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

Santosh T.Y.S.S, Kristina Kaiser, Matthias Grabmair
School of Computation, Information, and Technology;
Technical University of Munich, Germany
{santosh.tokala, kristina.kaiser, matthias.grabmair}@tum.de

Abstract

In this paper, we introduce CuSINeS, a negative sampling approach to enhance the performance of Statutory Article Retrieval (SAR). CuSINeS offers three key contributions. Firstly, it employs a curriculum-based negative sampling strategy guiding the model to focus on easier negatives initially and progressively tackle more difficult ones. Secondly, it leverages the hierarchical and sequential information derived from the structural organization of statutes to evaluate the difficulty of samples. Lastly, it introduces a dynamic semantic difficulty assessment using the being-trained model itself, surpassing conventional static methods like BM25, adapting the negatives to the model's evolving competence. Experimental results on a real-world expert-annotated SAR dataset validate the effectiveness of CuSINeS across four different baselines, demonstrating its versatility.

Keywords: Statute Retrieval, Curriculum Learning, Negative Sampling, Legislation Structure

1. Introduction

In an age marked by complex legal challenges, there's a growing imperative to bridge the gap between legal expertise and public comprehension (Ponce et al., 2019). Statutory article retrieval (SAR) involves finding relevant statutes for a legal question and is a vital initial step in legal assistance. Traditionally, SAR methods have been explored using the COLIEE Statute Law Corpus (Rabelo et al., 2021), containing questions linked to relevant articles from the Japanese Civil Code. However, these questions which are obtained from legal bar exam yes or no questions, are quite different from those posed by ordinary citizens, often being vague and underspecified. To address this, Louis and Spanakis 2022 developed the Belgian Statutory Article Retrieval Dataset (BSARD), featuring french legal questions from Belgian citizens labeled by legal experts with references to relevant articles from Belgian legislation, which we use in our study.

Traditional SAR techniques included BM25, TF-IDF (Yoshioka et al., 2018), Indri (Strohman et al., 2005) and Word Movers' Distance (Kusner et al., 2015). With the rise of pre-trained models, BERT and their ensembles have become popular (Kim et al., 2019; Rabelo et al., 2021, 2022). Recently, dense retrieval methods have gained attention (Louis and Spanakis, 2022) and were enhanced further through synthetic query generation and legal domain-oriented pre-training (Louis et al., 2023). Graph neural networks have been applied to enrich article representations by exploiting the interdependencies among articles within the topological structure of legislation, which consist of a network of interconnected statutes organized into codes, books, titles, chapters, and sections, forming a hierarchical and sequential framework.

Despite these improvements, a key aspect has been overlooked: how to construct high-quality negative samples for training SAR models. Prior methods have primarily relied on BM25-based semantic similarity to derive hard negatives. However, there have been no explicit efforts to utilize the structural organization of legislation for mining hard negatives. While Louis et al. 2023 used this structural information to derive article representations, our approach utilizes this to mine hard negatives, which is orthogonal to their work.

This structural organization reveals that the distant statutes, with greater shortest path, cover broader legal themes while the statutes with lesser shortest path deal with similar legal concepts. This makes negative articles distant from the candidate positive easier to distinguish, while the near ones are more difficult to distinguish. This insight into difficulty estimation based on structure complements the traditional semantic approach. Additionally, we enhance semantic difficulty estimation by dynamically assessing it with the retrieval model being trained. This goes beyond the static estimation derived from BM25, which is model-independent. This makes the model expose to those negatives during training based on its current competence.

Furthermore, inspired by curriculum learning, which suggests that learning often begins with simpler samples before gradually moving to more complex ones (Bengio et al., 2009), we introduce a curriculum-based scheduling of negative samples. This approach guides the model to focus on differentiating positive articles from easier negatives in

LREC-COLING 2024, pages 4266–4272
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

the initial learning stages and gradually transition to learning sophisticated reasoning from difficult negatives as the training advances. This helps the model find better local minima by mitigating the negative impact of difficult samples in the early stages of training (Hacohen and Weinshall, 2019).

Combining these three insights, we introduce CuSINeS, a Curriculum-driven Structure Induced Negative Sampling approach, which is model-agnostic and can be employed with training any SAR model. We apply CuSINeS on top of four SAR models on the BSARD dataset showcasing the versatility of our approach.

2. Preliminaries

Statutory Article Retrieval: Given a question q and corpus of statues $S = \{s_1, s_2, \ldots, s_m\}$, the task of SAR is to retrieve a smaller set of statutes S_q ($|S_q| << |S|$) ranked in terms of their relevancy to answer the query. We mainly deal with optimizing the recall of the SAR system acting as pre-fetcher, leaving the re-ranker component optimized for precision, for future.

