A Statutory Article Retrieval Dataset in French

Antoine Louis and **Gerasimos Spanakis**
Law & Tech Lab, Maastricht University
{a.louis, jerry.spanakis}@maastrichtuniversity.nl

Abstract

Statutory article retrieval is the task of automatically retrieving law articles relevant to a legal question. While recent advances in natural language processing have sparked considerable interest in many legal tasks, statutory article retrieval remains primarily untouched due to the scarcity of large-scale and high-quality annotated datasets. To address this bottleneck, we introduce the Belgian Statutory Article Retrieval Dataset (BSARD), which consists of 1,100+ French native legal questions labeled by experienced jurists with relevant articles from a corpus of 22,600+ Belgian law articles. Using BSARD, we benchmark several state-of-the-art retrieval approaches, including lexical and dense architectures, both in zero-shot and supervised setups. We find that fine-tuned dense retrieval models significantly outperform other systems. Our best performing baseline achieves 74.8% R@100, which is promising for the feasibility of the task and indicates there is still room for improvement. By the specificity of the domain and addressed task, BSARD presents a unique challenge problem for future research on legal information retrieval. Our dataset and source code are publicly available.

1 Introduction

Legal issues are an integral part of many people's lives (Ponce et al., 2019). However, the majority of citizens have little to no knowledge about their rights and fundamental legal processes (Balmer et al., 2010). As the Internet has become the primary source of information in response to life problems (Estabrook et al., 2007), people increasingly turn to search engines when faced with a legal issue (Denvir, 2016). Nevertheless, the quality of the search engine's legal help results is currently unsatisfactory, as top results mainly refer people to commercial websites that provide basic information as a way to advertise for-profit services (Hagan and Li, 2020). On average, only one in five persons obtain help from the Internet to clarify or solve their legal issue (Ponce et al., 2019). As a result, many vulnerable citizens who cannot afford a legal expert's costly assistance are left unprotected or even exploited. This barrier to accessing legal information creates a clear imbalance within the legal system, preventing the right to equal access to justice for all.

People do not need legal services in and of themselves; they need the ends that legal services can provide. Recent advances in natural language processing (NLP), combined with the increasing amount of digitized textual data in the legal domain, offer new possibilities to bridge the gap between people and the law. For example, legal judgment prediction (Aletras et al., 2016; Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018; Chen et al., 2019) may assist citizens in finding insightful patterns between their case and its outcome. Additionally, legal text summarization (Hachey and Grover, 2006; Bhattacharya et al., 2019) and automated contract review (Harkous et al., 2018; Lippi et al., 2019) may help people clarify long, complex, and ambiguous legal documents.

In this work, we focus on statutory article retrieval, which, given a legal question such as "*Is it legal to contract a lifetime lease?*", aims to return one or several relevant law articles from a body of legal statutes (Kim et al., 2019; Nguyen et al., 2020), as illustrated in Figure 1. A qualified statutory article retrieval system could provide a professional assisting service for unskilled humans and help empower the weaker parties when used for the public interest.

Finding relevant statutes to a legal question is a challenging task. Unlike traditional ad-hoc information retrieval (Craswell et al., 2020), statutory article retrieval deals with two types of language: common *natural* language for the questions and complex *legal* language for the statutes. This difference in language distribution greatly complicates

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6789 - 6803

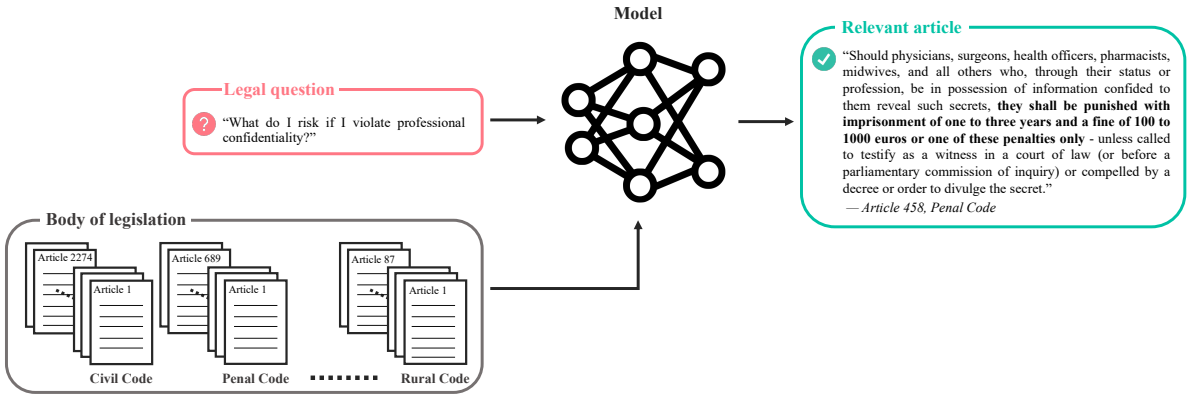


Figure 1: Illustration of the statutory article retrieval task performed on the Belgian Statutory Article Retrieval Dataset (BSARD), which consists of 1,100+ questions carefully labeled by legal experts with references to relevant articles from the Belgian legislation. With BSARD, models can learn to retrieve law articles relevant to a legal question. All examples we show in the paper are translated from French for illustration.

the retrieval task as it indirectly requires an inherent interpretation system that can translate a natural question from a non-expert to a legal question to be matched against statutes. For skilled legal experts, these interpretations come from their knowledge of a question's domain and their understanding of the legal concepts and processes involved. Nevertheless, an interpretation is rarely unique. Instead, it is the interpreter's subjective belief that gives meaning to the question and, accordingly, an idea of the domains in which the answer can be found. As a result, the same question can yield different paths to the desired outcome depending on its interpretation, making statutory article retrieval a difficult and time-consuming task.

