Knowledge-Based Legal Document Retrieval: A Case Study on Italian Civil Court Decisions

Knowledge-Based Legal Document Retrieval: A Case Study on Italian Civil Court Decisions

Valerio Bellandi[1], Silvana Castano[1], Paolo Ceravolo[1], Ernesto Damiani[1], Alfio Ferrara[1], Stefano Montanelli[1], Sergio Picascia[1], Antongiacomo Polimeno[1] and Davide Riva[1]

[1]*Università degli Studi di Milano, Department of Computer Science, Via Celoria, 18 - 20133 Milano, Italy*

Abstract

In this paper, we present a knowledge-based approach for legal document retrieval based on the organization of a textual data repository and on document embedding models. Pre-processed and embedded documents are iteratively classified at sentence level through a terminology extraction and concept formation cycle, using a zero-knowledge approach that offers a high degree of flexibility with regard to the integration of external knowledge and the variability of inputs, suitable to face the scarcity of annotated data and the specificity of terminology that feature the Italian legal domain document corpora.

Keywords

legal knowledge extraction, legal document retrieval, semantic search, zero-shot learning

1. Introduction

Document retrieval is a daily activity in a wide variety of domains. In the legal domain, retrieval of legal documents, like law articles and court decisions, is important for several categories of legal actors: practitioners (attorneys, lawyers), to support their professional activities; administrators (legislators, judges), to enforce law procedures; users (citizens, organizations), for information exploration and exploitation [1].

To provide effective retrieval functionalities, semantic approaches and knowledge-based systems are being proposed for the legal domain, combining Natural Language Processing (NLP) and context-aware embedding models for extracting and conceptualizing relevant terminology from documents. In particular, to cope with the variety of terminology adopted by judges in the production of legal documents such as court decisions, development of approaches providing word sense disambiguation and the ability to retrieve documents that refer to the same concept despite of the adopted terminology is essential in order to deal with synonyms, circumlocutions, polysemic terms and similar situations. In addition, applying general-purpose

EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, September 26–29, 2022, Bozen-Bolzano, IT

*Corresponding author.

valerio.bellandi@unimi.it (V. Bellandi); silvana.castano@unimi.it (S. Castano); paolo.ceravolo@unimi.it (P. Ceravolo); ernesto.damiani@unimi.it (E. Damiani); alfio.ferrara@unimi.it (A. Ferrara); stefano.montanelli@unimi.it (S. Montanelli); sergio.picascia@unimi.it (S. Picascia); antongiacomo.polimeno@unimi.it (A. Polimeno); davide.riva1@unimi.it (D. Riva)

supervised information retrieval techniques to legal documents would likely be ineffective since they would suffer from the lack of sufficiently large corpora of annotated documents.

In this paper, we present a knowledge-based approach for legal document retrieval based on the organization of a textual data repository and on document embedding models. Pre-processed and embedded documents are iteratively classified at sentence level through a terminology extraction and concept formation cycle. The approach relies on the ASKE (Automated System for Knowledge Extraction) engine, which iteratively solves a multilabel classification problem with zero initial knowledge, that is, without any annotation of the documents. A presentation of the ASKE functionalities is given in [2]. In this paper, we describe an application of using ASKE in a real case study of legal document retrieval from a repository of Italian court decisions as part of the *Next Generation UPP* (*NGUPP*) project, funded by the Italian Ministry of Justice, aiming at providing artificial intelligence and advanced information management techniques for digital transformation of Italian legal processes and digital justice in general. In particular, in this paper we present a case study addressing a practical task faced by law practitioners and administrators, namely the retrieval of past court decisions (so called "precedents") based on one or more text fragments in input, such as sentences, definitions, excerpts of articles. The objective is to retrieve the most pertinent documents (e.g., court decisions or sentences therein contained), i.e. precedents, for the input query.

The paper is organized as follows: Section 2 discusses the related work on legal information retrieval. In Section 3, we describe the proposed approach for legal document retrieval. Section 4 describes the real case study. Finally, Section 5 is devoted to ongoing and future work.

