Legal Case Retrieval: A Survey of the State of the Art

Yi Feng1, Chuanyi Li1, Vincent Ng2*

1State Key Laboratory for Novel Software Technology, Nanjing University, China
2Human Language Technology Research Institute, University of Texas at Dallas, USA
{fy, lcy}@nju.edu.cn, vince@hlt.utdallas.edu

- Presented by – Rajith Arulanandam
- Year – 2024
- Citations – 6
- https://aclanthology.org/2024.acl-long.350.pdf

Introduction to Legal Case Retrieval

- Legal Case Retrieval (LCR) is the task of finding relevant past cases for a current legal query.
- It is foundational for legal research, litigation preparation, and judicial decision-making.
- LCR lies at the intersection of Information Retrieval (IR), Natural Language Processing (NLP), and Legal Informatics [2][3][5].

[1] Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3), 289–299.
[2] Moens, M. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law, 9*, 29–57.
[3] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge University Press.
[5] Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law, 25*(1), 65–87.

Importance of LCR

- Enhances the efficiency of legal professionals.
- Helps standardize decision-making by referencing precedent [9].
- Can provide accessible legal support to non-experts.
- Supports consistency in judicial outcomes across jurisdictions.
- Aids in legal education and public awareness.

[9] Lewis, S. (2021). Precedent and the rule of law. *Oxford Journal of Legal Studies*, *41*(4), 873–898.

LCR as an Information Retrieval (IR) Task

- Early LCR models treated legal case retrieval as a traditional IR problem.
- Techniques used: **TF-IDF**, **BM25**, and **Vector Space Models** [1][2][5].
- Legal relevance ≠ textual similarity: LCR requires legal reasoning beyond keyword overlap [6][7].
- Example: "Theft of $998" vs. "Theft of $908" — same keywords, different legal classifications (grand vs. petty theft).
- Limitation: Lexical methods miss statutory thresholds and jurisdictional nuances.

[1] Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289-299.
[2] Moens, M. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law, 9*, 29–57.
[5] Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. Artificial Intelligence and Law, 25(1), 65-87.
[6] Shao, Y. et al. (2020). Bert-PLI for legal case retrieval. IJCAI.
[7] Ma, Y. et al. (2021). LeCaRD dataset. SIGIR.

Modeling Challenges

- LCR aims to retrieve legally relevant historical cases based on a new query case.
- Effective modeling is critical but faces multiple unique challenges.
- We categorize these into accuracy, efficiency, and user trust dimensions.Core challenges: legal understanding, time modeling, scalability, interpretability [8][9].

[8] Bhattacharya, P. et al. (2020). Hier-SPCNet: legal case similarity using statutes. SIGIR.
[9] Lewis, S. (2021). Precedent and the rule of law. Oxford Journal of Legal Studies, 41(4), 873–898.

Identifying Relevant Portions

- Legal documents are lengthy and detailed.Not all content is useful for determining similarity.Need to extract case elements (e.g., facts, legal issue, outcome).[10].
- Important to isolate factually and legally meaningful content [11].

[10] Hong, Z. et al. (2020). Legal feature enhanced semantic matching. IJCNN.
[11] Rabelo, J. et al. (2022). Semantic classification of case law. Springer.

Exploiting Legal Knowledge

- Legal similarity depends on legal definitions and statutes. External knowledge (e.g., laws) often required. Systems must understand legal terms and thresholds. Legal terms require statutory and external knowledge to interpret [8].

[8] Bhattacharya, P. et al. (2020). Hier-SPCNet: legal case similarity using statutes. SIGIR.

Processing Complex Cases

- Many cases involve multiple, interrelated events.Events may follow a temporal or causal order.[11].
- Requires models to reason across long, interlinked texts.

[11] Rabelo, J. et al. (2022). Semantic classification of case law. Springer.

Time Modeling

- Laws evolve over time, Legal interpretation changes over time, Precedents may be overruled or amended [9][12].
- Temporal awareness is crucial for accurate retrieval.

[9] Lewis, S. (2021). Precedent and the rule of law. Oxford Journal of Legal Studies.
[12] Veyseh, A. P. et al. (2023). Temporal robust retrieval. arXiv:2310.15678.

