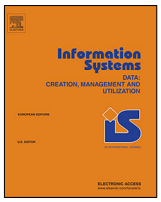
Information Systems 106 (2022) 101967

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is



Legal Information Retrieval systems: State-of-the-art and open issues

Carlo Sansone, Giancarlo Sperlí *

Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, Naples, Italy

ARTICLE INFO

Article history:
Received 15 October 2020
Received in revised form 18 October 2021
Accepted 7 November 2021
Available online 6 December 2021
Recommended by Andrea Tagarelli

Keywords:
Legal Information Retrieval
Artificial Intelligence
Natural Processing Language
Ontology

ABSTRACT

In the last years, the legal domain has been revolutionized by the use of Information and Communication Technologies, producing large amount of digital information. Legal practitioners' needs, then, in browsing these repositories has required to investigate more efficient retrieval methods, that assume more relevance because digital information is mostly unstructured. In this paper we analyze the state-of-the-art of artificial intelligence approaches for legal domain, focusing on Legal Information Retrieval systems based on Natural Language Processing, Machine Learning and Knowledge Extraction techniques. Finally, we also discuss challenges – mainly focusing on retrieving similar cases, statutes or paragraph for supporting latest cases' analysis – and open issues about Legal Information Retrieval systems.

1. Introduction

In the last thirty years, legal domain has been investigated by using different approaches based on Artificial Intelligence (AI) techniques [1,2] such as, for instance, theoretical model, non-monotonic deontic logic, rule-based techniques and expert systems. Their limits concern the number of rules that have to be manually defined for managing an entire domain (e.g., privacy policy or bank contracts) as well as maintaining cost of rules' set over time due to changing standards. In the last five years, the amount of legal documents available in open and machine-readable standard formats (e.g., Akoma Ntoso,[1] LegalXML,[2] LegalRuleML[3]) has exponentially increased due to recent Information and Communication Technologies (ICTs) technological advances. Legal domain is based on a technical language [3,4] that combines common terminologies with domain-specific terms (e.g., 'lease' can be either a place or the lease). In particular, legal documents can use an abstract, formal and legislative language designed or a judicial language that has a large narrative part, on the basis of the document's type.

1.1. Artificial Intelligence-based application in legal domain

There are several applications of AI in the legal domain that must respect the principles of transparency, non-discrimination, privacy-by-design, also according to ethical principles that are often recalled by the European Commission [5,6].

In the legal field there are many applications in which previous legal cases must be analyzed; for instance, it is possible to monitor changes in tax laws and regulations[4] or to use Legal Reasoning and checking compliance [7] for validating the quality of legislative documents. From data analytics point of view, main tasks consist in identifying the most relevant sentences [8,9] (where the concept of relevance is the outcome of legal interpretation), in the analysis of sentences for predicting future outcomes in similar legal cases [10,11], in the exploitation of AI for profiling/classification with the aim to provide probabilistic data of crime recidivism (e.g., *Compas*[5]). Furthermore, applications for the automatic evaluation of small claims is under study in Estonia,[6] or for predicting criminal activity in a certain area of the city.[7]

The legal department is surely the sector in which AI-based applications have been more widely applied. There are tools to support lawyer in drafting[8] and in the consistency control[9] of contracts based on automatically annotated patterns and document databases. It is also possible to use predictive tools based

* Corresponding author.
E-mail addresses: carlo.sansone@unina.it (C. Sansone), giancarlo.sperli@unina.it (G. Sperlí).

1 http://www.akomantoso.org/

2 http://www.legalxml.org/

3 https://www.oasis-open.org/committees/legalruleml/

4 https://www.prnewswire.com/news-releases/deloitte-and-signal-ai-collaborate-to-digitize-tax-regulation-monitoring-with-artificial-intelligence-300865683.html

5 https://www.equivant.com/

6 https://e-justice.europa.eu/content_small_claims-42-ee-en.do?member=1

7 https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

8 https://www.donna.legal/

9 http://www.contractroom.com

https://doi.org/10.1016/j.is.2021.101967
0306-4379/

on relevant previous legal cases[10,11] and e-Discovery tools to identify relevant cases through the myriad of documents[12] or for analyzing loan's relevant data from documents[13]. In particular, the former is a legal search engine that supports practitioners' activities by providing suggestions for common sentences, spelling corrections, specific case names, and filters. In turn, *Vaultedge* is a mortgage automation software with the aim to reduce loan production and due-diligence costs. The aim of the process is to improve borrower experience, also extracting relevant data with an interest confidence score from documents that are, subsequently, compared across different documents to identify discrepancies. More and more automatic bots collect information to reduce the investigation time[14] or support in understanding dispute's context[15]. Finally, AI algorithms are used for identifying solutions to deal with government bureaucracy[16] or to support users in patent applications[17].

