The Application of Natural Language Processing Technology in Legal Aid and Judicial Practice

Yexuan Yang *

School of Government Management, East China University of Political Science and Law, Shanghai, China

* Corresponding Author Email: 23073101121@ecupl.edu.cn

Abstract. Natural language processing (NLP) technology is an important constituent of artificial intelligence, focusing on the interaction between computers and human natural language, with the aim of enabling computers to understand, analyze, generate and process human languages. The fields of legal aid and judicial practice are also strongly related to texts, therefore, natural language processing technology has shown great potential in the legal field nowadays. In this study, the author reviews the development status, core technologies, application areas, and application value of natural language processing technology in the fields of legal aid and judicial practice. In addition, the study summarizes the difficulties and challenges faced by natural language processing technology in the field of legal aid and judicial practice. In particular, the current legal natural language processing technology still has technical problems, such as difficulty in processing complex and long legal texts, a lack of certain logical reasoning ability, insufficient understanding of non-professional questions, and a lack of public datasets, which leads to high model training costs, etc. In response to these problems, the study points out the need to make natural language processing models more adaptable to legal language, and proposes suggestions such as enhancing the popularity of models and open datasets.

Keywords: Natural language processing, legal aid, judicial practice, text information extraction.

1. Introduction

For many years, the legal field has been a key practice area for natural language processing (NLP) technology. In the traditional mode, the general public has a great demand for legal aid. However, formal legal aid is often expensive or difficult to obtain, which leads to a large number of groups being unable to access legal aid. It is estimated that 1.5 billion people worldwide are facing unresolved legal issues, while 4.5 billion people may be excluded from legal protection [1]. In addition, legal texts such as text archives of various stages of judicial case trials, official documents such as laws, regulations, and judicial interpretations, etc. [2], They are often complex and prolix, resulting in high reading and processing costs, which leads to a high demand for computer-aided assistance.

However, NLP technology has brought a turning point to this situation. The foundation of NLP technology is text, while the main components in the legal field, such as regulations, case law, contracts, and rules, are also composed of textual data. Therefore, the natural legal language processing technology has high research value and broad application prospects. Using natural language processing technology can process legal texts more quickly and accurately. Meanwhile, it is reasonable and operable to establish legal question and answer systems in the field of legal aid.

This paper will provide an overview of the development status, core technologies, application areas, values, current difficulties and challenges, prospects and conclusions of NLP technology in the fields of legal aid and judicial practice, in order to provide some guidance for new scholars to establish a knowledge framework in this field and assist other scholars studying the application status of natural legal language processing technology to understand the latest achievements in recent years, so as to quickly identify research directions worthy of in-depth study.

2. The Development Status of NLP Technology in the Fields of Legal Aid and Judicial Practice

As the rapidly developing field of NLP, Large Language Models (LLMs) were originally trained on extensive datasets, such as Wikipedia, which have demonstrated outstanding capabilities in various language tasks. Building on this success, these models are increasingly being applied in the field of natural language processing: To enhance the comprehension of legal documents' intricate language, it is necessary to adapt generic domain models to legal texts and subsequently train them further using specialized legal documents. This process enables the models to acquire a deeper understanding of the unique terminologies, phrases, and complex sentence structures frequently encountered in legal texts. The following summarizes some representative natural language processing models [3].

Firstly, Lawformer is a Chinese legal text language model based on Longformer, designed to handle long documents, which are common in legal data. Given the limitations of standard Pre-trained Language Models (PLM) in handling shorter tag sizes, Lawformer has extended sliding windows and global attention mechanisms through its innovative integration of sliding windows, which has broken through the technological bottleneck in long text processing, demonstrating remarkable applicability in legal artificial intelligence tasks. This model is particularly suitable for scenarios such as legal judicial Q&A, legal Q&A and so on. Lawformer conducts pre-training on a large corpus of Chinese legal documents, which covers two core case types: criminal and civil. It utilizes these attention techniques to integrate complex sequence dependencies between markers, thereby improving the performance of the model in legal specific tasks.

Secondly, SaulLM-7B is a new large language model designed for understanding and generating legal texts specifically, built on a Mistral architecture with 7 billion parameters. This model is trained on a vast English legal corpus, in order to face unique challenges in legal grammar and terminology. SaulLM-7B adopts a two-stage training method. Firstly, continuous pre-training is conducted on legal data of 30 billion characters, which is carefully designed. Then, an innovative approach to improving teaching was adopted, combining general guidance with specific legal guidance to enhance the model's performance in legal tasks.