Dense Retrieval (DR): They use a dual-encoder architecture (Karpukhin et al., 2020), where the relevance score is computed using dot product between encodings of query q and statute s_i as $f(q, s_i) = E_q(q) \cdot E_s(s_i)$ where E_q, E_s denote query and statute encoder to map each of them into a k-dimensional dense vector respectively.

Training with Negative Sampling: DR models are trained with contrastive loss whose objective is to pull the representations of the query q and relevant articles S_q together (as positives), while pushing apart irrelevant ones $S'_q = S - S_q$ (as negatives) (Lee et al., 2019). However, training with all the negatives is inscalable given larger corpus. To alleviate this issue, negative sampling has been employed where some irrelevant documents are sampled for each query during training making the final objective function as follows:

$$L(q, S_q, S'_q) = \sum_{p \in S_q} -log \frac{\exp(f(q,p)/\tau)}{\sum_{c \in \{p\} \cup S'_q} \exp(f(q,c)/\tau)} \quad (1)$$

where hyperparameter τ is a scalar temperature.

3. Our Method: CuSINeS

Unlike previous approaches that rely solely on hard negatives obtained through BM25 based on semantic relevance and employ them from the initial training stage (Louis and Spanakis, 2022; Louis et al., 2023), CuSINeS introduce a curriculum-based scheduler that exposes the model to easier negative samples before gradually introducing more challenging ones, facilitating the model to learn over time emulating the human learning process. Further, CuSINeS incorporates structural information to derive difficulty ranking in addition to the semantic one. Additionally, semantic-based ranking is updated dynamically using the model under training, based on its current competence.

3.1. Difficulty ranking of negatives

Semantic-based ranking We dynamically compute semantic difficulty of negative articles by assessing their semantic relevance to the query. A higher relevance score indicates a more difficult negative article. This dynamic ranking provides a more nuanced comprehension of the model's learning dynamics. While the model undergoes updates with each mini-batch iteration, we opt to refresh the difficulty rankings at each epoch to reduce the inference cost associated with continuous updates.

Structure-based ranking We leverage the statute structure to derive the difficulty ranks for negative articles. We consider two views based on structure: (i) Hierarchical view: We determine the difficulty of each negative article by measuring its proximity to the set of positive articles within the hierarchical graph. To create a ranked list of negatives for a given query, we follow these steps: First, we compute the shortest path distance between each positive article and every negative article within the hierarchical graph. Next, we determine the final distance for each negative article by selecting the minimum distance among all the distances to the positive articles. We rank these negatives based on distances, where a lower distance indicates a more difficult negative. (ii) Sequential View: Similar to the hierarchical view, the sequential view treats statutes as a linearized sequence. It calculates the distance score by considering the relative positional information between positive and negative articles in the sequential enumeration of articles and the difficulty rank is obtained similar to hierarchical view.

Combining multiple difficulty ranks While the semantic difficulty captures the interplay between queries and negative articles, the structural one reflects the relationship between positive and negative articles, highlighting their complementary nature. We unify these three sets of rankings through reciprocal rank fusion (RRF) (Cormack et al., 2009). For a query q and its corresponding set of negative articles S'_q, we generate rankings R using three methods (each providing a permutation on $1, \ldots, |S'_q|$). We then calculate the RRF score for each negative article and sort them to obtain the cumulative difficulty rankings.

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

Table 1: Comparison of CuSINeS with Baseline negative sampling strategy on four dense models. (L)CB denote (Legal)CamemBERT as encoder model.

3.2. Curriculum Scheduler

Based on cumulative difficulty ranking, we categorize these negatives into various difficulty-level buckets, ranging from easy to difficult. During the training process, we draw negative samples for each query from all the buckets. In initial iterations, a larger proportion of samples come from the easier buckets with smaller proportion from the difficult ones. As training progresses, the ratio gradually shifts, allocating a higher share of difficult samples in subsequent iterations. This adaptive scheduling enhances the model's ability to learn from a range of difficulty examples, akin to the Leitner system of spaced repetition that improves human learning.

4. Experiments

4.1. Dataset & Baselines

We use BSARD (Louis and Spanakis, 2022) containing 1108 french legal questions, with references to relevant articles from a corpus of 22,600 Belgian legal articles. A query can have multiple relevant legal articles.