2. Related work

Legal information retrieval (LIR) is the discipline that aims at extracting information from a corpus of legal documents, including case law decisions and legal codes. Since the digital transformation of these documents initiated, LIR has been of interest for both legal actors and information scientists. Boolean searches [3] were the first method applied to accomplish this task, followed by rule-based [4] and NLP-based [5] approaches. Other works focus on the exploitation of external resources, such as ontologies [6] or thesauri [7] combined with natural language processing techniques. With the advent of language models, more elaborated systems have been developed [8], which account for the contextual meaning of words and sentences rather than simply detecting their occurrences.

The main obstacle with legal documents is the lack of sufficient data, especially for languages different from English, that is a crucial aspect given the requirements of neural network-based language models. For this reason, one of the topic of interest for our work has been the zero-shot learning (ZSL) approach. ZSL is a problem setup in the field of machine learning, where a classifier is required to predict labels of examples extracted from classes that were never observed in the training phase. It was firstly referred to as *dataless classification* in 2008 [9] and has quickly become a subject of interest, particularly in the field of natural language processing. The great advantage of this approach consists in the resulting classifier being able to operate efficiently in a partially or totally unlabeled environment.

It is possible to classify ZSL techniques according to three different criteria, as explained

in [10]: the learning setting, the semantic space and the method. Firstly, ZSL can be applied on a completely unlabeled dataset, as in the original paper [9], or on a partially labeled one, like in [11]; with this last approach, called generalized ZSL, the goal of the classifier shifts to distinguishing between observation from already seen classes, and examples from unseen ones. Secondly, one may discern an engineered semantic space from a learned semantic space: the former is designed by humans and can be constructed upon a set of attributes [12] or a collection of keywords [13], while the latter is built on top of the results of a machine learning model, as in the case of a text-embedding space [14]. Finally, ZSL methods can be divided in instance-based [15], whose focus is on obtaining examples for unseen classes, and classifier-based [16], which instead focus on directly building a classifier for unlabeled instances.

Our approach relies on the ASKE engine to classify documents chunks and to extract knowledge from them. This process is enforced in a completely unsupervised environment, therefore eliminating the need for annotated data by operating in a text-embedding space. The employed instance-based method goes under the category of projection methods, which consists in labeling instances by collocating these examples in the same semantic space with class prototypes. A key component in ASKE is Sentence-BERT [17], a modification of BERT language model [18] that is specifically aimed at representing sentence meaning in a vector space. In the legal domain, a specific pre-trained BERT model, known as LEGAL-BERT, has been built by pre-training the original BERT model on several legal corpora [19]. Sentence-BERT, instead, has been used to retrieve legal documents by their embedding similarity [20]. By comparison, few models have been developed for Italian language. LamBERTa model [21] is built upon an Italian pre-trained BERT model by using the Italian Civil Code as a corpus to fine-tune it for article retrieval. By outperforming several other neural models, it proved a promising adaptation of BERT to Italian legal language. However, in our work, we preferred focusing on sentence representation by adopting a pre-trained multilingual Sentence-BERT model.

3. The proposed knowledge-based approach for legal document retrieval

As shown in Figure 1, the proposed approach is organized in three phases. Phase 1 is devoted to data acquisition, organizing a repository in which source documents and associated metadata are stored. Phase 2 deals with textual data preprocessing, where documents are split into chunks, queries are submitted by the user, and the two are mapped into a vector space by a pretrained document embedding model. Finally, in phase 3, the ASKE cycle is performed, consisting in zero-shot classification of documents, terminology enrichment and concept formation. Zero-shot classification is exploited for document retrieval by vector similarity, while the other two steps are used for knowledge extraction and query expansion. Retrieved documents are stored, together with terminological and conceptual knowledge extracted from text, in a graph data structure called ASKE Conceptual Graph (ACG).

