LREC-COLING 2024

CuSINeS: Curriculum-driven Structure Induced Negative Sampling for Statutory Article Retrieval

Santosh T.Y.S.S, Kristina Kaiser, Matthias Grabmair

Technical University of Munich, Germany

Statutory Article Retrieval (SAR)

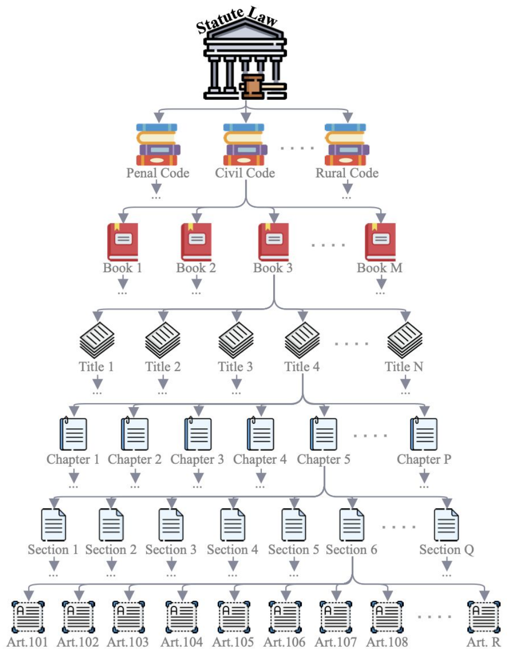
- Finding relevant statutes for a legal question
- SAR techniques have been explored with COLIEE Statute Law Corpus (Rabelo et al., 2021)
  - questions linked to relevant articles from the Japanese Civil Code
  - However, questions are obtained from legal bar exam yes or no questions
  - quite different from those posed by ordinary citizens, often being vague and underspecified
- Louis and Spanakis 2022 developed the Belgian Statutory Article Retrieval Dataset (BSARD)
  - French legal questions from Belgian citizens labeled by legal experts with references to relevant articles from Belgian legislation

Prior Works

- BM25, TF- IDF (Yoshioka et al., 2018), Indri (Strohman et al., 2005)
- Word Movers' Distance (Kusner et al., 2015).
- BERT and their ensembles (Kim et al., 2019; Rabelo et al., 2021, 2022).
- Dense retrieval methods (Louis and Spanakis, 2022)
- Synthetic query generation and legal domain-oriented pre-training (Louis et al., 2023).
- Graph neural networks to enrich article representations (Louis et al., 2023)
    - Using interdependencies among articles within the topological structure of legislation
- A key aspect that has been overlooked : How to construct high-quality negative samples for training SAR models ?
    - Prior works rely on BM25-based semantic similarity to derive hard negatives.

CuSINeS

- CuSiNeS introduce a curriculum-based scheduling of negative samples
    - Learn positive articles from easier negatives in the initial learning
    - Gradually transition to learning from difficult negatives.
- CuSiNeS utilizes topological structure of legislation to mine hard negatives
    - Distant statutes (greater shortest path) cover broader legal themes
        - Negative articles distant from the positive is easier to distinguish



- CuSiNeS dynamically assess semantic difficulty with the retrieval model being trained.
    - Beyond the static estimation derived from which is model-independent BM25
    - Expose to negatives during training based on model current competence.

Background

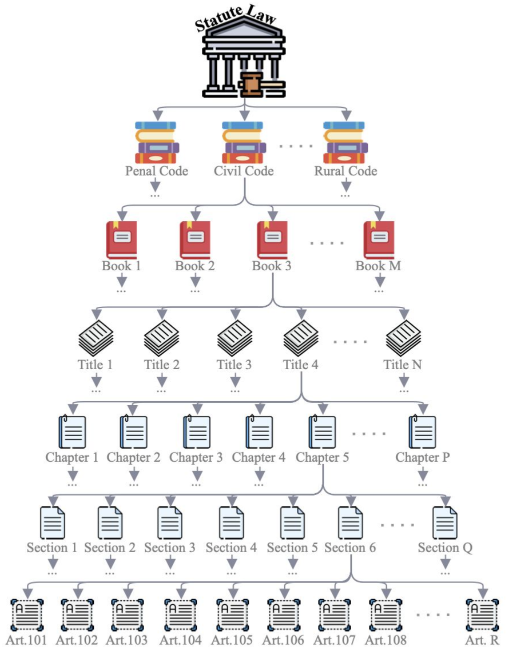
- Statutory Article Retrieval:
  - Given a question q and corpus of statues S = {s1,s2,...,sm}, the task is to retrieve a smaller set of statutes Sq (|Sq| << |S|) ranked in terms of their relevance to answer the query.
- Dense Retrieval (DR):
  - Dual-encoder architecture (Karpukhin et al., 2020) with query and statute encoder to map them into a dense vector
  - Relevance score is dot product between encodings of query and statute
  - **T**rained with contrastive loss - pull the representations of the query and relevant articles together (as positives), while pushing apart irrelevant ones (as negatives)
  - Negative sampling - some irrelevant documents are sampled for each query during training