Efficiency Challenges

- Legal texts are long; deep models are computationally expensive.
- Real-time or large-scale retrieval is difficult. Need efficient ranking mechanisms.[13].

[13] Zamani, H. et al. (2018). Neural ranking models with weak supervision. SIGIR.

Enabling Interpretability

- Users need to understand why a case is retrieved.
- Interpretability builds trust. Trustworthy legal AI needs transparency [10][14].
- Requires natural language explanations or visual cues.

[10] Hong, Z. et al. (2020). Legal feature enhanced semantic matching. IJCNN.
[14] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv:1901.04085.

Interactivity

- Lay users struggle to input legally precise queries.
- Systems must guide or correct user input.Interactive or conversational systems are ideal.
- User query refinement improves LCR outcomes [15].

[15] Westermann, H. et al. (2022). Chat-based interaction for legal case retrieval. arXiv:2212.11055.

Summary of Challenges

Challenge	Focus Area
Identifying Relevant Portions	Accuracy
Exploiting Legal Knowledge	Accuracy
Processing Complex Cases	Accuracy
Modeling Time	Accuracy
Ensuring Efficiency	Performance
Enabling Interpretability	Trust
Enabling Interactivity	Usability

Dataset Overview

- Legal corpora are foundational for building and evaluating LCR systems.
- A corpus is a structured collection of query cases, candidate cases, and relevance labels.
- Corpora differ by jurisdiction, language, size, and annotation method.
- Popular datasets: COLIEE, FIRE-IRLeD, CAIL2019, LeCaRD [7][16][17].

[7] Ma, Y. et al. (2021). LeCaRD dataset. SIGIR.
[16] Rabelo, J. et al. (2022). COLIEE 2021 overview. Springer.
[17] Kim, M. Y. et al. (2022). FIRE-IRLeD 2022. JSAI.

Dataset Construction

- **Common Law** (e.g., Canada, India): rely on precedents and citations.
- **Civil Law** (e.g., China): rely more on statutes and expert annotation.
- Annotation methods: automated (via citations) vs manual (by legal experts).[7][17].

[7] Ma, Y. et al. (2021). LeCaRD dataset. SIGIR.
[17] Kim, M. Y. et al. (2022). FIRE-IRLeD 2022. JSAI.

Comparison of several popularly used corpora for Legal Case Retrieval

Name	Language	Jurisdiction	# of queries	# of candidate cases/query	# of relevant cases/query
FIRE-IRLeD2017 (Mandal et al., 2017)	English	India	200	2000	5
COLIEE 2021 (Rabelo et al., 2022a)	English	Canada	900	4415	4.73
COLIEE 2022 (Kim et al., 2022a)	English	Canada	1198	2963	4.56
COLIEE 2023 (Li et al., 2023c)	English	Canada	1278	3635	4.18
CAIL2019-SCM (Xiao et al., 2019)	Chinese	China	8964	2	1
LeCaRD (Ma et al., 2021b)	Chinese	China	107	100	10.33

Table 1: Comparison of several popularly used corpora for Legal Case Retrieval.

Dataset Limitations

- **Annotation bias**: citations may not reflect legal similarity [4].
- **Lack of time-awareness**: most datasets are not temporally segmented.
- **Poor layman explanations**: limited understandable justifications. Few datasets include interpretability or time-awareness [18].
- **Improper usage**: using full case (facts + judgment) instead of just facts.

[4] Choi, S. J., & Gulati, G. M. (2008). Bias in judicial citations. Journal of Legal Studies.
[18] Shao, Y. et al. (2022). Understanding relevance judgments. ACM TOIS.

Evaluation Metrics

- Evaluation metrics help assess how well an LCR system performs.
- Two major types of models: **ranking-based** and **classification-based**.
- Metrics vary depending on the model type and task objective.
- MAP, NDCG@K for ranking [19]; F1, accuracy for classification [10].

[19] Zhu, J. et al. (2022). BERT-based ranking for legal cases. KSEM.
[10] Hong, Z. et al. (2020). Legal feature enhanced semantic matching. IJCNN.