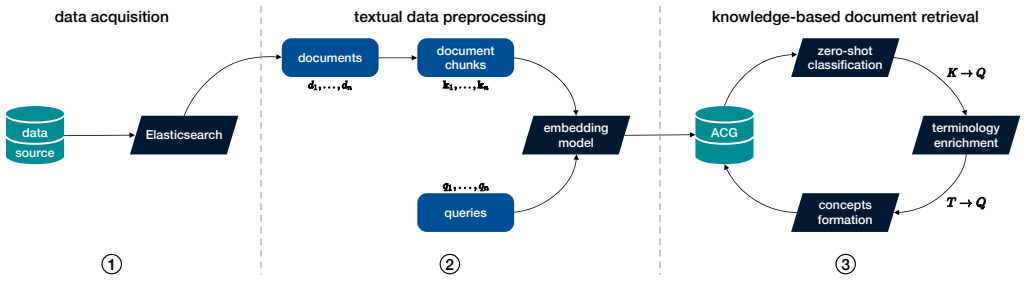

Figure 1: The proposed approach

3.1. Data acquisition

In this work, we consider a set of 1,059,358 public source legal documents with textual content and associated metadata provided by the Italian Ministry of Justice. Approximately 92% is made of court decisions, and the others are ancillary acts of the courts. The source documents were originally organized in folders, subdivided by district of the Italian administration of justice, containing csv files; a single row housed the metadata and text of a document. For each source document, 41 different metadata are provided; some of them, such as the path where text is stored or the id of the user who created the document, contain information used by the document management system of the Courts but poorly relevant for subsequent document retrieval. Other metadata are more content-related and provide useful information for subsequent phases of document retrieval, as per the list below, where numbers in parentheses are the percentages of documents in which they are filled in: the year when the trial started (100%) and the code assigned to it (100%), the year when the decision was enacted (100% of court decisions) and the code assigned to it (100% of court decisions), a code for the trial subject, called the "object code" (100%), the names of the claimant (99.2%) and defendant (98.9%), the name of the judge (100%), first instance or appeal (100%), court and office codes (100%), 8 pieces of data about the first instance trial in case of appeal. The district too can be considered part of the metadata, as, even if not coded in the records, can be deduced by the folder structure. The object code is a hierarchical classification of trial subjects, used by the Italian IT infrastructure of the courts, that considers both the general type (e.g. family) and the specificity of the case (e.g. divorce by mutual agreement).

Starting from source documents and associated metadata, we created an Elasticsearch indexed repository for the data, so that one can extract document sets for any combinations of the metadata for subsequent phases of the approach. The architecture of the repository is based on microservices, exposing specific API for document filtering.

3.2. Textual data preprocessing

Inputs of the textual data preprocessing phase are a corpus of documents $D = \{d_1, \ldots, d_n\}$ extracted from the repository organized in the previous phase, plus a set of queries $Q = \{q_1, \ldots, q_m\}$. Queries can be:

(i) lemmas;

(ii) text fragments, for instance definitions or excerpts of other documents;

(iii) concepts from an existing knowledge base.

While in case (ii) the query will correspond to the submitted fragment and in case (iii) to the concept definition, case (i) requires the system to retrieve all possible definitions of the input lemma from an external knowledge base, not specified by the user.

Documents are firstly tokenized and split into chunks $k_1, \ldots, k_N$. In our case, chunks correspond to sentences, but the analysis may be carried out at a different granularity level. Then stopwords are removed and the remaining terms are lemmatized. The preprocessing phase ends with sentence embedding, in which document chunks $k_1, \ldots, k_N$ and queries $q_1, \ldots, q_m$ are mapped into the same vector space using a pre-trained either language specific or multilingual Sentence-BERT [22] model.

3.3. Knowledge-based document retrieval

Document retrieval is carried out by enforcing the ASKE cycle, which iteratively solves a multilabel classification problem with zero initial knowledge, that is, without any annotation of the documents. The goal is to classify document chunks with query labels $q_1, \ldots, q_m$, which can correspond to words (the input lemmas in case (i) from the previous paragraph and the concept labels in case (iii)) or alphanumeric IDs (case (ii), in which summarizing the input fragments with an unambiguous label may not be possible).

The first step in the ASKE cycle is the zero-shot classification step, in which query labels are assigned to document chunks according to a similarity σ between the chunk embedding $\mathbf{k}$ and the query embedding $\mathbf{q}$:

$$f_Q(\mathbf{k}) = \{q \in Q : \sigma(\mathbf{k}, \mathbf{q}) \geq \alpha\}$$

where threshold α is treated as a hyperparameter: the closer to 1, the fewer query labels will be associated to each chunk. Such a setting may imply that not all chunks that are relevant for a query will be retrieved, while $\alpha \ll 1$ may result in the association of chunks that are not relevant for the query.