In the public administration few applications have been developed due to the difficulty of annotating corpus or conciliating public law and privacy regulation. Furthermore, public administration has found several obstacles (i.e. problems about intellectual property or data for training neural networks) in assigning public data to market players (e.g., Watson IBM or DeepMind). Problems and challenges has been discussed in the report published by experts' group of the Italian *Ministero dello Sviluppo Economico* (MISE) [12] and in the guidelines of the British government [13,14]. There are also applications that support computer-assisted settlement of disputes by increasingly using of sentiment analysis, facial recognition, facial expression analysis to provide an arbitration panel with useful information to better conduct the mediation.[18]

1.2. Main contributions

The legal domain has attracted a lot of interest both for its peculiarities and available resources to define novel document management methodologies. This domain is, in fact, characterized by a huge amount of digital documents, having implicit structures and a peculiar language with unique characteristics, influenced by different factors (i.e. document type, human subjectivity or country).

The large amount of digital documents has made the legal sector interesting for the development of specific methodologies based on natural language analysis (NLP) for the management, storage, indexing and retrieval of legal documents. A concise representation of the legal issues in a document is, then, required for legal practitioners [15] because legal documents have complex and semantically different structures with respect to the domains of interest [16]. In particular, legal documents often use ritual and rhetorical formulas; the former are easy to identify with respect to the latter ones. Legal texts use citations from other norms (e.g., legislative references, quotations from other judgments), therefore some text analysis techniques are strongly limited in their effect if they do not also consider the quoted portion of the text. It is, then, important to design a system to automatic retrieve information or legal documents. This problem can be modeled as an information search task, where a description of the current situation (query) will be used to query the system to retrieve the most suitable information than the input query. However, the relevance of different pieces of information strongly depends on its possible analysis and the needs of the user who will have to use it. For this reason the concept of *semantic search* has assumed a key role about legal documents analysis. One of the main techniques retrieves results on the basis of key words or phrases [17]. Furthermore, different AI models has been designed for several decades due to the growing amount of digital legal documents, also arising an increasing number of open issues such as document classification, information analysis and semantic search.

In this paper we provide an overview about the state-of-the-art of artificial intelligence approaches for legal domain, focusing on Legal Information Retrieval systems using Natural Language Processing (NLP), Machine Learning (ML) and Knowledge Extraction (KE) techniques. In the last thirty years, this topic has been widely studied in different conferences addressing research in Artificial Intelligence and Law as well as it has been the subject of several workshops hosted in different conferences. In particular, the International Conference on Artificial Intelligence and Law (ICAIL) is a biennial conference organized by the International Association for Artificial Intelligence and Law (IAAIL) in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI) since 1987 with the objective of investigating different methodologies and technology demonstrations about the use of Artificial Intelligence models in the legal domain. Furthermore, different competitions about Legal Information Retrieval tasks are emerging to analyze, for instance, the identification of similar cases or sequences of them to support the resolution of new cases (i.e. COLIEE 2019 [18]) or mainly focused on retrieval phase of previous cases or statutes (i.e. FIRE2019 [19]). In the last years, the Competition on Legal Information Extraction/Entailment (COLIEE)[19,20,21] concerns different tasks about the legal case or statute retrieval task as well as identifying paragraph from existing cases that entails the decision of a new case (*Legal Case Entailment* task) or retrieving relevant articles with respect to a given question (*Legal Question Answering* task). Different surveys have been further designed for analyzing this topic from specific point of view as well as legal ontologies [20], digital transformation of legal documents [21] or a cross-language information retrieval [22]. In particular, the motivation of the paper is related to the *Information Legal Management* concept according to [21], which is only focused on the changing nature of the legal information profession. On the other hand, our paper is mainly focused on the analysis of main methodologies and techniques for supporting Legal Information Retrieval (LIR) task. Furthermore, we aim to investigate LIR systems from different point-of-views, also providing a taxonomy of approaches, with respect to [20], which is only focused on the legal ontologies, and to [22], which mainly investigates cross-language information retrieval, a specific sub-field of information retrieval, without focusing on legal data.