Thirdly, Legal-LM is a professional language model specifically tailored for Chinese legal consulting, enhanced through a knowledge graph (KG) to address challenges in specific areas such as data accuracy and non-professional user interaction. This framework includes multiple steps: extensive pre-training on a rich legal text corpus combined with a legal KG, then using keyword extraction and direct preference optimization to improve the level of response, and data retrieval and response validation using an external legal knowledge base. This multifaceted method makes sure that Legal-LM can not only understand incomprehensible legal language, but also provide accurate and user-friendly legal advice.

Overall, NLP technology is still capable of generating text to answer legal questions, drafting regulations, and simulating legal reasoning nowadays. At the same time, it can also achieve a large degree of automation in the fields of contract review and case prediction, reducing human errors and labor costs. In addition, lawyers and legal professionals are able to reduce workload, improve work efficiency and reduce mistakes in the process of making decision to the greatest extent.

3. The Current Status of NLP Applications in Legal Analytics

3.1. Core Technologies

Pre-trained language models [4] refer to models that perform self-supervised learning of language representations on large amounts of textual data, often built on Transformers and featuring autoregressive, autoencoder, or encoder decoder architectures [5]. In the specialized PLMs for the legal field, Lawformer is based on the Longformer architecture for pre-training long legal documents, optimizing its ability to handle lengthy legal texts. SAILER is a structure aware PLM that is able to enhance the understanding of legal text structure by using logical connections in legal documents' different chapters. For example, Legal Prompt Engineering (LPE) is the process of guiding and assisting LLMs to perform Natural Legal Language Processing (NLLP) skills. It is common to combine LPE with LLM in legal judgment prediction tasks to handle lengthy legal documents.

The classification technology of Legal Statute Identification (LSI) refers to transforming legal clause retrieval into a classification task. It treats each legal clause as a unique tag, and fine-tunes Transformer models such as LSI-STARD to achieve accurate matching between queries and clauses, which is suitable for rapid localization of large-scale legal clauses [4].

Retrieval-Augmented Generation (RAG) technology refers to dynamically retrieving external knowledge bases during generation in order to improve the knowledge coverage and factual consistency of generated content. In the legal field, it is common to combine retrieval models such as BM25, Dense STAR and LLMs, and then relevant legal provisions retrieved are integrated into prompts as external knowledge to enhance the accuracy of LLMs in tasks such as legal Q&A and judgment prediction, so that model illusions are reduced [4].

The three-step method of recall, query decomposition, and filter [6] is applicable to general questions raised by non-legal professionals. This method transforms these general issues into professional legal questions and determines the most relevant legal provisions. Through establishing widely applicable legal principles (major premise) and applying them to the specific facts of the case (minor premise), one can draw conclusions. The first step is to recall, which is to determine the broad legal department category to which the problem belongs such as civil law, criminal law, or administrative law. Then gradually refine it to more specific legal fields, such as contract law or tort law in civil law, as well as more specific types, thereby narrowing the scope of legal article retrieval to specific chapters of relevant legal departments. The second step is query decomposition, which is to transform the informal language of non-legal professionals into legal facts or statements that comply with legal norms. The third step is filtering, which involves matching the legal facts obtained from the previous step with the legal provisions recalled in the first step. By calculating the intersection between the subsets of legal provisions corresponding to each legal fact and the initial collection of recall articles, the most relevant legal provisions are ultimately determined to ensure their relevance and comprehensiveness, providing a legal basis for resolving queries.

3.2. Fields

3.2.1. Legal Question Answering

The legal question answering system uses technologies such as neural networks, Transformer models and so on to automatically answer legal queries. It combines regulations, cases, and legal knowledge graphs to generate accurate answers, helping professional and non-professional users obtain legal information and assisting legal research and decision-making.

For example, Sovrano et al. proposed Disco-LQA, which is a discourse based legal question and answer system in 2024. Its main function is to answer legal questions through instance speech components, such as basic language units and abstract representation of meaning. The method can help recovers identify the most relational parts of speech, thereby improving retrieval accuracy. In addition, they introduced the Q4EU dataset, which contains more than 70 questions and 200 answers about six European regulations, resulting in improved performance in legal Q&A [7].