Besides, statutory law is not a stack of independent articles to be treated as complete sources of information on their own – unlike news or recipes. Instead, it is a structured and hierarchical collection of legal provisions that have whole meaning only when considered in their overall context, i.e., together with the supplementary information from their neighboring articles, the fields and sub-fields they belong to, and their place in the hierarchy of the law. For instance, the answer to the question "*Can I terminate an employment contract?*" will most often be found in labor law. However, this is not necessarily true if an employer is contracting a self-employed worker to carry out a specific task, in which case the answer probably lies at the higher level of contract law. This example illustrates the importance of considering the question's context and understanding the hierarchical structure of the law when looking for relevant statutory articles.

In order to study whether retrieval models can approximate the efficiency and reliability of legal experts, we need a suitable labeled dataset. However, such datasets are difficult to obtain considering that, although statutory provisions are generally publicly accessible (yet often not in a machine-readable format), the questions posed by citizens are not.

This work presents a novel French native expert-annotated statutory article retrieval dataset as its main contribution. Our Belgian Statutory Article Retrieval Dataset (BSARD) consists of more than 1,100 legal questions posed by Belgian citizens and labeled by legal experts with references to relevant articles from a corpus of around 22,600 Belgian law articles. As a second contribution, we establish strong baselines on BSARD by comparing diverse state-of-the-art retrieval approaches from lexical and dense architectures. Our results show that fine-tuned dense retrieval models significantly outperform other approaches yet suggest ample opportunity for improvement. We publicly release our dataset and source code at `https://github.com/maastrichtlawtech/bsard`.

2 Related Work

Due to the increasing digitization of textual legal data, the NLP community has recently introduced more and more datasets to help researchers build reliable models on several legal tasks. For instance, Fawei et al. (2016) introduced a legal question answering (LQA) dataset with 400 multi-choices questions based on the US national bar exam. Similarly, Zhong et al. (2020) released an LQA dataset based on the Chinese bar exam consisting of 26,365

multiple-choice questions, together with a database of evidence that includes 3,382 Chinese legal provisions and the content of the national examination counseling book.

Furthermore, Duan et al. (2019) proposed a legal reading comprehension dataset with 52,000 question-answer pairs crafted on the fact descriptions of 10,000 cases from the Supreme People's Court of China. On a different note, Xiao et al. (2018) presented a dataset for legal judgment prediction (LJP) with around 2.68 million Chinese criminal cases annotated with 183 law articles and 202 charges. Likewise, Chalkidis et al. (2019a) introduced an LJP dataset consisting of 11,478 English cases from the European Court of Human Rights labeled with the associated final decision.

Meanwhile, Xiao et al. (2019) introduced a dataset for similar case matching with 8,964 triplets of cases published by the Supreme People's Court of China, and Chalkidis et al. (2019b) released a text classification dataset containing 57,000 English EU legislative documents tagged with 4,271 labels from the European Vocabulary. Additionally, Manor and Li (2019) introduced a legal text summarization dataset consisting of 446 sets of contract sections and corresponding reference summaries, and Holzenberger et al. (2020) presented a statutory reasoning dataset based on US tax law.

Recently, Hendrycks et al. (2021) proposed a dataset for legal contract review that includes 510 contracts annotated with 41 different clauses for a total of 13,101 annotations. In the same vein, Borchmann et al. (2020) introduced a semantic retrieval dataset for contract discovery with more than 2,500 annotations in around 600 documents. Lastly, the COLIEE Case Law Corpus (Rabelo et al., 2020) is a case law retrieval and entailment dataset that includes 650 base cases from the Federal Court of Canada, each with 200 candidate cases to be identified as relevant to the base case.

Regarding statutory article retrieval, the only other publicly available dataset is the COLIEE Statute Law Corpus (Rabelo et al., 2020). It comprises 696 questions from the Japanese legal bar exam labeled with references to relevant articles from the Japanese Civil Code, where both the questions and articles have been translated from Japanese to English. However, this dataset focuses on legal bar exam question answering, which is quite different from legal questions posed by ordinary citizens. While the latter tend to be vague and straightforward, bar exam questions are meant for aspiring lawyers and are thus specific and advanced. Besides, the dataset only contains closed questions (i.e., questions with "yes" or "no" answers) and considers almost 30 times fewer law articles than BSARD does. Also, unlike BSARD, the data are not native sentences but instead translated from a foreign language with a completely different legal system.[1] As a result, the translated dataset may not accurately reflect the logic of the original legal system and language. These limitations suggest the need for a novel large-scale citizen-centric native dataset for statutory article retrieval, which is the core contribution of the present work.

3 The Belgian Statutory Article Retrieval Dataset

3.1 Dataset Collection

We create our dataset in four stages: (i) compiling a large corpus of Belgian law articles, (ii) gathering legal questions with references to relevant law articles, (iii) refining these questions, and (iv) matching the references to the corresponding articles from our corpus.

Law articles collection. In civil law jurisdictions, a legal code is a type of legislation that purports to exhaustively cover a whole area of law, such as criminal law or tax law, by gathering and restating all the written laws in that area into a unique book. Hence, these books constitute valuable resources to collect many law articles on various subjects. We consider 32 publicly available Belgian codes, as presented in Table 3 of Appendix A. Together with the legal articles, we extract the corresponding headings of the sections in which these articles appear (i.e., book, part, act, chapter, section, and subsection names). These headings provide an overview of each article's subject. As preprocessing, we use regular expressions to clean up the articles of specific wording indicating a change in part of the article by a past law (e.g., nested brackets, superscripts, or footnotes). Additionally, we identify and remove the articles repealed by past laws but still present in the codes. Eventually, we end up with a corpus $\mathcal{C} = \{a_1, \cdots, a_N\}$

[1] Japan is a civil law country that relies predominantly on the rules written down in statutes, whereas most English-speaking countries (e.g., US, UK, Canada, and Australia) have a common law system that relies predominantly on past judicial decisions, known as precedents.

of $N = 22,633$ articles that we use as our basic retrieval units.