This survey has been designed for legal practitioners, also including experts in data science, artificial intelligence and legal informatics interesting in the analysis of legal information.

The paper is organized as follows. Section 2 analyzes Legal Information Retrieval (LIR) systems by investigating main concepts, principles and tasks while Section 3 classifies them in Natural Language Processing, legal ontology and deep learning based approaches, that have been respectively investigated in Section 3.1, 3.2 and 3.3. Challenges and open issues about

10 https://intraspexion.com/softwarepoc
11 https://www.courtquant.com/
12 https://www.judicata.com/
13 https://www.vaultedge.com/
14 https://casestatus.com/
15 https://www.legalmation.com/
16 https://mypadilla.com/
17 https://patentpal.com/
18 https://kleros.io/
19 https://sites.ualberta.ca/~rabelo/COLIEE2019/
20 https://sites.ualberta.ca/~rabelo/COLIEE2020/
21 https://sites.ualberta.ca/~rabelo/COLIEE2021/

Legal Information Retrieval systems have been discussed in Section 4 while Section 5 provides a discussion about findings of the analysis.

2. Legal Information Retrieval

The legal domain has been revolutionized by the use of Information and Communication Technologies (ICTs), that provide benefits both in the legal practice and in the professional attitudes and skills. The increasing application of these technologies has produced large document repositories [23]. More in details, the large amount of *Electronically Stored Information* (ESI) has required the design of methodologies for their processing in order to classify legal content and to retrieve relevant information. Legal practitioners' needs, then, in browsing these large amount of electronic information has required to investigate more efficient retrieval methods, also analyzing statistical methods, that assume more relevance due to the digital information are mostly heterogeneous and have long document size [24], having to take in account legal hierarchy, temporal aspects, importance of quotations, etc [25]. In particular, Van Opijnen et al. [25] underlined the main needs arising from the legal domain that IR systems must satisfy, characterizing legal information according to:

1. *Volume*: in terms of number of documents sharing by both public and private repositories;
2. *Document size*: concerning the length of each document which tends to be longer than other domains;
3. *Structure*: that is very specific for each type of document;
4. *Heterogeneity*: in terms of large variety of document types;
5. *Self-contained documents*: having specific authority since they contain the domain itself
6. *Legal hierarchy*: in terms of a hierarchical organization with regard to documents types and their authority
7. *Temporal aspects*: concerning the search of legal document history or of an applicable law with respect to the analyzed context;
8. *Citations' relevance*: that are more an integral part of the text and arguments with respect to other domains;
9. *Legal terminology*: whose vocabulary are very specific and rich;
10. *Audience*: which is very diversified in terms of different levels of legal knowledge and skills;
11. *Personal data*: since legal memory is often built on names of persons and places;
12. *Multilingualism and multi-jurisdictionality*: as civil laws have a variety of languages based on the countries where they are issued;
13. *Sparseness of legal resources*: because legal information can be found in a variety of resources that have different access regimes and formats.

2.1. Legal Information Retrieval task and main characteristics

In the last years the *Legal Information Retrieval* (LIR) has become one of the main topics in the legal domain [1], whose aim is to model information search as well as it has been performed by legal practitioners to identify useful information for their job [26]. In particular, the LIR covers different tasks, such as electronic document modeling (Discovery of Electronically Stored Information (DESI) or e-discovery), patent retrieval and/or recovery of previous cases.

Mitra and Craswell [27] have defined the main characteristics of Information Retrieval System: (i) semantic Understanding; ii) robustness with respect to different inputs; (iii) robustness to the varying corpus of documents; (iv) robustness as the number of documents increases; (v) robustness to input errors; (vi) sensitivity to context. Specifically, a LIR system should be able to retrieve semantic information from a large data repository, respecting the following constraints: (i) robust to the different ways in which a user might describe their information needs (query); (ii) retrieve information according to the semantic meaning of the examined documents; (iii) retrieve document sequence that interests a legal practitioner. Furthermore, LIR relies on the definition of the relevance concept [25], which considers bibliographic relevance, authority and source quality as relevant criteria.