3.2.2. Legal Judgment Prediction

Legal judgment prediction often appears in the application scenarios of civil law. Based on the facts of the case and legal provisions, by describing the case and relevant legal provisions, it is able to use deep learning models to predict the judgment results, so that it can assist judges and lawyers in evaluating the direction of the case, and finally improve the efficiency of the judicial process.

For example, in Canadian labor law, the compensation received by an employee upon termination is linked to multiple factors such as the employee's age, service's length, nature of work, and the difficulty of re-employment. Therefore, it is very difficult for ordinary people to know their statutory compensation results. However, the RoBERTa basic model designed by Jason Lam et al. trained the model based on 4 million case data, resulting in a prediction accuracy of 69% for compensation amounts. This provides a reference for many employees and reduces unnecessary disputes [8].

3.2.3. Legal Text Classification

Legal text classification refers to the use of NLP technology to classify legal documents into predefined categories based on their content, such as civil, criminal, administrative cases, etc., in order to help legal practitioners quickly locate relevant cases, simplify the legal research process, and decision-making process. In general tasks, documents are often assigned one or more categories from a set of predefined options, including various forms such as binary classification, multi classification, and multi label classification, etc. While document classification typically falls within the realm of large-scale multi-label text classification in legal text classification tasks, the complexity of the task is immense, with the label space encompassing thousands of potential categories. Each document under scrutiny requires meticulous analysis to accurately assign multiple labels, reflecting the diverse and often nuanced legal issues present [3].

3.2.4. Legal Document Summary

Legal document summary refers to compressing legal texts into clear and informative summaries. It is worth noting that in legal texts, the unique structure and specialized content must be taken into account, such as numbering, citations, and complex and long legal language. Therefore, in the natural language processing of legal document summaries, it is necessary to enhance the hierarchical nature of the files.

Legal text abstracts can be implemented through the methods of ROUGE metrics and BERT score. Specifically, ROUGE metrics directly identify and extract the most critical information from the text, with a high degree of preservation of the original sentence. The BERT score is used to generate new sentences to retell important information from the original text. According to a study on approximately 40 million pending cases in the Indian court system in 2024, the BART, T5, PEGASUS, ROBERT, Legal PEGASUS, and Legal BERT models can be used for abstract summarization. TextRank, LexRank, LSA, Summarizer BERT, and kl sum can be used when extracting abstracts. [9] In addition, a legal document summary also includes legal argument mining (LAM). It means that it automatically extracts argument structures from legal documents, uses deep learning models to identify argument logic, assists in the sorting and analysis of arguments in legal research, and improves the efficiency of argument extraction.

3.2.5. Legal Named Entity Recognition

Legal named entity recognition refers to the use of NLP technology to extract entities unique to legal texts, such as legal provisions, courts, regulatory names, procedural terms, etc., in order to help construct legal documents and enhance legal information retrieval. It should be noted that named entity recognition in the legal field must address the complex language and structured format of legal documents, and require systems and methods tailored to the legal context. For instance, Ç etinda ğ C et al. proposed a sophisticated NER model tailored for Turkish legal texts. They developed a custom-made corpus and experimented with multiple NER architectures. These architectures were based on conditional random fields and bidirectional long-short-term memories (BiLSTMs) to effectively address the challenges of the task. Additionally, they explored various combinations of word

embeddings, including GloVe, Morph2Vec, and techniques for neural network-based character feature extraction. These advanced techniques were utilized to enhance the model's ability to recognize and classify legal entities within the Turkish texts. This facilitates subsequent information retrieval and analysis [10].

4. Challenges

Firstly, the terminology and language structure of laws and regulations are relatively complex and lengthy, and often contain words that are different from commonly used language, such as prohibited expressions [11]. These characteristics greatly increase the requirements for retrieval models. For example, Song et al. found that in a dataset containing 50000 US cases, about 35% of the cases had 3000 words or more [4], compared to news reports, which typically have only a few hundred words. Therefore, this places high demands on legal NLP systems, especially those based on pre-trained language models.

Secondly, the legal field involves a large amount of legal reasoning work, rather than just considering the similarity of texts. Therefore, in the process of natural language processing, multiple factors such as different legal terms, relationships between fields, and specific legal principles need to be considered, resulting in a large amount of data to be processed [6].