Baselines We derive following baselines from Louis and Spanakis 2022; Louis et al. 2023 (i) BM25 (Robertson et al., 1995) as sparse retrieval baseline. (ii) Dense Retrieval (DR) model where BERT model is used as query encoder and hierarchical version of BERT is used as article encoder to account for longer length of articles. This hierarchical version splits longer text into various segments, obtains [CLS] representation for each segment and passes them through another transformer layer to obtain the final representation through max pooling. (iii) DR+GNN where dense retrieval model is augmented with graph attention network, a variant of graph neural network, to enrich article representations by fusing information from a legislative graph constructed from hierarchical organization of statutes. We experiment with two initializations in each of the dense models: one with the French CamemBERT (Martin et al., 2020) and other with the LegalCamemBERT (Louis et al., 2023) which is further pre-trained on BSARD corpus. All these baselines employ BM25 for negative mining with a fixed training schedule i.e. use these hard negatives from the initial training stage. We apply our negative sampling method, CuSINeS, to these four dense models. CuSINeS is model-agnostic and can be integrated into the training of any model.

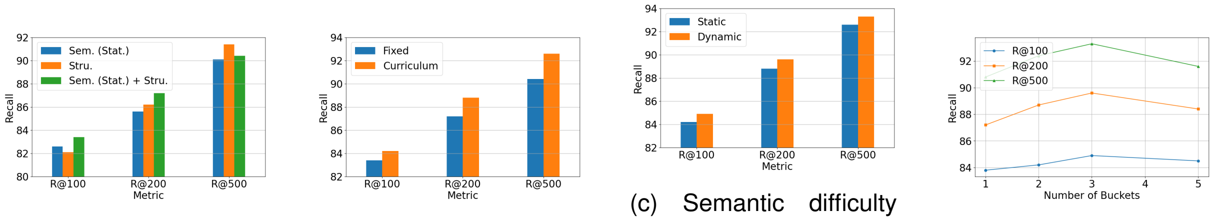
4.2. Implementation Details

We adhere to the baseline configuration outlined in previous work by Louis et al. 2023. In our hierarchical article encoder, we initialize the second-level encoder with a two-layer transformer encoder featuring a hidden dimension of 768, an intermediate dimension of 3072, 12 heads, 0.1 dropout rate, and the GeLU non-linearity function. Our training process for DR spans 15 epochs with a batch size of 24, employing AdamW optimizer (Loshchilov and Hutter, 2018) with hyperparameters β_1 = 0.9, β_2 = 0.999, ϵ = 1e-7, a weight decay of 0.01, and a learning rate warm-up for the first 5% of training steps, reaching a maximum value of 2e-5, after which linear decay is applied. For GNN, we conduct 25 epochs of training with a batch size of 512, using the AdamW optimizer with a learning rate of 2e-4. The best model is determined based on performance evaluation on the validation set. For our adaptive curriculum strategy, we sample 0.7%, 0.2%, 0.1% from easy to difficult buckets in the initial 5 epochs, 0.15%, 0.7%, 0.15% in the next 5 epochs and 0.1%, 0.2%. 0.7% in the last 5 epochs.

4.3. Performance comparison

Following previous work, we evaluate the retriever's performance using Recall@k (R@k) (k=100,200,500), Mean Average Precision (MAP) and Mean R-Precision (MRP). R@K measures the proportion of relevant articles in the top-k candidates, with results averaged across all instances. MAP and MRP provide the mean of average precision and R-Precision scores for each query where average precision is the average of Precision@k scores for every rank position of each relevant document and Precision@k represents the proportion of relevant documents in the top-k candidates. R-Precision indicates the proportion of the relevant articles in the top-k ranked ones where k is the exact number of relevant articles for that query. Higher scores in these metrics indicate better performance.

From Table 1, comparing CuSINeS to the baseline with BM25-based fixed negative sampling strategy across all four models, CuSINeS consistently outperforms the baseline demonstrating its efficacy. This can be attributed to (i) incorporation of the



(a) Structural information in negative mining

(b) Training scheduler: Fixed vs Curriculum

(c) Semantic difficulty ranking: Static vs Dynamic

(d) Number of Buckets for curriculum scheduling

Figure 1: Analysis of sub-components of CuSINeS

structural information through hierarchical and sequential proximity to derive the difficulty ranking of negatives. (ii) curriculum-based negative schedule rather than providing hard negatives from the start as in baseline. This easy-to-difficult curriculum helps to learn coarse-grained distinctions between the articles initially and then progressively move towards finer-grained nuances. (iii) dynamic criterion of semantic-based difficulty ranking using the model that is being fine-tuned rather than using static BM25 which is based on term matching and independent of the model being pre-trained.