Questions collection. We partner with Droits Quotidiens (DQ),[2] a Belgian organization whose mission is to clarify the law for laypeople. Each year, DQ receives and collects around 4,000 emails from Belgian citizens asking for advice on a personal legal issue. Thanks to these emails, its team of six experienced jurists keeps abreast of Belgium's most common legal issues and addresses them as comprehensively as possible on its website. Each jurist is an expert in a specific field (e.g., "family", "housing", or "work") and is responsible for answering all questions related to that field. Given their qualifications and years of experience in providing legal advice in their respective fields, the experts can be considered competent enough to always (eventually) retrieve the correct articles to a given question.

In practice, their legal clarification process consists of four steps. First, they identify the most frequently asked questions on a common legal issue. Then, they define a new anonymized "model" question on that issue expressed in natural language terms, i.e., as close as possible as if a layperson had asked it. Next, they search the Belgian law for articles that help answer the model question and reference them. Finally, they answer the question using the retrieved relevant articles in a way a layperson can understand. These model questions, legal references, and answers are further categorized before being posted on DQ's website (e.g., the question "*What is the seizure of goods?*" is tagged under the "*Money* → *Debt recovery*" category). With their consent, we collect more than 3,200 model questions together with their references to relevant law articles and categorization tags.

Assuming it takes a jurist between 5 to 20 minutes to find the relevant articles to a given question and categorize the latter. An estimate of the pecuniary value of those labeled questions is over €105,000 – 3,200 questions, each requiring 10 minutes to label, assuming a rate of €200 per hour.

Questions refinement. We find that around one-third of the collected questions are duplicates. However, these duplicated questions come with different categorization tags, some of which providing additional context that can be used to refine the questions. For example, the question "*Should I install fire detectors?*" appears four times in total, under the following tags: "*Housing* → *Rent* → *I am a {tenant, landlord}* → *In {Wallonia, Brussels}*". We distinguish between the tags with one or a few words indicating a question *subject* (e.g., "*housing*" and "*rent*") and those that provide *context* about a personal situation or location as short descriptive sentences (e.g., "*I am tenant in Brussels.*"). If any, we append the contextual sentence tags in front of the questions, which solves most of the duplicates problem and improves the overall quality of the questions by making them more specific.

Questions filtering. The questions collected are annotated with plain text references to relevant law articles (e.g., "*Article 8 of the Civil Code*"). We use regular expressions to parse these references and match them to the corresponding articles from our corpus. First, we filter out questions whose references are not articles (e.g., an entire decree or order). Then, we remove questions with references to legal acts other than codes of law (e.g., decrees, directives, or ordinances). Next, we ignore questions with references to codes other than those we initially considered. We eventually end up with 1,108 questions, each carefully labeled with the ids of the corresponding relevant law articles from our corpus. Finally, we split the dataset into training/test sets with 886 and 222 questions, respectively.

3.2 Dataset Analysis

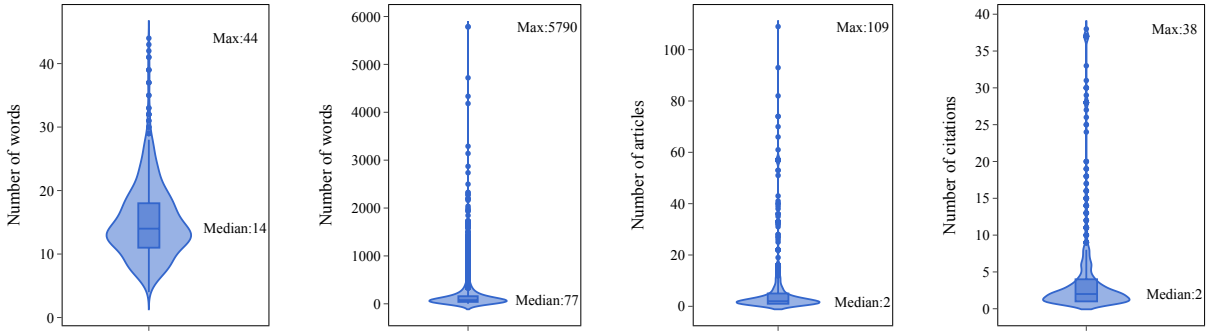
To provide more insight, we describe quantitative and qualitative observations about BSARD. Specifically, we explore (i) the diversity in questions and articles, (ii) the relationship between questions and their relevant articles, and (iii) the type of reasoning required to retrieve relevant articles.

Diversity. The 22,633 law articles that constitute our corpus have been collected from 32 Belgian codes covering a large number of legal topics, as presented in Table 3 of Appendix A. The articles have a median length of 77 words, but 142 articles exceed 1,000 words (the lengthiest one being up to 5,790 words), as illustrated in Figure 2b. These long articles are mostly *general provisions*, i.e., articles that appear at the beginning of a code and define many terms and concepts later mentioned in the code. The questions are between 5 and 44 words long, with a median of 14 words, as shown in Figure 2a. They cover a wide range of topics, with around 85% of them being either about family,

[2] https://droitsquotidiens.be/

General topic	Percentage	Subtopics	Example
Family	30.6%	Marriage, parentage, divorce, etc.	*When is there a guardianship?*
Housing	27.4%	Rental, flatshare, insalubrity, etc.	*Who should repair the common wall?*
Money	16.0%	Debts, insurance, taxes, etc.	*What is the seizure of goods?*
Justice	13.6%	Proceedings, crimes, legal aid, etc.	*How does the appeal process work?*
Foreigners	5.7%	Citizenship, illegal stay, etc.	*Can I come to Belgium to get married?*
Social security	3.5%	Pensions, pregnancy, health, etc.	*Am I dismissed during my pregnancy?*
Work	3.2%	Breach of contract, injuries, etc.	*Can I miss work to visit the doctor?*

Table 1: Distribution of question topics in BSARD.



(a) Question length.

(b) Article length.

(c) Number of relevant articles per question.

(d) Number of citations per relevant article.

Figure 2: Statistics of BSARD.

housing, money, or justice, while the remaining 15% concern either social security, foreigners, or work, as described in Table 1.