In terms of legal aid, the search content of non-professionals often lacks precise legal terminology and includes a large number of vague expressions of legal concepts, greatly increasing the difficulty of regulatory retrieval tasks.

Finally, natural language processing models are often built on large-scale, high-quality datasets. However, legal cases often involve personal data information and may even involve ethical issues [12], and public datasets are difficult to obtain, which also poses significant obstacles to the establishment of legal natural language processing models.

5. Critical Discussions

Firstly, in terms of technology, the author suggests enhancing natural language processing capabilities and improving the ability to handle complex and lengthy languages. At the same time, train the model and enhance its ability to associate texts that are far apart but have logical relationships, thus better adapting to the requirements of the legal field.

Secondly, in order to better serve the legal aid field, it is necessary to focus on developing non-professional semantic understanding systems and strengthening the technical level of converting colloquial expressions into professional legal expressions. Therefore, a wider range of non-professionals can use a common language to provide legal advice through the big language model.

Finally, more public datasets should be opened up and corpora should be continuously updated [13], providing greater data endorsement for natural language processing while protecting personal privacy, and jointly building an open-source ecosystem.

6. Conclusion

This article briefly summarizes the application of natural language processing technology in the fields of legal aid and judicial practice in recent years, and summarizes some representative NLP models in the legal field. In addition, this article sorts out and summarizes the core technologies, directions, and values of NLP used in the legal field. Finally, this article summarizes the challenges faced by natural language processing in the current legal field, and looks forward to its future development in response to these challenges.

This article provides a general description of the current status of NLP technology in the field of legal aid and judicial practice, helping scholars who want to understand this field quickly establish a cognitive framework. It is also beneficial for practitioners engaged in judicial or legal aid to understand methods and practical tools for solving the difficulties they encounter in their practice. In

addition, this paper can also help scholars who have a better understanding of this field to learn about the latest achievements in recent years, so as to quickly identify research directions worth exploring.

With the popularization of artificial intelligence and people's attention and importance to their own rights, the intelligence of the judiciary and the improvement of popularization of the law are indispensable. People should use legal technology tools to make legal services convenient and accessible, make the judicial process fairer and more efficient, and reduce human errors and biases.

References

[1] Louis A, van Dijck G, Spanakis G. Interpretable long-form legal question answering with retrieval-augmented large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38 (20): 22266-22275.

[2] An Z W, Lai Y X, Feng Y S. Natural language understanding for legal documents. Journal of Chinese Information Processing, 2022, 36 (08): 1-11.

[3] Ariai F, Mackenzie J, Demartini G. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv preprint arXiv:2410.21306, 2024.

[4] Song D, Gao S, He B, et al. On the effectiveness of pre-trained language models for legal natural language processing: An empirical study. IEEE Access, 2022, 10: 75835-75858.

[5] Zhang Z F, Ma H W, Jiang X. A review of bias evaluation and mitigation for pre-trained language models. Journal of Software Guide, 2025, 1-7.

[6] Su W, Hu Y, Xie A, et al. STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. Findings of the Association for Computational Linguistics: EMNLP 2024, 2024: 10658-10671.

[7] Sovrano F, Palmirani M, Sapienza S, Pistone V. DiscoLQA: Zero-shot discourse-based legal question answering on European legislation. Artificial Intelligence and Law, 2024.

[8] Lam J, Chen Y, Zulkernine F, et al. Legal text analytics for reasonable notice period prediction. Journal of Computational and Cognitive Engineering, 2025.

[9] Kumar H, Jayanth P. Large language models for Indian legal text summarisation. 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, 2024: 1-5.

[10] Çetindağ C, Yazıcıoğlu B, Koç A. Named-entity recognition in Turkish legal texts. Natural Language Engineering, 2023, 29 (3): 615-642.

[11] Trautmann D, Petrova A, Schilder F. Legal prompt engineering for multilingual legal judgement prediction. arXiv preprint arXiv:2212.02199, 2022.

[12] Quevedo E, Cerny T, Rodriguez A, et al. Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. IEEE Access, 2023, 12: 145286-145317.

[13] Katz D M, Hartung D, Gerlach L, et al. Natural language processing in the legal domain. arXiv preprint arXiv:2302.12039, 2023.