However, one of the main challenges concerns flexibility of LIR systems with respect to the documents' corpus under examination, also ensuring relevant results according to the user query. First approaches were mainly focused on the classification of information sources according to legal concepts. Nevertheless, the main issues is related to the search results' variety, as, for instance, covering a wide range of possible legal interpretations.

Finally, Fig. 1 provides the number of published papers between 2016 and 2020 about the legal information topic, in which it is easy to note that there is an increase of 5% (from 21% to about 26%) of the ones published on conferences and journals of computer science, proving the relevance of designing a Legal Information Retrieval systems in order to support legal practitioners' needs.

2.1.1. Applications of Legal Information Retrieval systems

Legal Information Retrieval relies on both quantitative (i.e. amount and type of processed data) and qualitative information [25]; in fact, lawyer's tasks include research, drafting, negotiation, consulting, management and argumentation. For this reason, an ideal LIR system should explicitly model the search complexity about legal information. Furthermore, these systems must consider their distinctive features, which, in addition to their enormous volume, include includes documents' size, their structure and heterogeneity types in conjunction with legal hierarchy, temporal aspects, importance of quotations, etc. [25]. For instance, *LexisNexis*[22] and *Westlaw*[23] are popular commercial service providers of legal research, also offering legal, regulatory and business information and analysis for supporting practitioners in the decision making process. Other services, such as the *KeyCite* system in *Westlaw*, has been developed by different firms for tracking citation number and its impact. These systems, whose comparison has been discussed in Table 1, are mainly based on the Boolean information search model, which is used by most search systems.

2.2. Cross-language information retrieval

An important challenge in the legal field concerns the *Cross-Language Information Retrieval* (CLIR) (see [22] for an overview on this topic), a sub-field of information retrieval, since civil laws are written in the language of the country in which they were promulgated. In particular, the aim of this task is to support the retrieval of multi-language documents for supporting legal practitioners' activities. Bonab et al. [28] developed a cross-lingual embedding method, named *Smart Shuffling*, that relies on statistical word alignment approaches to leverage dictionaries for producing dense representation with the aim to overcome neural CLIR issues because the latter mainly translate query terms into related terms in the target language. In turn, Li et al. [29] designed a text representation model based on adversarial learning for target retrieval with respect to which the main CLIR approaches, focusing on models of text representation (i.e. latent semantic

[22] https://www.lexisnexis.com/en-us/gateway.page
[23] https://legal.thomsonreuters.com/en/products/westlaw

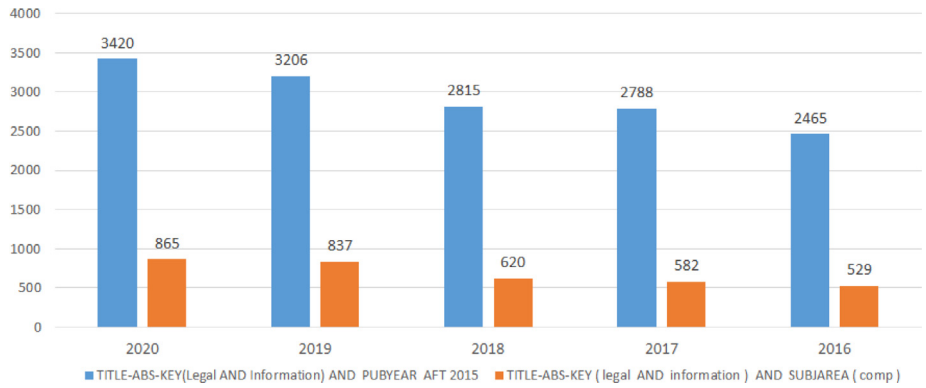


Fig. 1. Number of retrieved papers about *Legal Information Retrieval* between 2016–2020 by performing different queries: (i)*TITLE-ABS-KEY(Legal AND Information) AND PUBYEAR AFT 2015* and (ii)*TITLE-ABS-KEY (Legal AND Information) AND SUBJAREA (comp) AND PUBYEAR AFT 2015*.

Table 1
Comparison between *LexisNexis* and *Westlaw*[a].