Question-article relationship. Questions might have one or several relevant legal articles. Overall, 75% of the questions have less than five relevant articles, 18% have between 5 and 20, and the remaining 7% have more than 20 with a maximum of 109, as seen in Figure 2c. The latter often have complex and indirect answers that demand extensive reasoning over a whole code section, which explains these large numbers of relevant articles. Furthermore, an article deemed relevant to one question might also be for others. Therefore, we calculate for each unique article deemed relevant to at least one question the total number of times it is cited as a legal reference across all questions. As a result, we find that the median number of citations for those articles is 2, and less than 25% of them are cited more than five times, as illustrated in Figure 2d. Hence, out of the 22,633 articles, only 1,612 are referred to as relevant to at least one question in the dataset, and around 80% of these 1,612 articles come from either the Civil Code, Judicial Code, Criminal Investigation Code, or Penal Code. Meanwhile, 18 out of the 32 codes have less than five articles mentioned as relevant to at least one question, which can be explained by the fact that those codes focus less on individuals and their concerns.

4 Models

Formally speaking, a statutory article retrieval system $R : (q, \mathcal{C}) \to \mathcal{F}$ is a function that takes as input a question q along with a corpus of law articles $\mathcal{C}$, and returns a much smaller filter set $\mathcal{F} \subset \mathcal{C}$ of the supposedly relevant articles, ranked by decreasing order of relevance. For a fixed $k = |\mathcal{F}| \ll |\mathcal{C}|$, the retriever can be evaluated in isolation with multiple rank-based metrics (see Section 5.1). The following section describes the retrieval models we use as a benchmark for the task.

4.1 Lexical Models

Traditionally, lexical approaches have been the de facto standard for textual information retrieval due to their robustness and efficiency. Given a query q and an article a, a lexical model assigns to the pair (q, a) a score $s_L : (q, a) \to \mathbb{R}_+$ by computing the sum, over the query terms, of the weights of each query term $t \in q$ in the article, i.e.,

$$s_L(q, a) = \sum_{t \in q} w(t, a). \quad (1)$$

First, we use the TF-IDF weighting scheme, in which

$$w(t,a) = \mathrm{tf}(t,a) \cdot \log \frac{|\mathcal{C}|}{\mathrm{df}(t)}, \tag{2}$$

where the term frequency tf is the number of occurrences of term t in article a, and the document frequency df is the number of articles within the corpus that contain term t. Then, we experiment with the BM25 weighting formula (Robertson et al., 1994), defined as

$$w(t,a) = \frac{\mathrm{tf}(t,a)\cdot(k_1+1)}{\mathrm{tf}(t,a)+k_1\cdot\left(1-b+b\cdot\frac{|a|}{avgal}\right)} \cdot \log \frac{|\mathcal{C}| - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5}, \tag{3}$$

where $k_1 \in \mathbb{R}_+$ and $b \in [0,1]$ are constant parameters to be fixed, $|a|$ is the article length, and $avgal$ is the average article length in the collection.

During inference, we compute a score for each article in corpus $\mathcal{C}$ and return the k articles with the highest scores as the top-k most relevant results to the input query.

4.2 Dense Models

Lexical approaches suffer from the lexical gap problem (Berger et al., 2000) and can only retrieve articles containing keywords present in the query. To overcome this limitation, recent work (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021) relies on neural-based architectures to capture semantic relationships between the query and documents. The most commonly used approach is based on a bi-encoder model (Gillick et al., 2018) that maps queries and documents into dense vector representations. Formally, a dense retriever calculates a relevance score $s_D : (q,a) \rightarrow \mathbb{R}_+$ between question q and article a by the similarity of their respective embeddings $\boldsymbol{h}_q, \boldsymbol{h}_a \in \mathbb{R}^d$, i.e.,

$$s_D(q,a) = \mathrm{sim}\left(\boldsymbol{h}_q, \boldsymbol{h}_a\right), \tag{4}$$

where $\mathrm{sim} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a similarity function such as dot product or cosine similarity. Typically, these embeddings result from a pooling operation on the output representations of a word embedding model:

$$\begin{aligned} \boldsymbol{h}_q &= \mathrm{pool}\left(f(q;\boldsymbol{\theta}_1)\right), \text{and} \\ \boldsymbol{h}_a &= \mathrm{pool}\left(f(a;\boldsymbol{\theta}_2)\right), \end{aligned} \tag{5}$$

where model $f(\cdot;\boldsymbol{\theta}_i) : \mathcal{W}^n \rightarrow \mathbb{R}^{n\times d}$ with parameters $\boldsymbol{\theta}_i$ maps an input text sequence of n terms from vocabulary $\mathcal{W}$ to d-dimensional real-valued word vectors. The pooling operation $\mathrm{pool} : \mathbb{R}^{n\times d} \rightarrow \mathbb{R}^d$ uses the output word embeddings to distill a global representation for the text passage – using either mean, max, or `[CLS]` pooling.

Note that the bi-encoder architecture comes with two flavors: (i) *siamese* (Reimers and Gurevych, 2019; Xiong et al., 2021), which uses a unique word embedding model (i.e., $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$) that maps the query and article together in a shared dense vector space, and (ii) *two-tower* (Yang et al., 2020; Karpukhin et al., 2020), which use two independent word embedding models that encode the query and article separately into different embedding spaces.

During inference, the articles are pre-encoded offline, and their representations are stored in an index structure. Then, given an input query, an exact search is performed by computing the similarities between the query representation and all pre-encoded article representations. The resulting scores are used to rank the articles such that the k articles that have the highest similarities with the query are returned as the top-k results.

4.2.1 Zero-Shot Evaluation

First, we study the effectiveness of siamese bi-encoders in a zero-shot evaluation setup, i.e., pre-trained word embedding models are applied out-of-the-box without any additional fine-tuning. We experiment with two types of widely-used word embedding models: (i) models that learned context-*independent* word representations, namely word2vec (Mikolov et al., 2013a,b) and fastText (Bojanowski et al., 2017), and (ii) models that learned context-*dependent* word embeddings, namely RoBERTa (Liu et al., 2019).