While the classification step already produces a set of document chunks that are pertinent to the input queries, the retrieval is also refined by a knowledge extraction and query expansion mechanism consisting of two further steps, performed at each cycle iteration: terminology extraction and concept formation. Given a set of chunks K_q classified with label q, the terminology extraction step considers term lemmas $t \in k, \forall k \in K_q$, it takes and embeds all their possible definitions $s_{t,1}, \ldots, s_{t,L}$ from a predefined knowledge base (e.g. WordNet), and performs term

sense disambiguation by retaining only the definition whose embedding $\mathbf{s}_t^*$ is the closest to the query embedding $\mathbf{q}$. Finally, for each query q, only terms that satisfy the following equation are extracted:

$$T(q) = \{t \in k, \forall k \in K_q : \sigma(\mathbf{s}_t^*, \mathbf{q}) + \sigma(\mathbf{s}_t^*, \overline{\mathbf{K}_q}) \geq \beta\}$$

where β is again a hyperparameter, and $\overline{\mathbf{K}_q} = \frac{1}{|K_q|} \sum_{k \in K_q} \mathbf{k}$ is the centroid of the embeddings of chunks that have been labelled with q.

All terms are finally clustered in what we called the "concept formation" step, and the resulting clusters of terms, referred to as "concepts" and represented by their centroid in the embedding space, will be added to the set of queries in the following cycle iteration. For concept formation, any clustering algorithm will fit, producing different but comparable results, as long as the number of clusters is not fixed a priori and the clusters corresponding to the user-defined queries are always retained. The newly formed concepts are considered as "derived" from the initial queries or from concepts formed at previous iterations.

Terminology extraction and concept formation contribute to query expansion in two ways. First, they add derived concepts to the set of initial queries, enabling the extraction of additional information from the retrieved documents. Secondly, they modify the initial queries themselves. Indeed, at the first iteration a query q is represented solely by its embedding, while at later iterations it will be represented by the centroid of its embedding and the embeddings of the terms clustered with q.

All concepts, document chunks and extracted terms are stored in a graph a data structure called ACG, ASKE Conceptual Graph. ACG includes term-to-chunk belonging, term-to-concept (or term-to-query) relatedness, concept-to-chunk (or query-to-chunk) labelling, and concept-to-concept (or query-to-concept) derivation relationships. The data contained in the ACG at the end of every iteration constitute the input to the subsequent iteration: in particular, concepts are used as new queries. This mechanism enables iterative refinement of the retrieval operation by exploiting knowledge contained in the documents themselves.

In the end, the ACG will contain a set of final concepts, a vocabulary of extracted terms, and the collection of document chunks classified against initial query labels $q_1, \ldots, q_m$ and against new concepts obtained as the result of the ASKE cycle.

4. Application to a real case study of "precedents" retrieval

In this section, we describe a real case study of using ASKE in the legal domain to retrieve the precedents for a given case. This case study has been chosen as it corresponds to a significant and frequent task that judges commonly face in their decision process.

We built a web prototype in order to demonstrate the application of ASKE approach for legal document retrieval. The prototype works as a search engine, which takes in input an initial query (i.e., the input text expressing the precedent search) and which returns a list of document chunks ordered by pertinence as the output. To define the corpus of source documents for the case study, in the project team we decided to focus on the matter of `unfair competition`. The underlying repository is accessed to extract all the documents related to `unfair competition`. A first extraction exploits metadata only, by querying the Elasticsearch

repository using the 8 object codes related to the unfair competition. In this way, an initial corpus of 779 court decisions is built. To overcome situations where court decisions referring to `unfair competition` might have been associated with a different object code metadata (this occurs, for instance, for documents where the decision deals with several different matters), a second repository extraction has been performed to retrieve also documents containing unfair competition in the document text. As a result, a final corpus of 3171 documents is formed. The prototype is designed to support several datasets, such as court decisions related to different fields (e.g., civil law, criminal law), to provide a comprehensive and flexible tool. In the case study, we run the prototype on the *unfair competition* dataset to search related precedents over past court decisions, on the basis of one or more text fragments provided as input by the user. Three main "retrieval-by-input" patterns are envisaged, namely:

1. retrieval by law provision;
2. retrieval by decision chunk;
3. retrieval by keyword.