CuSINeS - Difficulty Ranking of Negatives

- Semantic-based ranking
    - Unlike static model independent BM25
    - Dynamically compute semantic difficulty of negative articles using the model being trained
        - Based on semantic relevance score to the query
    - A higher relevance score indicates a more difficult negative article
    - Captures model's learning dynamics
    - Refresh the difficulty rankings at each epoch
        - Reduce the inference cost associated with continuous updates

CuSINeS - Difficulty Ranking of Negatives

- Structure-based ranking : Two views based on structure:
  - Hierarchical view:
    - Difficulty of each negative article is measured by its proximity to the set of positive articles within the hierarchical graph.
    - Based on the shortest path distance between each positive article and every negative article within the hierarchical graph.
  - Sequential View:
    - Treats statutes as a linearized sequence.
    - Calculates the distance by considering the relative position in the sequential enumeration of articles.



CuSINeS - Combining Different Difficulty Rankings

- Semantic difficulty captures the interplay between queries and negative articles
- Structural difficulty reflects the relationship between positive and negative articles,
- To capture the complementary information, unify these rankings through reciprocal rank fusion (RRF) (Cormack et al., 2009).

$$RRFscore(d) = \sum_{i} \frac{1}{k + r_i(d)},$$

CuSINeS - Curriculum Scheduler

- Based on cumulative difficulty ranking, categorize negatives into various difficulty-level buckets, ranging from easy to difficult.
- In initial training, more samples from the easier buckets with lesser from the difficult ones.
- As training progresses, the ratio gradually shifts, with more of difficult samples
- This adaptive scheduling akin to the Leitner system of spaced repetition that improves human learning.

Dataset & Metrics

- BSARD dataset (Louis and Spanakis, 2022)
  - 1108 french legal questions
  - Corpus of 22,600 Belgian legal articles.
  - A query can have multiple relevant legal articles.
- Metrics:
  - Recall@k (R@k) (k=100,200,500) : Proportion of relevant articles in the top-k
  - Mean Average Precision (MAP) : Average of Precision@k scores for every rank position k of each relevant document
    - Precision@k is the proportion of relevant documents in the top-k candidates
  - Mean R-Precision (MRP)
    - R- Precision is the proportion of the relevant articles in the top-k ranked ones (k is the exact number of relevant articles for that query)

Base Models

- BM25 (Robertson et al., 1995)
- Dense Retrieval (DR) model
  - Query encoder - BERT, Article encoder - Hierarchical version of BERT
- DR+GNN
  - DR model with Graph Attention Network to enrich article representations by using legislative graph topology constructed from hierarchical organization of statutes.
- Two initializations of DR
  - CamemBERT (Martin et al., 2020)
  - LegalCamemBERT (Louis et al., 2023)