Westlaw	LexisNexis
+ Legal database becoming something of a de facto standard in legal research.	+ News Integration exploiting alliance with Down Jones from 2007.
+ 54% of the market for fee-based legal research service subscriptions, also offering free access to law students.	+ Free access to law students
- Expensive w.r.t competitor	+ Cheapest w.r.t competitor

[a]https://lac-group.com/blog/westlaw-vs-lexisnexis-better/.

analysis and vector space model), do not seem to be optimized. In particular, the authors aimed to better preserve the cross-language semantics by combining a matching and translation constraints, which respectively captures main characteristics of cross-language ranking and bridges language gaps, with an adversarial learning strategies. A CLIR technique based on continuous space models, named *external-data composition neural network* (XCNN), has been developed by Gupta et al. [30]. In particular, this model relies on a composition function on top of a deep neural network in order to learn language's continuous space representation while its extension to cross language is obtained by training the model with a small set of parallel data.

In legal domain, Zhebel et al. [31] provided a first attempt to investigate different CLIR approaches, starting from mediator-based method to modern procedures of distributive semantics, on the United Nation (UN) library for retrieving similar documents in different language. Nevertheless the problem of cross-lingual information retrieval in the legal domain is still an open challenge due to the need to study international laws and practices for improving legislation.

3. Classification of Legal Information Retrieval systems

In our opinion, we can classify LIR system according to methodology used for analyzing and retrieving information in the legal domain in the following three categories: *NLP*,*Ontology* and *Deep Learning*-based approaches.

3.1. Natural language processing based techniques

Legal text analysis for extracting useful information has become a challenging task, attracting wide attention in the literature. In particular, the majority of the existing techniques are based on the deontic logic [32], that distinguishes permission or right from obligation and their denials, or on the Hohfeldian system [33], in which each term is combined with an oppositive and a correlative one, for analyzing laws' semantics. The concepts, introduced by deontic logic and Hofeldian system, have been used for different applications in the legal field, as well as legal requirement and compliance rules specification. In particular, two of the first researches stands on the information extraction from legal texts have been proposed by Giorgini et al. [34] and Breaux et al. [35] according to the principles of *deontic* logic. In particular, the former formalizes concepts in terms of rules to model security requirements based on ownership, permission and delegation formalizing new concept through Datalog to verify the correctness and consistency of these requirements. In turn, the latter performs a two-step methodology by using Semantic Parameterization[24], which, firstly, manually identifies rules statement that are, successively, used for extracting rights and obligations from regulations. Nevertheless, these approaches require a high effort to build rules, which are often manually defined.

Furthermore, different approaches based on Natural Language Processing (NLP) have been designed for improving semantic understanding of legal documents. More in detail, they are mainly focused on *Part-Of-Speech* (POS) tagging and *Named-Entity Recognition* (NER) techniques, aiming to assign grammatical tags (i.e. verb, noun or adjective) to text content and to identify keywords related to real-world entities respectively. Classic search engines return an ordered list of documents relevant with respect to a list of keywords, representing the user query in input. The development of NLP-based methodologies for the analysis, indexing and retrieve of large repository of documents has made possible the definition of new approaches to improve the efficiency and consistency of legal systems. In particular, the *Legal Artificial Intelligence* (LegalAI) concerns the use of Natural Processing Language to deal with different legal task by using data-driven methods, as shown in [36] that investigates current state and future direction of LegalAI.

24 Semantic Parameterization has been developed using Grounded theory, stating that the theory, that is inferred from a specific dataset, is valid only for that dataset.

The first approaches focused on document classification to support the retrieval task through two phases: the former concerns feature generation and selection and the latter performs a supervised classification on the basis of identified features. In [37], the authors designed a system for classifying provisions' type in a normative text to easily retrieve these documents. In particular, the *Provision Automatic Classifier* based on *Multiclass Support Vector Machine* has been trained over a vocabulary of words, obtained after a stemming operation and a feature selection, that are weighted according to three different types of weight (binary, term-frequency and term frequency–inverse document frequency (TF–IDF)). In [38], the authors developed a python-based package *LexNLP*, combining NLP pipeline and machine learning algorithms for information extraction from legal and regulatory documents. In particular, this framework aims to segment document and identifying key section, also extracting over eighteen types of structured information and Named Entity Recognition. Soria et al. [39] proposed an automated process for legal texts based on NLP-based pipeline – that is composed by tokenization, normalization, abbreviation and multi-word expression, lemmatized and POS-tagging – whose aim is to classify paragraphs of law according to their normative content and to extract fragments on the basis of semantic rules. A machine-learning analysis has further been designed in [40] for supporting legal practitioners in their needs according to a two steps methodology. Firstly, the logistic regression algorithm has been used to classify sentences on the basis of their keywords into five categories based on contract court decisions and, then, the similarity between legal cases has been computed by using decision trees. The main limitations of these approaches concern the size of the vocabulary, due to the high length of legal documents and the variability of domain-related terminology, and the resulting generated noise in the analysis as well as the lack of context analysis in the classification phase. Furthermore, the variability of language based on document type as well as their structure – containing long and complex sentences, punctuation lists, mix of punctuation and alpha-numeric references – poses different challenges in their analysis and feature extraction phase.