Overall, dense models outperformed BM25, addressing the lexical gap problem. Legal pretraining and the GNN on top of DR to enrich article representations from the legislative structure continued to demonstrate the same trend even with CuSINeS sampling. Notably, CuSINeS demonstrated improvements over GNN-based approaches, highlighting the complementary nature of structure-based sampling compared to structure-based representation learning through graph networks.

4.4. Ablation Study

We study the effectiveness of each of the sub-components in CuSINeS by ablating on DR+GNN (LegalCamemBERT) model.

Incorporating structural information for negative mining We use the model with fixed training schedule with BM25-based negatives (baseline). To demonstrate the effect of structural information, we change the negative mining strategy from semantic BM25 distance to structure-based distances computed using sequential and hierarchical views. Further, we also combine these difficulty rankings using RRF. Figure 1a demonstrates structure-based negatives are more informative than semantic ones with improvements on R@k at higher k-values. Combining both of them validates their complementary nature with R@k improvement on lower k-values. This result demonstrates that legislation structure information can be leveraged to mine hard negatives to improve SAR performance.

Training Scheduler Taking the best model from previous ablation, we change the fixed scheduler to curriculum scheduler for negative sampling with three difficulty buckets where the model witnesses negatives progressively from easy to difficult. Fig. 1b demonstrates that the curriculum-based scheduling achieved better recall indicating this progressive exposure lends the model to find better local optima. This result underscores the need for better ranking criteria to design effective curricula, mirroring the way humans learn, to further enhance model performance in such complex tasks.

Semantic Difficulty Ranking We pick the best model from previous setup and ablate the static BM25-based ranking criterion for semantic view with dynamic criterion computed using the being-trained model. From Fig. 1c, we observe dynamic semantic ranking criterion helps the model to improve performance, illustrating that computing difficulty dynamically helps to design adaptive curricula needed for the model based on its competence at the current training step.

Number of Buckets We ablate on number of buckets which determines the scale of curricula progression. From Fig. 1d, we observe that the performance first improves with the increase in number of buckets (upto 3) and then drops as the buckets further increase. This indicates that a moderate number of buckets strike a balance. Too few buckets mix diverse difficulties hindering training, while more buckets narrow focus to specific difficulties, risking forgotten patterns as training advances.

5. Conclusion

We improved the SAR performance with CuSINeS, our model-agnostic negative sampling method. It leverages structural information from statutes to assess negative sample difficulty and dynamically update semantic difficulty computed from the model in training. Additionally, curriculum-based training schedule further boosts performance. Our experiments on BSARD illustrate each CuSINeS component's contributions, inspiring further research in leveraging legal code structure for enhanced modeling and developing better difficulty assessment methods for curricula design in various legal tasks.

6. Limitations

Our experimental contributions are contextual to the Belgian legislation, whose statutes are organized in a topological structure. We believe our negative sampling approach, CuSINeS, is general and could potentially be applied to most jurisdictions with structurally organized statutes. However, the performance of CuSINeS may vary across jurisdictions due to differences in legal nature, semantic difficulty, and statute organization. We leave the exploration of our method in other jurisdictions as future work. Additionally, it is worth noting that constructing such datasets can be expensive and challenging. The BSARD dataset used for evaluation introduces a linguistic bias as Belgium is a multilingual country with French, Dutch, and German speakers, but the provided legal questions and provisions are only available in French (Louis et al., 2023).

Our work focuses on the first stage of the retrieval system, optimizing for recall. To make this system practically useful, it would require a reranking component to sort the retrieved articles by importance, optimizing precision and pinpointing the exact set of statutes needed to answer each question. Furthermore, to achieve the goal of increasing accessibility, the system should not only retrieve relevant articles but also possess the capability to simplify these legal texts, making them more comprehensible to laypeople.

Furthermore, CuSINeS employs curriculum learning to determine the ordering of negative articles for every query during training. Future work can explore understanding query difficulty based on legal complexity, underspecification, or other factors to design even more effective curricula based on queries.

7. Ethics Statement

We experiment with a publicly available SAR dataset, BSARD (Louis and Spanakis, 2022). We are conscious that, by adapting pre-trained encoders, our models inherit any biases they contain and therefore should naturally be scrutinized against applicable equal treatment imperatives regarding their performance, behavior and intended use. Apart from these, we do not foresee any harm incurred by our proposed method.

8. Bibliographical References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Mi-Young Kim, Juliano Rabelo, and Randy Goebel. 2019. Statute law information retrieval and entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 283–289.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803.

Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 2023. Finding the law: Enhancing statutory article retrieval via graph neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2753–2768.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Alejandro Ponce, Sarah Chamness Long, Elizabeth Andersen, Camilo Gutierrez Patino, Matthew Harman, Jorge A Morales, Ted Piccone, Natalia Rodriguez Cajamarca, Adriana Stephan, Kirssy Gonzalez, et al. 2019. Global insights on access to justice 2019: Findings from the world justice project general population poll in 101 countries. *World Justice Project*, page 1.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. Coliee 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*, pages 196–210. Springer.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligent analysis*, volume 2, pages 2–6. Washington, DC.

Masaharu Yoshioka, Yoshinobu Kano, Naoki Kiyota, and Ken Satoh. 2018. Overview of japanese statute law retrieval and entailment task at coliee-2018. In *The Proceedings of the 12th International Workshop on Juris-Informatics (JURISIN2018)*, pages 117–128.

A. Appendix

A.1. Ablation Study

Effect of Incorporating structural information for negative mining: Tab. 2 showcases the impact of utilizing structural information on all four models. We create three variants of each model employing difficulty rankings obtained from (a) BM25-based Semantic criterion, (b) Hierarchical- and Sequential-based Structural criterion (c) Combining all of them using RRF. All these variants use fixed training schedule. Table 3 shows that using the structure of statutes yields improvements demonstrating that difficulty assessment can be derived using structural information. Combining both structural and semantic criteria further boosts recall. This highlights the potential of integrating structural insights for more effective negative mining strategies.

Effect of Training Scheduler: The results in Table 3 compare two training schedules for each of the four models. One uses a fixed schedule, where the model encounters difficult negatives right from the start of training iterations. The other employs a curriculum schedule, gradually introducing difficult negatives and starting with easier ones. All variants use rankings from three lists based on semantics, hierarchy and sequence. Table 3 shows that the curriculum-based schedule outperforms, indicating that gradually exposing the model to difficulty leads to better results.

Effect of Semantic Difficulty Ranking : In Table 4, we compare two variants of each of the four models based on semantic difficulty ranking. In one variant, the ranking is obtained using BM25 which remains static, and in the other, it's dynamically derived using the model undergoing training. All variants follow a curriculum schedule and use difficulty rankings from three lists. Table 4 clearly shows that incorporating dynamic ranking criteria helps improve model performance. This demonstrates that dynamically computing difficulty rankings allows the model to derive its learning curriculum-based on its current training progress, leading to enhanced results.

Method	Encoder Model	Training Schedule	R@100	R@200	R@500
DR	CamemBERT	Semantic (Static)	77.1	81.8	86.7
		Structural	78.8	83.1	89.3
		Semantic (Static) + Structural	77.5	84.3	89.7
DR + GNN		Semantic (Static)	80.2	83.2	88.6
		Structural	80.9	84.8	89.1
		Semantic (Static) + Structural	80.8	85.2	90.4
DR	Legal CamemBERT	Semantic (Static)	79.8	83.9	88.9
		Structural	79.9	84.9	90.6
		Semantic (Static) + Structural	78.7	84.4	89.6
DR + GNN		Semantic (Static)	82.6	85.6	90.1
		Structural	82.1	86.2	91.4
		Semantic (Static) + Structural	83.4	87.2	90.4

Table 2: Effect of Incorporating Structural Information on all four models.

Method	Encoder Model	Training Schedule	R@100	R@200	R@500
DR	CamemBERT	Fixed	77.5	84.3	89.7
		Curriculum	80.8	85.9	90.6
DR + GNN		Fixed	80.8	85.2	90.4
		Curriculum	82.5	87.7	91.6
DR	LegalCamemBERT	Fixed	78.7	84.4	89.6
		Curriculum	81.5	85.9	91.5
DR+GNN		Fixed	83.4	87.2	90.4
		Curriculum	84.2	88.8	92.6

Table 3: Effect of Training Schedule on all four models.

Method	Encoder Model	Semantic Negatives	R@100	R@200	R@500
DR	CamemBERT	Static	80.8	85.9	90.6
		Dynamic	82.6	86.6	91.6
DR + GNN		Static	82.5	87.7	91.6
		Dynamic	83.2	88.1	92.6
DR	Legal CamemBERT	Static	81.5	85.9	91.5
		Dynamic	83.7	87.5	92.3
DR+GNN		Static	84.2	88.8	92.6
		Dynamic	84.9	89.6	93.3

Table 4: Effect of Semantic Difficulty Ranking on all four models.