RoBERTa can process texts up to a maximum input length of 512 tokens. Although alternative models exist to alleviate this limitation (Beltagy et al., 2020; Ainslie et al., 2020), they have all been trained on English text, and there are no French equivalents available yet. Therefore, we use a simple workaround that splits the text into overlapping chunks and passes each chunk in turn to the embedding model. To form the chunks, we consider contiguous text sequences of 200 tokens with an overlap of 20 tokens between consecutive chunks.

For all zero-shot models, we use mean pooling on all word embeddings of the passage to extract a global representation for the latter and cosine similarity to score passage representations.

4.2.2 Training

Thereafter, we train our own siamese and two-tower RoBERTa-based bi-encoder models on BSARD. Let $\mathcal{D} = \{\langle q_i, a_i^+ \rangle\}_{i=1}^N$ be the training data where each of the N instances consists of a query q_i associated with a relevant (positive) article a_i^+. Using in-batch negatives (Chen et al., 2017; Henderson et al., 2017), we can create a training set $\mathcal{T} = \{\langle q_i, a_i^+, \mathcal{A}_i^- \rangle\}_{i=1}^N$ where $\mathcal{A}_i^-$ is a set of negative articles for question q_i constructed by considering the articles paired with the other questions from the same mini-batch. For each training instance, we contrastively optimize the negative log-likelihood of each positive article against their negative articles, i.e.,

$$L\left(q_i, a_i^+, \mathcal{A}_i^-\right) = -\log \frac{\exp\left(s_D(q_i, a_i^+)/\tau\right)}{\sum_{a \in \mathcal{A}_i^- \cup \{a_i^+\}} \exp\left(s_D(q_i, a)/\tau\right)}, \quad (6)$$

where $\tau > 0$ is a temperature parameter to be set. This contrastive loss allows learning embedding functions such that relevant question-article pairs will have a higher score than irrelevant ones.

To deal with articles longer than 512 tokens, we use the same workaround as in the zero-shot evaluation and split the long sequences into overlapping chunks of 200 tokens with a window size of 20. However, this time, we limit the size of the articles to the first 1,000 words due to limited GPU memory. Although not ideal, doing so remains reasonable given that only 0.6% of the articles in our corpus have more than 1,000 words, as mentioned in Section 3.2. Each chunk is prefixed by the `[CLS]` token, and we extract a global representation for the whole article by averaging the output `[CLS]` token embeddings of the different chunks. Here, we use the dot product to compute similarities as it gives slightly better results than cosine.

5 Experiments

We now describe the setup we use for experiments and evaluate the performance of our models.

5.1 Experimental Setup

Metrics. We use three standard information retrieval metrics (Manning et al., 2008) to evaluate performance, namely the (macro-averaged) recall@k (R@k), mean average precision@k (MAP@k), and mean reciprocal rank@k (MRR@k). Appendix B gives a detailed description of these metrics in the context of statutory article retrieval. We deliberately omit to report the precision@k given that questions have a variable number of relevant articles (see Figure 2c), which makes it senseless to report it at a fixed k – questions with r relevant articles will always have P@$k < 1$ if $k > r$. For the same reason, k should be large enough for the recall@k. Hence, we use $k \in \{100, 200, 500\}$ for our evaluation.

French word embedding models. Our focus is on a non-English dataset, so we experiment with French variants of the models mentioned above. Specifically, we use a 500-dimensional skip-gram word2vec model pre-trained on a crawled French corpus (Fauconnier, 2015), a 300-dimensional CBOW fastText model pre-trained on French Web data (Grave et al., 2018), and a French RoBERTa model, namely CamemBERT (Martin et al., 2020), pre-trained on 147GB of French web pages filtered from Common Crawl.[3]

Hyper-parameters & schedule. For BM25, we optimize the parameters on BSARD training set and find $k_1 = 1.0$ and $b = 0.6$ to perform best. Regarding the bi-encoder models, we optimize the contrastive loss using a batch size of 22 question-article pairs and a temperature of 0.05 for 100 epochs, which is approximately 20,500 steps. We use AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.01, learning rate warm up over the first 500 steps, and linear decay of the learning rate. Training is performed on a single Tesla V100 GPU with 32 GBs of memory and evaluation on a server with a dual 20 core Intel(R) Xeon(R) E5-2698 v4 CPU @2.20GHz and 512 GBs of RAM.

5.2 Results

In Table 2, we report the retrieval performance of our models on the BSARD test set. Overall, the trained bi-encoder models significantly outperform all the other baselines. The two-tower model improves over its siamese variant on recall@100 but performs similarly on the other metrics. Although BM25 underperforms the trained bi-encoders significantly, its performance indicates that it is still a strong baseline for domain-specific retrieval. These results are consistent with those obtained on other in-domain datasets (Thakur et al., 2021).

[3] https://commoncrawl.org/

Train	Model	Encoder(s)	Params	Latency	R@100	R@200	R@500	MAP@100	MRR@100
✗	TF-IDF	-	-	827	40.13	50.44	59.34	8.69	12.98
✗	BM25 (official)	-	-	1342	51.33	56.78	64.71	16.04	24.59
✗	Siamese bi-encoder	word2vec	-	4	49.41	61.76	71.57	12.90	21.49
✗	Siamese bi-encoder	fastText	-	3	32.93	41.33	49.26	6.29	11.78
✗	Siamese bi-encoder	CamemBERT	-	27	4.21	6.00	12.82	0.50	2.04
✓	Siamese bi-encoder	CamemBERT	110M	28	71.63	**78.38**	**83.77**	35.44	**43.52**
✓	Two-tower bi-encoder	CamemBERT	220M	26	**74.78**	78.04	83.39	**35.67**	42.46

Table 2: Retrieval performance (in percent) and query latency (in milliseconds) of various information retrieval approaches on the test set. The best results are marked in bold.