In turn, other approaches aimed to extract relevant concepts or terms from legal documents corpora through the definition of specific rules and constraints. Van Engers et al. [41] discussed about an automatic framework based on linguistic techniques aiming to model extracted norms and concepts into a formal model. In particular, the authors exploited Juridical Language Constructs (JLC) to formalize legal knowledge by defining a subset of legal phrases based on NLP constructs. Approaches based on Natural Processing Language and semantic parsing tree have been proposed in [42] and in [43] for Dutch and Italian regulations respectively. In particular, the former developed a semi-automatic approach that firstly makes explicit references into structured text and, successively, creates models for each individual statement, that are, finally, integrated with each others to build a final model. In turn, the latter designed a pattern-oriented approach aiming to define a formal model of the arguments in the modificatory provision and actions' taxonomy generated by the normative modification for providing formal instruments to the NLP tools. Kiyavitskaya et al. [44] developed a tool for extracting requirements from standards by identifying normative concepts fragments. In particular, each document is parsed for generating a document parse tree and a structural grammar while a domain dependent annotation schema is used to infer annotations, that are selected with respect to a predefined database schema template. In [45], the authors designed a linguistic approach for building deontological rules from regulation combining rule-based and syntatic approach. This approach aims to deal with the problem of syntatic parser in the analysis of document containing long and complex sentences, punctuation lists and alpha-numeric and punctuation references, that are typically in legal documents. In [46] a framework has been developed to extract information from legislative texts combining pattern matching techniques and syntactic lexicon patterns. In particular, the proposed approach relies on syntatic dependencies between extracted terms with a syntatic pattern because pattern matching approach are more robust when length and complex sentences, that are typically in the legal documents, are analyzed. Finally, a SVM binary classifier has been trained for identifying active and passive role, corresponding, respectively, to the active agent and the beneficiary of the analyzed norm, and the related involved object. Sleimi et al. [47] designed a semantic framework based on NLP pipeline for extracting semantic metadata from legal provisions with the aim to support legal experts in the systematic analysis of legal requirements. In particular, the authors firstly performs a classical NLP pipeline (Tokenizer, Sentence Splitter, POS Tagger and Named Entity Recognition) on legal content and they successively performs heuristic analysis based on pattern matching about the defined parse trees. Furthermore, Dragoni et al. [48] proposed an automated framework by combining Natural Processing Language approaches for extracting rules from legal texts. In particular, the authors developed a linguistic information approach based on WordNet[25] for extracting rules from legal documents, also investigating dependencies between content chunks through a logic-based extraction. The main limitations of these approaches concern the definition of the rules and the annotations, which are often manually generated and domain-specific, to extract the relevant information. Therefore, a second limitation concerns rules' updates that often have to integrate with legislative changes that modify the meaning of some terms as well as introduce new concepts to be identified. Furthermore, another challenge has been introduced by the complexity of legal documents' structure that makes it difficult to localize concepts and identify context.

Other approaches aimed to improve the obtained results by combining NLP-based methodologies with lexical resources. Wikipedia has been used in [49,50], where the authors aimed to enhance relevant entities' ranking in response to a given user query and to identify entity search in legal documents, respectively. In particular, both approaches use Wikipedia as a knowledge base, on the one hand, to improve entity ranking by combining content analysis, page linking and category and, on the other hand, it has been used as pivot for identifying web entities with the aim to reduce the analyzed problem to the entity ranking in Wikipedia. Furthermore, Schuhmacher et al. [51] designed a supervised learning approaches to deal with ranking problem as a pairwise classification task, also minimizing the number of discordant pairs, on the basis of feature computed on documents, entity mentions and knowledge base entities. Another ontology-based approach for references' resolution has been proposed by Prokofyev et al. [52]. They developed *SANAPHOR* system that firstly extracts entities, noun phrases and candidate co-references and, successively, an inverted index built on a knowledge base has been used to type noun phrases, whose semantic relatedness is used to split and merge co-reference clusters with the aim to improve the references' resolution. In turn, Flitz [53] proposed a strategy based on data representation of Austrian court decisions and legal norms for building a knowledge graph to deal with search problem. In [54], Natural Language Processing methodology has been also applied in the European project *Lynx* in order to build a Multilingual Legal Knowledge graph for considering semantic information and references to legal documents. Gifford [55] developed *LexrideLaw*, a legal search engine that allows to retrieve appellate cases through relational keywords