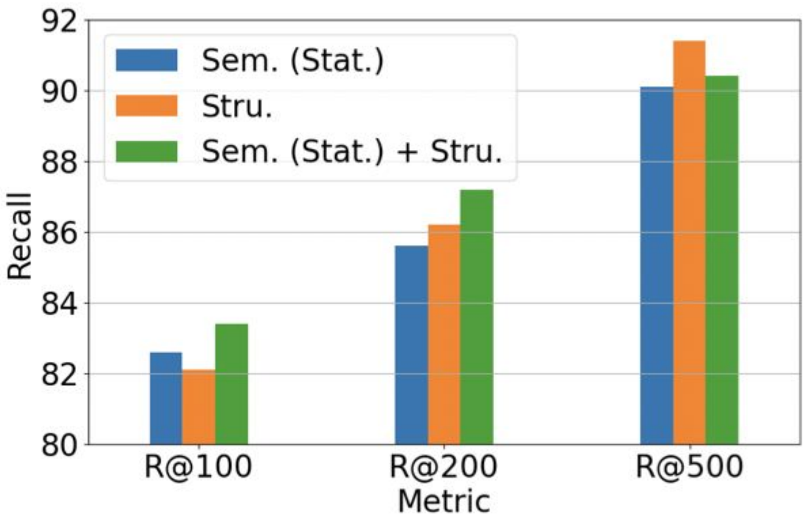
Results

Method		R@			MAP	MRP
		100	200	500		
BM25	Baseline	49.3	57.3	63	16.8	13.6
DR CB	Baseline	77.1	81.8	86.7	35.6	28.8
	CuSINeS	82.6	86.6	91.6	38	29.1
DR+GNN CB	Baseline	80.2	83.2	88.6	39.2	32.6
	CuSINeS	83.2	88.1	92.6	42.2	33.4
DR LCB	Baseline	79.8	83.9	88.9	39.5	31.3
	CuSINeS	83.7	87.5	92.3	41.2	32.1
DR+GNN LCB	Baseline	82.6	85.6	90.1	44.6	35.8
	CuSINeS	84.9	89.6	93.3	46.2	36.2

CuSINeS improvement can be attributed to

1. Legislative Structure topology based negative mining
2. Curriculum-based negative schedule
3. Dynamic criterion of semantic-based difficulty ranking

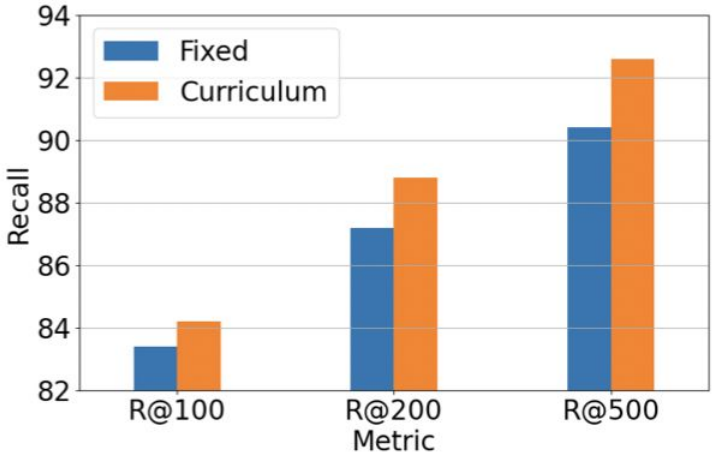
Influence of Structure and Semantic Difficulty



Structure-based negatives are more informative than semantic ones with improvements on R@k at higher k-values.

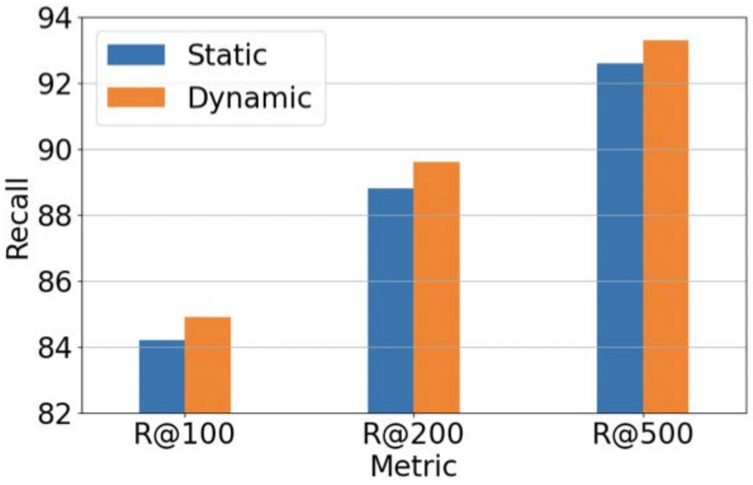
Combining both validates their complementary nature at lower k-values.

Influence of Fixed vs Curriculum Schedule



Easy-to-difficult curriculum helps to learn coarse-grained distinctions between the articles initially and then progressively move towards finer-grained nuances.

Influence of Static vs Dynamic Semantic Difficulty



Dynamic semantic ranking helps to improve performance

Computing difficulty dynamically helps to design adaptive curricula based on its competence at the current training step.

Conclusion

- Improved the SAR performance on BSARD dataset with CuSINeS, our model-agnostic negative sampling method.
- CuSINeS leverages
  - Structural information from statutes to assess difficulty for negative sampling
  - Dynamically update semantic difficulty based on the model in training.
  - Curriculum-based training schedule
- Inspiring further research to leverage legal code structure for enhanced modeling and developing better difficulty assessment methods for curricula design in various legal tasks.