Ranking Based Models

- **Precision@K**: Proportion of relevant cases among top K retrieved.
- **Recall@K**: Proportion of all relevant cases retrieved in top K.
- **Accuracy@K**: At least one relevant case in top K? [1][2]
- **F1@K**: Harmonic mean of Precision@K and Recall@K.
- **NDCG@K (Normalized Discounted Cumulative Gain)**: Gives more credit to relevant cases ranked higher [1].
- **MAP (Mean Average Precision)**: Average of precision values where relevant cases appear [1][19].

[1] Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289–299.
[2] Moens, M. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law, 9*, 29-57.
[19] Zhu, J. et al. (2022). BERT-based ranking for legal cases. *KSEM*.

Approaches to LCR - Overview

- LCR approaches have evolved from simple lexical models to advanced deep learning frameworks.
- Two main categories:
  - Traditional Approaches
  - Neural Approaches
- Each has strengths and limitations.

Traditional IR Approaches

- Based on lexical and statistical features like TF-IDF and BM25 [1][2].
- Use hand-crafted features or document vectors (e.g., doc2vec).
- Examples:
  - Vector Space Model [1]
  - BM25 [2]
- Fast and interpretable but limited in semantic understanding.TF-IDF, BM25, and lexical features provide efficient baseline [5][20].

[1] Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289–299.
[2] Moens, M. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law, 9*, 29–57.
[5] Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval.
[20] Schafer, B., et al. (2020). Information retrieval in legal systems. Information & Communications Technology Law.

Basic Neural Models

- Move from feature engineering to representation learning [21][22].
- Use pre-trained language models (PLMs) like BERT, RoBERTa.
- Learn semantic relationships between cases. Compute similarity between encoded vectors.
- BERT, LSTM encode documents as contextual embeddings [22].

[21] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers. *NAACL*.
[22] Vaswani, A., et al. (2017). Attention is All You Need. *NeurIPS*.

Attention-Based Models

- Use attention to highlight key legal phrases [10][22].
- Improves focus on relevant content: e.g., facts, legal statutes.
- Enables more precise comparisons.
- Token-level attention reveals important parts for decision-making [23].

[10] Hong, Z. et al. (2020). Legal feature enhanced semantic matching. *IJCNN*.
[22] Vaswani, A., et al. (2017). Attention is All You Need. NeurIPS.

Sentence-Level Models

- Break down documents into smaller units [19][6].
- Match sentences or paragraphs independently.
- Improves accuracy for long, multi-event cases.
- Compare sentence pairs to reduce document complexity [23].

[19] Zhu, J. et al. (2022). BERT-based ranking for legal cases. *KSEM*.
[6] Shao, Y. et al. (2020). BERT-PLI for legal case retrieval. *IJCAI*.
[23] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.

Coarse-to-Fine Approaches

- Combine traditional IR and deep models [7][19].
- Step 1: Coarse retrieval (e.g., BM25).
- Step 2: Fine-grained ranking (e.g., Legal-BERT).
- First filter with BM25, then rank with deep models [24].

[24] Xiong, L., et al. (2021). Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ICLR.

Knowledge-Rich Models

- Integrate external legal knowledge like statutes or legal graphs [8][25].
- Models like Hier-SPCNet use statute-citation graphs [8].
- Legal-BERT and Lawformer pre-trained on legal corpora [25].

[8] Bhattacharya, P. et al. (2020). Hier-SPCNet. SIGIR.
[11] Rabelo, J. et al. (2022). Semantic classification of case law. Springer.
[25] Chalkidis, I., et al. (2020). LEGAL-BERT: Pretrained Transformers for Legal NLP. Findings of EMNLP.

Interpretable and Interactive LCR

- Interpretability: Explain why a case is retrieved [18][15].
- Interactivity: Assist users with query refinement [15].
- Helps build trust and usability.

[8] Bhattacharya, P. et al. (2020). Hier-SPCNet. SIGIR.
[15] Westermann, H. et al. (2022). Chat-based interaction for legal case retrieval. *arXiv:2212.11055*.

SOTA Systems

- Many LCR systems have been benchmarked using datasets like COLIEE, FIRE-IRLeD, LeCaRD, and CAIL.
- Systems are evaluated on precision, recall, F1, MAP, and NDCG.
- Trade-offs exist between performance, interpretability, and efficiency.
- Hybrid IR + neural models are state-of-the-art [16][20].

[16] Rabelo, J. et al. (2022). COLIEE 2021 overview. Springer.
[20] Schafer, B., et al. (2020). Information retrieval in legal systems. *Information & Communications Technology Law*.