[25] https://wordnet.princeton.edu/

Table 2
Synthesis of the main Natural Processing Language based techniques.

Approach	Category	Model	Dataset	LIR Task	Cons
[34]	Rule-based	A framework has been designed for modeling and analyzing security requirements, distinguishing notion of delegation of execution and permission.	Italian public administration dataset	Information Extraction	High effort to build rules, which are often manually defined.
[35]	Rule-based	A methodology based on semantic models has been designed for extracting rights and obligations from regulations.	Rights and obligation	Information Extraction	
[37]	Rule-based	The authors developed a methodology based on semantic analysis for classifying portion of a normative text.	Italian legislative text	Document Classification	Rules and the annotations are often manually generated and domain-specific, to extract the relevant information.
[39]	Rule-based	The authors provide a semantic-based analysis to classify legal text's fragment.	Italian legislative text	Document Classification	
[40]	Rule-based	A machine learning-based analysis for classifying sentences of contract court decision according to types.	Supreme Court of Alabama Court of Civil Appeals of Alabama U.S. District Court of Alabama	Document Classification	
[41]	NLP-based	A methodology based on linguistic techniques has been developed for extracting concept and norm from legal texts.	Dutch Income Tax Law	Information Extraction	The size of the vocabulary, due to the high length of legal documents and the variability of domain-related terminology, and the resulting generated noise in the analysis as well as the lack of context analysis in the classification phase.
[42]	NLP-based	A natural language analysis for parsing sentences in a parse tree combining both generic and law-specific attributes for suggesting fragments for sentences.	Dutch regulations	Information Extraction	
[43]	NLP-based	An approach based on linguistic speech has been designed for managing semi-automatically the consolidation process.	Italian regulations	Information Extraction	
[44]	NLP-based	A methodology based on natural language analysis has been designed for automatic extraction of rights and obligations from regulatory compliance.	HIPPA Privacy Rules	Information Extraction	
[45]	NLP-based	A NLP-based methodology has been developed for extracting rules from regulations.	US Federal Code of Regulations	Information Extraction	
[46]	NLP-based	An approach based on syntactic dependencies between terms has been developed for extracting semantic relation from legislative text.	Annotated legal text	Information Extraction	
[48]	NLP-based	A combination of NLP approaches has been designed to extract rules from legal documents.	Australian Telecommunications Consumer Protections Code	Information Extraction	
[38]	NLP-based	A *LexNLP* framework has been designed by combining NLP pipeline and machine learning algorithm for legal information extraction.	Legal documents	Information Extraction	
[47]	NLP-based	A framework based on NLP pipeline designed for extracting semantic metadata from legal provisions.	Legal documents	Information Extraction	
[49]	NLP and Ontology-based	An approach based on categories and graph-based structure of Wikipedia has been designed for retrieving entities.	INEX 2007 entity ranking track	Information Extraction	The availability of few domain-specific ontologies and the complexity of their managing and updating
[50]	NLP and Ontology-based	The authors developed an approach based on ontologies for retrieving entities.	INEX 2007 entity ranking track	Information Extraction	
[51]	NLP and Ontology-based	The authors designed an approach combining text and ontologies.	Ranking Entities for Web Queries dataset	Information Extraction	
[52]	NLP and Ontology-based	The authors propose an approach based on knowledge graph for co-reference resolution in textual content.	CoNLL-2012 Shared Task on Co-reference Resolution dataset	Information Extraction	
[53]	NLP and Ontology-based	A NLP-based approach has been proposed to model legal data into a knowledge-graph for supporting different tasks.	Austrian court decisions	Information Extraction	
[54]	NLP and Ontology-based	A NLP-based methodology has been designed to build a Multilingual Legal Knowledge graph for considering semantic information and references to legal documents.	Legal documents	Information Extraction	
[55]	NLP and Ontology-based	*LexrideLaw* has been designed to retrieve appellate cases through relational keywords or performing a query expansion operation in a litigation issues ontology.	Legal documents	Information Extraction	

