applications. We formulate the original task as a zero-shot list-wise entailment prediction and evaluate pre-trained LLMs of various sizes with diverse prompt designs to measure the capability of these models in legal reasoning.

We utilize the top-performing systems from COLIEE competitions between 2020 and 2024 as our baseline for comparison, alongside the zero-shot performance of established open-source LLMs. Extensive evaluation demonstrates that our proposed system significantly outperforms previous methods, consistently surpassing the baseline by a notable margin. Additionally, our system surpasses the zero-shot predictions of LLMs by a substantial margin, maintaining an average 3% performance gap in F1 score over the best-performing Llama3 70B LLM. By focusing on entailment retrieval, our system demonstrates a robust capability to identify entailment information within the top five candidates, achieving an average recall of approximately 90%. These results highlight the effectiveness of our approach and the promising potential of the proposed system for real-world applications.

In our analysis section, we examine the cost-effectiveness of ColBERT-UOT, as well as compare the alignment characteristics of the proposed sparse keyword alignment with the baseline approach. The analysis reveals that by focusing on the interaction of semantically and contextually similar keyword pairs between the query and the document, our proposed alignment strategy enhances the retrieval performance of ColBERT for legal texts. In this section, we also evaluate the performance of the two best-performing LLMs in legal case entailment under different prompt designs. Our findings indicate that although LLMs are sensitive to prompt formulation, they exhibit promising zero-shot performance in legal entailment scenarios.

In summary, this thesis investigates the task of legal case entailment and introduces a two-stage framework focused on entailment information retrieval. Based on this framework, our system, which consists of the ColBERT-UOT candidate retrieval model and the MonoT5 entailment prediction model, demonstrates superior performance compared to previous methods on the COLIEE datasets. The benchmarking study also underscores the potential

of LLMs in legal case entailment, despite their sensitivity to prompt design. For future work, we suggest focusing on the development of specialized LLMs tailored to the legal domain, leveraging extensive legal corpora to further support legal professionals and accelerate the analysis process of legal documents.