Regarding the zero-shot evaluation of siamese bi-encoder models, we find that directly using the embeddings of a pre-trained CamemBERT model without optimizing for the IR task gives poor results. Reimers and Gurevych (2019) noted similar findings for the task of semantic textual similarity. Furthermore, we observe that the word2vec-based bi-encoder significantly outperforms the fastText and BERT-based models, suggesting that pre-trained word-level embeddings are more appropriate for the task than character-level or subword-level embeddings when used out of the box.

Although promising, these results suggest ample opportunity for improvement compared to a skilled legal expert who can eventually retrieve all relevant articles to any question and thus get perfect scores.

6 Discussion

This section discusses the limitations and broader impacts of our dataset.

6.1 Limitations

As our dataset aims to give researchers a well-defined benchmark to evaluate existing and future legal information retrieval models, certain limitations need to be borne in mind to avoid drawing erroneous conclusions.

First, the corpus of articles is limited to those collected from the 32 Belgian codes described in Table 3 of Appendix A, which does not cover the entire Belgian law as thousands of articles from decrees, directives, and ordinances are missing. During the dataset construction, all references to these uncollected articles are ignored, which causes some questions to end up with only a fraction of their initial number of relevant articles. This information loss implies that the answer contained in the remaining relevant articles might be incomplete, although it is still appropriate.

Additionally, it is essential to note that not all legal questions can be answered with statutes alone. For instance, the question "*Can I evict my tenants if they make too much noise?*" might not have a detailed answer within the statutory law that quantifies a specific noise threshold at which eviction is allowed. Instead, the landlord should probably rely more on case law and find precedents similar to their current situation (e.g., the tenant makes two parties a week until 2 am). Hence, some questions are better suited than others to the statutory article retrieval task, and the domain of the less suitable ones remains to be determined.

6.2 Broader Impacts

In addition to helping advance the state-of-the-art in retrieving statutes relevant to a legal question, BSARD-based models could improve the efficiency of the legal information retrieval process in the context of legal research, therefore enabling researchers to devote themselves to more thoughtful parts of their research.

Furthermore, BSARD can become a starting point of new *open-source* legal information search tools so that the socially weaker parties to disputes can benefit from a free professional assisting service. However, there are risks that the dataset will not be used exclusively for the public interest but perhaps also for profit as part of proprietary search tools developed by companies. Since this would reinforce rather than solve the problem of access to legal information and justice for all, we decided to distribute BSARD under a license with a non-commercial clause.

Other potential negative societal impacts could involve using models trained on BSARD to misuse or find gaps within the governmental laws or use the latter not to defend oneself but to deliberately damage people or companies instead. Of course, we discourage anyone from developing models that aim to perform the latter actions.

7 Conclusion

In this paper, we present the Belgian Statutory Article Retrieval Dataset (BSARD), a citizen-centric French native dataset for statutory article retrieval. Within a larger effort to bridge the gap between people and the law, BSARD provides a means of evaluating and developing models capable of retrieving law articles relevant to a legal question posed by a layperson. We benchmark several strong information retrieval baselines that show promise for the feasibility of the task yet indicate room for improvement. In the future, we plan to build retrieval models that can handle lengthy statutory articles and inherently exploit the hierarchy of the law. In closing, we hope that our work sparks interest in developing practical and reliable statutory article retrieval models to help improve access to justice for all.

Acknowledgments

This research is partially supported by the Sector Plan Digital Legal Studies of the Dutch Ministry of Education, Culture, and Science. In addition, this research was made possible, in part, using the Data Science Research Infrastructure (DSRI) hosted at Maastricht University.

References

Joshua Ainslie, Santiago Ontañón, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 268–284. Association for Computational Linguistics.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Nigel J Balmer, Alexy Buck, Ash Patel, Catrina Denvir, and Pascoe Pleasence. 2010. Knowledge, capability and the experience of rights problems. *London: PLEnet*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adam L. Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu O. Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–199. ACM.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval - 41st European Conference on IR Research*, volume 11437 of *Lecture Notes in Computer Science*, pages 413–428. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lukasz Borchmann, Dawid Wisniewski, Andrzej Gretkowski, Izabela Kosmala, Dawid Jurkiewicz, Lukasz Szalkiewicz, Gabriela Palka, Karol Kaczmarek, Agnieszka Kaliska, and Filip Gralinski. 2020. Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4254–4268. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in english. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4317–4323. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019b. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6314–6322. Association for Computational Linguistics.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6361–6366. Association for Computational Linguistics.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820.

Catrina Denvir. 2016. Online and in the know? Public legal education, young people and the internet. *Computers & Education*, 92-93:204–220.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, Heng Wang, and Zhiyuan Liu. 2019. CJRC: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *18th China National Conference on Chinese Computational Linguistics*, volume 11856 of *Lecture Notes in Computer Science*, pages 439–451. Springer.

Leigh S Estabrook, G Evans Witt, and Harrison Rainie. 2007. *Information searches that solve problems: How people use the Internet, libraries, and government agencies when they need help*. Pew Internet & American Life Project.

Jean-Philippe Fauconnier. 2015. French word embeddings.

Biralatei Fawei, Adam Z. Wyner, and Jeff Z. Pan. 2016. Passing a USA national bar exam: a first corpus for experimentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3373–3378. European Language Resources Association (ELRA).

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomás Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*. European Language Resources Association (ELRA).

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

Margaret Hagan and Yue Li. 2020. Legal help search audit: Are search engines effective brokers of legal information? *Available at SSRN 3623333*.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium*, pages 531–548. USENIX Association.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated nlp dataset for legal contract review. In *Advances in Neural Information Processing Systems 31*.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, volume 2645 of *CEUR Workshop Proceedings*, pages 31–38. CEUR-WS.org.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020*, pages 6769–6781. Association for Computational Linguistics.

Mi-Young Kim, Juliano Rabelo, and Randy Goebel. 2019. Statute law information retrieval and entailment. In *Proceedings of the 6th Competition on Legal Information Retrieval and Entailment Workshop in association with the Seventeenth International Conference on Artificial Intelligence and Law*, pages 283–289. ACM.