For the purposes of the case study: pattern (1) is enforced by using provisions from Italian Civil Code, pattern (2) is enforced by using chunks from rulings of the Italian Supreme Court; pattern (3) is enforced by using sets of keywords suggested by legal experts participating in the NGUPP project team. ASKE was applied choosing a multilingual Sentence-BERT model[1] and thresholds $\alpha = 0.2$ and $\beta = 0.2$, which are only partly relevant if we focus on highest similarity results.

Figure 2 shows results ordered by pertinence. Each row in the table represents a pertinent document chunk for which, from left to right, the following information is given to user:

- *query id*: the query to which the chunks has been assigned; since the user can define multiple queries, each query is enumerated by order of definition;
- *text*: the plain text of the document chunk;
- *metadata*: the district, date and id related to the court decision document the chunk belongs to;
- *pertinence*: the a percentage value of similarity between the chunk and the input query.

The user can click on a chunk row, to open an external page showing the full text of the original document from which the chunk has been extracted.

4.1. Preliminary analysis of results

With the help of legal and linguistics experts in the project team, we tried to simulate the typical process that a judge is likely to undertake when looking for precedents. We were advised to use parts of a law or of a court decision as search queries and we evaluated the resulting output, corresponding to pattern (1) of the possible retrieval-by-input patterns previously envisaged.

[1] `distiluse-base-multilingual-cased-v2` from www.sbert.net/docs/pretrained_models.html

Classe	Testo	Distretto	Data	N°Sentenza	Pertinenza
filter					
Q1	"usare nomi o segni distintivi idonei a produrre confusione con i nomi o i segni distintivi legittimamente usati da altri, ovvero imitare servilmente i prodotti di un concorrente, ovvero ancora compiere con qualsiasi altro mezzo atti idonei a creare confusione con i prodotti e con l' attività di un concorrente".	AN	2021-04-21	547	94.90 %
Q1	Ai sensi dell'art. 2598 n. 1 cc commette concorrenza sleale chiunque " usa nomi o segni distintivi idonei a produrre confusione con i nomi o con i segni distintivi legittimamente usati da altri, o imita servilmente i prodotti di un concorrente, o compie con qualsiasi altro mezzo atti idonei a creare confusione con i prodotti e con l'attività di un concorrente".		2014-10-27	14098	93.28 %

Figure 2: Example of the results retrieved over *unfair competition dataset*

Before proceeding further, a couple of remarks are needed. Each of the following retrieval tests has been performed on the original documents in Italian. For the sake of clarity and understandability, hereafter, some text translations, made by the authors, are provided to the reader. Furthermore, for brevity, below we report only on retrieved text chunks with the highest pertinence score.

Test (1): example of result-by-law provision, that is, retrieval is performed using an entire law or part of it as input query. In this test, we give as input article 2598, paragraph 2, of the Italian Civil Code instituting the second type of unfair competition:

> *"[...] any person who disseminates information and appreciations on the products and activities of a competitor, capable of discredit them, or takes possession of the merits of a competitor's products or business."*

As it can be seen from the two retrieved chunks below, ASKE is able to retrieve as most relevant results the ones containing a reference to the input law text, regardless of the way in which this reference is made, either explicitly or implicitly.

> *"Therefore, neither would be the additional case referred to in article 2598 comma 2 of the Civil Code, which refers to the unfair nature of competition in the act or behaviour of those who disseminate news and appreciations of the products and activities of a competitor, capable of discredit them, or takes possession of the merits of a competitor's products or business."*

> *"In order to qualify unfair competition for denigration, it is not necessary that the news and appreciations spread among the public relate specifically to the competitor's products, since they may also have as object also circumstances or opinions more in*

general inherent to the activity of the latter, and therefore also to its organization or to the way of acting of the entrepreneur in the professional field (with the exclusion, therefore, of its strictly personal and private sphere), whose knowledge by third parties is in any case likely to adversely affect the consideration that the company enjoys among consumers."