Summary of SOTA Systems

Dataset	System	Key Features
COLIEE 2021	Ma et al. [7]	Coarse-to-fine + BERT-PLI entailment
COLIEE 2022	Rabelo et al. [11]	Paragraph histograms + classifier
COLIEE 2023	Li et al. [7]	Lexical + semantic + LightGBM + filtering
FIRE 2017	Sampath & Durairaj [17]	Case elements + CNN + binary classifier
CAIL	Bi et al. [18]	Legal knowledge graph + node embeddings
LeCaRD	Zhu et al. [19]	Query type + BM25 + BERT ranking

Ethical Considerations

- Legal decisions impact lives—ethical design is essential.
- LCR systems must be fair, unbiased, interpretable, and respectful of due process.
- Focus areas: bias, transparency, responsibility, and misuse.
- Biases in citation, race, and annotation affect fairness [4][15].

[4] Choi, S. J., & Gulati, G. M. (2008). Bias in judicial citations. Journal of Legal Studies.
[15] Westermann, H. et al. (2022). Chat-based legal AI interaction. arXiv.

Principle	Implication for LCR Systems
Fairness	Avoid bias in data and model predictions
Transparency	Provide understandable reasoning
Responsibility	Support, not supplant, human decision-making
Privacy	Protect sensitive information
Equity	Ensure accessibility across user groups

Summary of Ethical Principles

Concluding Remarks

- Legal Case Retrieval (LCR) remains a challenging yet promising domain.
- Advances in AI, legal corpora, and evaluation metrics have driven progress.
- Ethical and practical deployment remains a top priority.
- LCR demands precision, legal sensitivity, and interpretability.
- Models must incorporate legal knowledge, time awareness, and fairness.
- Focus on time-aware, interpretable, and fair systems [9][12].

[9] Lewis, S. (2021). Precedent and the rule of law.
[12] Veyseh, A. P. et al. (2023). Temporal robust retrieval. arXiv.

References

- [1] Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289–299.

[2] Moens, M. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence and Law, 9*, 29–57.

[3] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge University Press.

[4] Choi, S. J., & Gulati, G. M. (2008). Bias in judicial citations. *Journal of Legal Studies, 37*(1), 87–129.

[5] Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law, 25*(1), 65–87.

[6] Shao, Y. et al. (2020). BERT-PLI for legal case retrieval. *IJCAI*.

[7] Ma, Y. et al. (2021). LeCaRD dataset. *SIGIR*.

[8] Bhattacharya, P. et al. (2020). Hier-SPCNet: legal case similarity using statutes. *SIGIR*.

[9] Lewis, S. (2021). Precedent and the rule of law. *Oxford Journal of Legal Studies, 41*(4), 873–898.

[10] Hong, Z. et al. (2020). Legal feature enhanced semantic matching. *IJCNN*.

[11] Rabelo, J. et al. (2022). Semantic classification of case law. *Springer*.

[12] Veyseh, A. P. et al. (2023). Temporal robust retrieval. *arXiv:2310.15678*.

[13] Zamani, H. et al. (2018). Neural ranking models with weak supervision. *SIGIR*.

[14] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv:1901.04085*.

[15] Westermann, H. et al. (2022). Chat-based interaction for legal case retrieval. *arXiv:2212.11055*.

[16] Rabelo, J. et al. (2022). COLIEE 2021 overview. *Springer*.

[17] Kim, M. Y. et al. (2022). FIRE-IRLeD 2022. *JSAI*.

[18] Shao, Y. et al. (2022). Understanding relevance judgments. *ACM TOIS*.

[19] Zhu, J. et al. (2022). BERT-based ranking for legal cases. *KSEM*.

[20] Schafer, B., et al. (2020). Information retrieval in legal systems. *Information & Communications Technology Law*.

[21] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers. *NAACL*.

[22] Vaswani, A., et al. (2017). Attention is All You Need. *NeurIPS*.

[23] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP*.

[24] Xiong, L., et al. (2021). Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *ICLR*.

[25] Chalkidis, I., et al. (2020). LEGAL-BERT: Pretrained Transformers for Legal NLP. *Findings of EMNLP*.

Thank You

Q and A