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

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing: System Demonstrations, pages 175–184. Association for Computational Linguistics.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11. Association for Computational Linguistics.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric 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, ACL 2020*, pages 7203–7219. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Ha-Thanh Nguyen, Hai-Yen Thi Vuong, Phuong Minh Nguyen, Tran Binh Dang, Quan Minh Bui, Vu Trong Sinh, Chau Minh Nguyen, Vu D. Tran, Ken Satoh, and Minh Le Nguyen. 2020. JNLP team: Deep learning for legal processing in COLIEE 2020. *CoRR*, abs/2011.08071.

Alejandro Ponce, Sarah Chamness Long, Elizabeth Andersen, Camilo Gutierrez Patino, Matthew Harman, Jorge A Morales, Ted Piccone, Natalia Rodriguez Cajamarca, Adriana Stephan, Kirssy Gonzalez, Jennifer VanRiper, Alicia Evangelides, Rachel Martin, Priya Khosla, Lindsey Bock, Erin Campbell, Emily Gray, Amy Gryskiewicz, Ayyub Ibrahim, Leslie Solis, Gabriel Hearn-Desautels, and Francesca Tinucci. 2019. *Global Insights on Access to Justice 2019: Findings from the World Justice Project General Population Poll in 101 Countries*. World Justice Project.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. COLIEE 2020: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops*, volume 12758 of *Lecture Notes in Computer Science*, pages 196–210. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3980–3990. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. CAIL2019-SCM: A dataset of similar case matching in legal domain. *CoRR*, abs/1911.08962.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego,

Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020*, pages 87–94. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34(05), pages 9701–9708. AAAI Press.

Appendix

A Legal Codes

Table 3 presents a detailed summary of the 32 publicly available Belgian codes collected for BSARD.

B Evaluation Metrics

Let $\mathrm{rel}_q(a) \in \{0, 1\}$ be the binary relevance label of article a for question q, and $\langle i, a\rangle \in \mathcal{F}_q$ a result tuple (article a at rank i) from the filter set $\mathcal{F}_q \subset \mathcal{C}$ of ranked articles retrieved for question q.

Recall. The *recall* R_q is the fraction of relevant articles retrieved for query q w.r.t. the total number of relevant articles in the corpus $\mathcal{C}$, i.e.,

$$\mathrm{R}_q = \frac{\sum_{\langle i,a\rangle \in \mathcal{F}_q} \mathrm{rel}_q(a)}{\sum_{a \in \mathcal{C}} \mathrm{rel}_q(a)}. \tag{7}$$

Reciprocal rank. The *reciprocal rank* (RR_q) calculates the reciprocal of the rank at which the first relevant article is retrieved, i.e.,

$$\mathrm{RR}_q = \max_{\langle i,a\rangle \in \mathcal{F}_q} \frac{\mathrm{rel}_q(a)}{i}. \tag{8}$$

Average precision. The *average precision* AP_q is the mean of the precision value obtained after each relevant article is retrieved, that is

$$\mathrm{AP}_q = \frac{\sum_{\langle i,a\rangle \in \mathcal{F}_q} \mathrm{P}_{q,i} \times \mathrm{rel}_q(a)}{\sum_{a \in \mathcal{C}} \mathrm{rel}_q(a)}, \tag{9}$$

where $\mathrm{P}_{q,j}$ is the *precision* computed at rank j for query q, i.e., the fraction of relevant articles retrieved for query q w.r.t. the total number of articles in the retrieved set $\{\mathcal{F}_q\}_{i=1}^{j}$:

$$\mathrm{P}_{q,j} = \frac{\sum_{\langle i,a\rangle \in \{\mathcal{F}_q\}_{i=1}^{j}} \mathrm{rel}_q(a)}{\left|\{\mathcal{F}_q\}_{i=1}^{j}\right|}. \tag{10}$$

We report the macro-averaged *recall* (R), *mean reciprocal rank* (MRR), and *mean average precision* (MAP), which are the average values of the corresponding metrics over a set of n queries. Note that as those metrics are computed for a filter set of size $k = |\mathcal{F}_q| \ll |\mathcal{C}|$ (and not on the entire list of articles in $\mathcal{C}$), we report them with the suffix "@k".

C Dataset Documentation

C.1 Dataset Nutrition Labels

As a first way to document our dataset, we provide the *dataset nutrition labels* (Holland et al., 2018) for BSARD in Table 4.

C.2 Data Statement

In addition to the data nutrition labels, we include the *data statement* (Bender and Friedman, 2018) for BSARD, which provides detailed context on the dataset so that researchers, developers, and users can understand how models built upon it might generalize, be appropriately deployed, and potentially reflect bias or exclusion.

Curation rationale. All law articles from the selected Belgian codes were included in our dataset, except those revoked (identifiable because mentioned before the article or empty content) and those with a duplicate number within the same code (namely, the articles from Act V, Book III of the Civil Code; from Sections 2, 2bis, and 3 of Chapter II, Act VIII, Book III of the Civil Code; from Act XVIII, Book III of the Civil Code; from the Preliminary Act of the Code of Criminal Instruction; from the Appendix of the Judicial Code). Not including the latter articles did not pose a vital concern because none of them were mentioned as relevant to any of the questions in our dataset. Regarding the questions, all those that referenced at least one of the articles from our corpus were included in the dataset.

Language variety. The questions and legal articles were collected in French (fr-BE) as spoken in Wallonia and Brussels-Capital region.

