ECtHR-PCR: A Dataset for Precedent Understanding and Prior Case Retrieval in the European Court of Human Rights

Santosh T.Y.S.S, Rashid Gustav Haddad, Matthias Grabmair
Technical University of Munich, Germany

LREC-COLING 2024

Introduction

- Common law jurisdictions rely on existing case decisions as a vital source of law
  - Growing demand for PCR systems to aid practitioners

- Problems with existing datasets such as COLIEE, IRLeD
  - Query comprise both factual and reasoning sections with just citations text suppressed
    - Can result in exact text matching in cases due to verbatim quotations
  - In a realistic scenario, the reasoning section of a case is often available only after the final verdict has been delivered, while only the factual aspects are accessible prior to the verdict.
  - Artificially created negative pool of candidates to select from not simulating realistic scenario

Introduction

- Curate PCR dataset for European Court of Human Rights
  - ECHR's case law documents separate the facts from arguments
    - ensure queries used for PCR do not contain the argument/reasoning
  - We assess both lexical and dense retrieval based approaches employing different negative sampling strategies for PCR task
- What factors constitute the *ratio decidendi* (binding reasons for a decision that have an impact on subsequent cases)
  - Halsbury 1907 : Judge's reasoning and arguments are what bind
  - Goodhart 1930 : Analogy between the facts of the precedent and the current case.
- We empirically test Halsbury's and Goodhart's views in practice using our PCR dataset.

ECHR PCR dataset

- Document Collection & Filtering:
  - Scrapped HUDOC, retained only English documents
- Parsing documents
  - Parse each document into facts, law section
- Citation extraction
  - Using various regex & heuristics to obtain citations
- Mapping citations to documents

- 15,729 judgements
- Chronologically split
  - Training (9.7k, 1960–2014)
  - Validation (2.1k, 2015–2017)
  - Test (3.2k, 2018–2022)

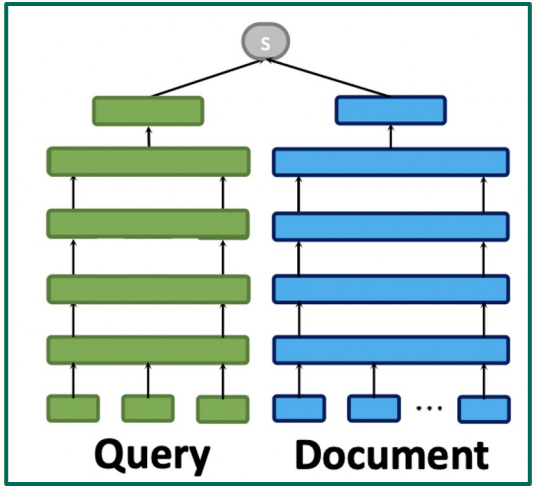
Dataset	ECtHR-PCR			IRLeD	COLIEE	
Split	Train	Valid	Test	Test	Train	Test
#Queries	9787	2186	3231	200	898	300
Avg. #Candidates per query	5283.22	11374.96	14102.01	2000	4415	1564
Avg. #relevant Doc per query	9.61	12.97	12	5	4.68	-
Avg. #words in query	1706.39	1765.11	1743.57	7883.41	4628.42	5327.08
Avg. #words in candidate	5530.43	6075.86	5887.91	7377.77	4777.98	4976.06

Models

- BM25
- Dense Models: Hierarchical BERT model to account for longer text
    - Negative sampling
        - Random
        - Random + BM25
        - ANCE
            - Select top-k negatives ranked by the dense retrieval model which is in training.
- Uni vs Bi Encoder: Same encoder both query and document vs different ones

Metrics

- Recall@k
- MAP
- Mean Rank
- Median rank



Results

- BM25 performs better than DR-uniencoder and competitive with our biencoder models
- Biencoder model outperforms the uniencoder and BM25 model
  - Differing semantics between queries and documents
  - Queries contain only the factual statements
  - Documents contain both the factual statements and the reasoning section.

	Recall@k (↑)					Mean	Median
Model	**50**	**100**	**500**	**1000**	**MAP (↑)**	**Rank (↓)**	**Rank (↓)**
BM25	22.14	27.82	47.8	60.38	9.65	1945.73	1218.07
DR-Rand-Uniencoder	19.33	26.19	47.61	58.9	7.28	1827.08	1388.38
DR-Rand-Biencoder	20.36	29.26	56.03	67.31	6.72	1676.55	1387.8

Results

- *Surprisingly*, *DR-Rand* better than DR-BM25+Rand and the DPR-ANCE model
  - Using difficulty-based hard negatives lowered performance compared to random
  - hard-negative selection strategies may end up in selecting relevant documents that simply have not been cited (false negatives)

	Recall@k (↑)					Mean	Median
Model	**50**	**100**	**500**	**1000**	**MAP (↑)**	**Rank (↓)**	**Rank (↓)**
DR-Rand	20.36	29.26	56.03	67.31	6.72	1676.55	1387.8
DR-BM25+Rand	13.65	18.8	38.51	51.72	5.35	2275.41	1944.51
DR-ANCE	15.38	22.63	45.97	57.4	4.9	2101.06	1703.8

ECtHR: Halsbury or Goodhart's view?