Test (2): example of result-by-decision chunk, that is, retrieval is performed using a part of the court decision as input query. In particular, this test used a court decision chunk referring to the first of the three types of unfair competition identified by the Italian Civil Code (Article 2598, paragraph 1):

"The reproduction of what has been or could have been the subject of patent for model of utility or ornament cannot, by itself, normally supplement the extremes of unfair competition for slavish imitation, being necessary, for the survival of the latter, that the presentation of the goods - on the basis of a comparative examination of similar products in relation to the degree of diligence and capacity of the average consumer, to whom the goods are intended to be allocated - is carried out in such a way as to mislead the consumer, so that he, wishing to buy the goods of a certain producer, may confuse it with that of a competitor."

Although the evaluation of results retrieved by ASKE for this kind of query could be not immediate, since one needs to understand the meaning and the context of these chunks, the law experts of the project team positively judged the relevance/pertinence of retrieved results, by demonstrating the effectiveness of the proposed approach at this preliminary stage of evaluation, i.e., manual validation by legal experts. In particular, the two most pertinent chunks retrieved are reported below:

"In order to be considered integrated in the case of unfair competition for slavish imitation, it is also necessary to verify the existence of the imitation and the distinctive character of the model, namely its ability to link the product to a certain company, and the ability of imitation behaviour to create confusion in the average consumer regarding the origin of the product."

"Indeed, anti-competitive protection against slavish imitation may be more effective than the one provided by the legislation on the protection of models, since the exclusion of counterfeiting of an ornamental model justified by the fact that what is claimed to be unlawful presents individual character, does not preclude verification, with regard to the same products, of unfair competition for slavish imitation given that the latter is subject to the diligence of the average consumer and not to the higher one required for the informed user in the context of the assessment of individual character."

Finally, a third test has been conducted to experiment the retrieval-by-keyword pattern, defined in Section 4. In this case, keywords like "denigration" (related to Article 2598, paragraph 2) and "slavish imitation" (Article 2598, paragraph 1) have been provided as input query, as suggested by legal experts. Results were less satisfactory than previous cases, in that our

approach aims at capturing the contextual meaning provided by a portion of text, which can not be equally captured through the ASKE cycle from simple keywords or from a combination of few keywords.

As mentioned in Section 3.2, the system may also be allowed to retrieve all definitions of a lemma from an external knowledge base. The following ones are the most similar chunks for the lemmas "imitation" and "denigration" (in Italian) using WordNet as knowledge base to retrieve the displayed definitions:

> Imitation (1): *copying (or trying to copy) the actions of someone else*
>
> Chunk: *"to distinguish the original from the copy."*
>
> Imitation (2): *something copied or derived from an original*
>
> Chunk: *"to distinguish the original from the copy."*
>
> Imitation (3): *the doctrine that representations of nature or human behavior should be accurate imitations*
>
> Chunk: *"The nature, unfavourable or not, of the exemption must be evaluated at the moment when the parties foresee it."*
>
> Denigration (1): *the act of speaking contemptuously of*
>
> Chunk: *"seizure minutes on record"*
>
> Denigration (2): *a communication that belittles somebody or something*
>
> Chunk: *"Communications by certified e-mail"*
>
> Denigration (3): *a false accusation of an offense or a malicious misrepresentation of someone's words or actions*
>
> Chunk: *"on the non-existence of any illicit fact or malicious behavior ascribable to __"*

As in the case of the keywords alone, the results are less satisfying than the case of retrieval by law provision or decision chunk. Not only the similarity scores are far lower, achieving at most 64.2%, but the system appears also to be more likely to retrieve short chunks and to misunderstand the context (as in cases for "denigration"). Moreover, here input queries are not necessarily pertinent to the legal domain, due to the usage of a non-legal knowledge base. All in all, we found that query type 1 (law provisions) and 2 (decision chunks) are generally preferable.

5. Ongoing and future work

In this paper we addressed the problem of document search in the legal domain by introducing a zero-knowledge approach that offers a high degree of flexibility with regard to the integration of external knowledge and the variability of inputs, suitable to face the scarcity of annotated

data and the specificity of terminology typical of the legal domain in general, and of the Italian legal domain in particular.