Authority	Code	#Articles	#Relevant
Federal	Judicial Code	2285	429
	Code of Economic Law	2032	98
	Civil Code	1961	568
	Code of Workplace Welfare	1287	25
	Code of Companies and Associations	1194	0
	Code of Local Democracy and Decentralization	1159	3
	Navigation Code	977	0
	Code of Criminal Instruction	719	155
	Penal Code	689	154
	Social Penal Code	307	23
	Forestry Code	261	0
	Railway Code	260	0
	Electoral Code	218	0
	The Constitution	208	5
	Code of Various Rights and Taxes	191	0
	Code of Private International Law	135	4
	Consular Code	100	0
	Rural Code	87	12
	Military Penal Code	66	1
	Code of Belgian Nationality	31	8
Regional	Walloon Code of Social Action and Health	3650	40
	Walloon Code of the Environment	1270	22
	Walloon Code of Territorial Development	796	0
	Walloon Public Service Code	597	0
	Walloon Code of Agriculture	461	0
	Brussels Spatial Planning Code	401	1
	Walloon Code of Basic and Secondary Education	310	0
	Walloon Code of Sustainable Housing	286	20
	Brussels Housing Code	279	44
	Brussels Code of Air, Climate and Energy Management	208	0
	Walloon Animal Welfare Code	108	0
	Brussels Municipal Electoral Code	100	0
Total		22633	1612

Table 3: Summary of the number of articles collected (after pre-processing) from each of the Belgian codes considered for BSARD, as well as the number of articles found to be relevant for at least one of the legal questions.

Speaker demographic. Speakers were not directly approached for inclusion in this dataset and thus could not be asked for demographic information. Questions were collected, anonymized, and reformulated by Droits Quotidiens. Therefore, no direct information about the speakers' age and gender distribution or socioeconomic status is available. However, it is expected that most, but not all, of the speakers are adults (18+ years), speak French as a native language, and live in Wallonia or Brussels-Capital region.

Annotator demographic. A total of six Belgian jurists from Droits Quotidiens contributed to annotating the questions. All have a law degree from a Belgian university and years of experience in providing legal advice and clarifications of the law. They range in age from 30-60 years, including one man and five women, gave their ethnicity as white European, speak French as a native language, and represent upper middle class based on income levels.

Speech situation. The questions were written between 2018 and 2021 and collected in May 2021. They represent informal, asynchronous, edited, written language that does not exceed 44 words. No question contains hateful, aggressive, or inap-

Data Facts

Belgian Statutory Article Retrieval Dataset (BSARD)

Metadata

Filename	articles_fr.csv* questions_fr_train.csv† questions_fr_test.csv‡
Format	CSV
Url	https://doi.org/10.5281/zenodo.5217310
Domain	natural language processing
Keywords	information retrieval, law
Type	tabular
Rows	22633*, 886†, 222‡
Columns	6*, 6†, 6‡
Missing	none
License	CC BY-NC-SA 4.0
Released	August 2021
Range	N/A.
Description	This dataset is a collection of French native legal questions posed by Belgian citizens and law articles from the Belgian legislation. The articles come from 32 publicly available Belgian codes. Each question is labeled by one or several relevant articles from the corpus. The annotations were done by a team of experienced Belgian jurists.

Provenance

Source

Belgian legislation (https://www.ejustice.just.fgov.be/loi/loi.htm)

Droits Quotidiens (https://droitsquotidiens.be)

Author

Name	Antoine Louis
Email	a.louis@maastrichtuniversity.nl

Variables

id*	A unique ID number for the article.
article*	The full content of the article.
code*	The code to which the article belongs.
article_no*	The article number in the code.
description*	The concatenated headings of the article.
law_type*	Either "regional" or "national" law.
id†,‡	A unique ID number for the question.
question†,‡	The content of the question.
category†,‡	The general topic of the question.
subcategory†,‡	The precise topic of the question.
extra_description†,‡	Extra categorization tags of the question.
article_ids†,‡	A list of article IDs relevant to the question.

Table 4: Dataset nutrition labels for BSARD.

propriate language as they were all reviewed and reworded by Droits Quotidiens to be neutral, anonymous, and comprehensive. All the legal articles were written between 1804 and 2021 and collected in May 2021. They represent strong, formal, written language containing up to 5,790 words.

Text characteristics. Many articles complement or rely on other articles in the same or another code and thus contain (sometimes lengthy) legal references, which might be seen as noisy data.

Recording quality. N/A.

Other. N/A.

Provenance appendix. N/A.

C.3 Intended Uses

The dataset is intended to be used by researchers to build and evaluate models on retrieving law articles relevant to an input legal question. Therefore, it should not be regarded as a reliable source of legal information at this point in time, as both the questions and articles correspond to an outdated version of the Belgian law from May 2021 (time of dataset collection). In the latter case, the user is advised to consult daily updated official legal resources (e.g., the Belgian Official Gazette).

C.4 Hosting

We provide access to BSARD on Hugging Face Datasets (Lhoest et al., 2021) at https://huggingface.co/datasets/antoiloui/bsard. Additionally, the dataset is hosted on Zenodo at https://doi.org/10.5281/zenodo.5217310.

C.5 Data Format

The dataset is stored as CSV files and can be read using standard libraries (e.g., the built-in `csv` mod-

ule in Python) or the 🤗 `datasets` library:

```
from datasets import load_dataset
data = load_dataset("antoiloui/bsard")
```

C.6 Reproducibility

We ensure the reproducibility of the experimental results by releasing our code on Github at https://github.com/maastrichtlawtech/bsard.

C.7 Licensing

The dataset is publicly distributed under a CC BY-NC-SA 4.0 license, which allows sharing freely (i.e., copy and redistribute) and adapt (i.e., remix, transform, and build upon) the material on the conditions that the latter is used for non-commercial purposes only, proper attribution is given (i.e., appropriate credit, link to the license, and an indication of changes), and the same license as the original is used if one distributes an adapted version of the material. In addition, the code to reproduce the experimental results of the paper is released under the MIT license.

C.8 Maintenance

The dataset will be supported and maintained by the Law & Tech Lab at Maastricht University. Any updates to the dataset will be communicated via the Github repository. All questions and comments about the dataset can be sent to Antoine Louis: a.louis@maastrichtuniversity.nl. Other contacts can be found at https://maastrichtuniversity.nl/law-and-tech-people.