- Using law section of the document alone turns better than using the facts section alone
- Evidence supporting Halsbury's view in the ECtHR compared to Goodhart's view, aligning with the findings of Valvoda et al. 2021

Model	Document Elements	Recall@k (↑)				MAP (↑)	Mean Rank (↓)	Median Rank (↓)
		50	100	500	1000			
BM25	Facts	19.54	24.81	41.98	52.55	8.24	2577.65	1802.26
DR-Rand - UniEncoder	Facts	18.89	25.21	44.33	54.07	7.23	2205.01	1688.33
DR-Rand- BiEncoder	Facts	17.04	24.65	49.76	60.8	5.41	2225.31	1797.27
BM25	Law	22.72	28.67	52.25	62.47	10.26	1824.59	1053.12
DR-Rand - UniEncoder	Law	19.22	26.62	50.28	62.38	7.31	1503.02	1034
DR-Rand- BiEncoder	Law	23.72	32.23	56.08	66.85	7.87	1572.15	1217.87

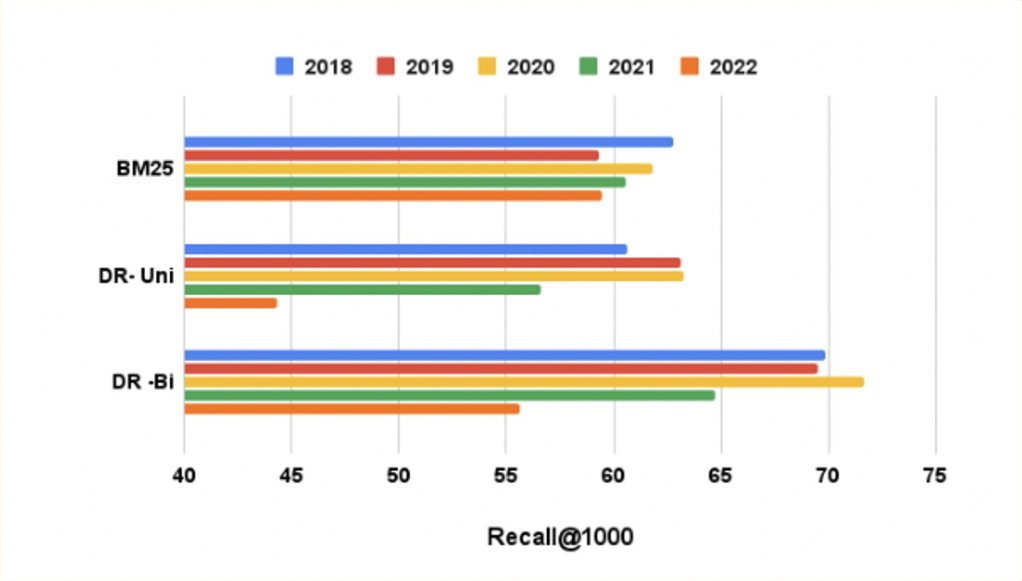
ECtHR: Halsbury or Goodhart's view?

- Using the law section alone proves more effective than using the entire document.
  - Important facts are discussed in the law section, helping model to focus on the relevant factual information presented in the law section.
  - Adding facts tends to shift the model's focus with unnecessary additional information

Model	Document Elements	Recall@k (↑)				MAP (↑)	Mean Rank (↓)	Median Rank (↓)
		50	100	500	1000			
BM25	Facts&Law	22.14	27.82	47.8	60.38	9.65	1945.73	1218.07
DR-Rand-Uniencoder	Facts&Law	19.33	26.19	47.61	58.9	7.28	1827.08	1388.38
DR-Rand-Biencoder	Facts&Law	20.36	29.26	56.03	67.31	6.72	1676.55	1387.8
BM25	Law	22.72	28.67	52.25	62.47	10.26	1824.59	1053.12
DR-Rand - UniEncoder	Law	19.22	26.62	50.28	62.38	7.31	1503.02	1034
DR-Rand- BiEncoder	Law	23.72	32.23	56.08	66.85	7.87	1572.15	1217.87

Temporal Degradation

- Dense models deteriorate over time
  - Due to temporal distributional shift
  - Difference between the training and the test data distribution
  - due to addition of new case documents to candidate pool over time.



Conclusion

- Present ECtHR-PCR, prior case retrieval dataset for jurisdiction of European Court of Human Rights.
- Assess various retrieval baselines, both lexical-based and dense retrieval models
- Difficulty-based negative sampling underperforms random negative sampling
- Dense models degrade with time, while BM25 is temporally robust
- Need to develop temporally robust retrieval models
- Empirically examined the contested Halsbury's and Goodhart's view on what constitutes a ratio and witnessed Halsbury's view that reasoning and arguments hold more weight in ECtHR

Future Work

- Solely focuses on text alone: Need to leverage citation network
- Leverage impact factor or influence score of a case
- Deal with dynamic evolution of law, by capturing temporal nature of precedents
- Deduce the reasoning process on why a document needs citation thus making more interpretable