The work is being developed in the context of an ongoing project and will be further refined and extended. The integration of the approach with external knowledge bases and dictionaries that specifically relate to the legal domain is one practical line for future work, while fine-tuning the underlying language model for the search task will depend on the availability of an annotated corpus. Such data may not only give the possibility of fine-tuning the model, but also enable deeper, mathematically rigorous evaluation, which is often affected by the characteristics of the field of application. The lack of annotated data suitable for our task prevented us from evaluating the performances of the proposed approach in quantitative terms. Despite that, we are currently working on performing tests on datasets devoted to the evaluation of document classification tasks, such as the BBC News dataset [23], a collection of 2225 articles from five topical areas: business, entertainment, politics, sport, and tech. The metrics used for the evaluation is the weighted F1 score, which takes into consideration the different sizes of the classes, reaching a value of 0.89; this can be considered a promising result, given the unsupervised nature of our approach.

Continuation of the collaboration with experts and practitioners of related disciplines, such as law and linguistics, will be crucial to improve our approach in the directions outlined above, and it may prove helpful to discover new improvements to better cope and capture the specificity of the legal documents and the rules commonly adopted by legal actors for their construction.

Acknowledgements

This paper is partially funded by the Next Generation UPP project within the PON programme of the Italian Ministry of Justice.

References

[1] H. Surden, Artificial intelligence and law: An overview, Georgia State University Law Review 35 (2019) 19–22.

[2] A. Ferrara, S. Picascia, D. Riva, Context-Aware Knowledge Extraction from Legal Documents through Zero-Shot Classification, in: Proc. of the 1st ER Int. Workshop on Digital Justice, Digital Law, and Conceptual Modeling (JUSMOD22), Hyderabad, India, 2022.

[3] D. C. Blair, M. E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, Commun. ACM 28 (1985) 289–299. URL: https://doi.org/10.1145/3166.3197. doi:10.1145/3166.3197.

[4] W. Y. Mok, J. R. Mok, Legal machine-learning analysis: First steps towards a.i. assisted legal research, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 266–267. URL: https://doi.org/10.1145/3322640.3326737. doi:10.1145/3322640.3326737.

[5] N. Zeni, N. Kiyavitskaya, L. Mich, J. Cordy, J. Mylopoulos, Gaiust: supporting the extraction of rights and obligations for regulatory compliance, Requirements Engineering 20 (2015) 1–22. doi:10.1007/s00766-013-0181-8.

[6] S. Castano, A. Ferrara, M. Falduti, S. Montanelli, Crime knowledge extraction: An ontology-driven approach for detecting abstract terms in case law decisions, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 179–183. URL: https://doi.org/10.1145/3322640.3326730. doi:10.1145/3322640.3326730.
[7] M. Klein, W. Steenbergen, E. Uijttenbroek, A. Lodder, F. Harmelen, Thesaurus-based retrieval of case law., 2006, pp. 61–70.
[8] W. Hu, S. Zhao, Q. Zhao, H. Sun, X. Hu, R. Guo, Y. Li, Y. Cui, L. Ma, BERT_LF: A similar case retrieval method based on legal facts, Wireless Communications and Mobile Computing 2022 (2022) 1–9. URL: https://doi.org/10.1155/2022/2511147. doi:10.1155/2022/2511147.
[9] M.-W. Chang, L.-A. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: Dataless classification., in: Aaai, volume 2, 2008, pp. 830–835.
[10] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–37.
[11] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2251–2265.
[12] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 951–958.
[13] R. Qiao, L. Liu, C. Shen, A. Van Den Hengel, Less is more: zero-shot learning from online textual documents with noise suppression, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2249–2257.
[14] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 69–77.
[15] X. Xu, T. Hospedales, S. Gong, Transductive zero-shot action recognition by word-vector embedding, International Journal of Computer Vision 123 (2017) 309–333.
[16] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, Advances in neural information processing systems 26 (2013).
[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).
[20] S. Villata, et al., Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents, in: Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, volume 334, IOS Press, 2020, p. 164.
[21] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language

representation models with application to the italian civil code, Artificial Intelligence and Law (2021) 1–57.
[22] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[23] D. Greene, P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in: Proc. 23rd International Conference on Machine learning (ICML'06), ACM Press, 2006, pp. 377–384.