or performing a query expansion operation in a litigation issues ontology. Nevertheless, the availability of few domain-specific ontologies and the complexity of their managing and updating poses some limitations inherent in the applicability of such approaches in real domains. Furthermore, regulatory changes may also require the introduction of new concept or changing ontology entities and their relationships, which is a task done manually making it difficult to update ontologies.

Finally, Table 2 provides a synthesis of the main Natural Processing Language based techniques, also focusing on their limits. In particular, the analyzed approaches have been categorized into Rule, NLP and NLP and ontology based ones according to their methodology used to addressed two main LIR tasks (document classification and information extraction).

3.2. Legal ontology based techniques

Legal systems deal with a large number of concepts, specific terms, legislative documents, frequent amendments and annotations. In the last decades, several methodologies have been proposed to store and retrieve legal documents and the related information, also dealing with the continuous growth of digital data in the legal domain. More in detail, artificial intelligence and the legal field have investigated reasoning techniques based on rules and ontology to support the information search phase.

Ontology represents a "*formal and explicit specification of a shared conceptualization*" [56], that can be built according to two strategies: bottom-up and top-down [57]:

1. *Bottom-Up*: it analyzes most specific concepts for building an ontology. More in detail, the first step of ontology building concerns linguistic study based on existing modules (documents, reports, etc.) in order to extract relevant domain concepts and the relationships between them, also using semi-automatic tools for document analysis. This approach allows the definition of ontologies at high level, that are not often reusable, being domain-specific, also increasing inconsistencies' risk. An example has been discussed in [58], where the authors designed a method, called *Terminae*, based on NLP for building bottom-up ontology. In particular, concepts occurring in the text are extracted for building different micro-ontologies, that are firstly aligned to find a correspondence between concepts and they are successively merged into a unified core ontology. Another example of bottom-up strategy has been discussed by Zhang et al. [59], who built a case ontology for supporting a Chinese legal consultation system in the retrieval of relevant cases or judgments.

2. *Top-down*: it analyzes most generic concepts for building a specialized structure. In this strategy, the process of ontology building begins with the analysis of relevant information sources relevant for the domain under examination while the modeling of higher-level concepts will be refined in the next steps. These approaches are manually performed by domain experts, that define reusable and shareable high-level ontologies [57]. An example has been discussed in [60], where the authors designed a top-down ontology building methodology for modeling and sharing knowledge among users with the aim to support different tasks.

Leone et al. [61] analyzed state-of-the-art in legal ontologies, focusing on distincive features, for assisting users and law experts in choosing the legal ontology more appropriate for their aim. The *Top Ontology of the Law* ontology [62] describes a particular "*view of law as a dynamic system of states of things, which are connected by events and rules*". In turn, *CausatiOnt* [63] is an as-is ontology[26] developed to conceptualize and address legal liability issues, that relies on three main classes: *Category*, *entities* and *dimension*.

Several ontologies have been designed to model the heterogeneity of the legal domain, also encouraging the reuse of legal information, in order to support decision-making process (as discussed in [20]). Modern LIR systems, therefore, infers on domain ontology by using their constructs (i.e. class hierarchy, object and relationship property), also modeling semantic object in the legal documents for performing further inference. Different prototypes and tools based on ontologies have been proposed for legal cases simulations in the criminal field (i.e. *LEGIS* [64–66] and *CORBS* [67]) and benefits about the Dutch unemployment law (*FRAMER* [68]). The former is based on the *Unified Foundation Ontology* (UFO) [69], that provides constraints and rules for modeling ontologically valid models through a rich axiomatization of the chosen vocabulary. In particular, *UFO*'s categories include both the identity principle, that allows to judge two entities as being the same, as well as the rigidity concept, that determines which type can be instantiated with respect to the context. In turn, *CORBS* is a criminal ontology, defined by a set of logic rules using the *Semantic Web Rule Language* (SWRL)[27], with the aim to perform rule-based reasoning for supporting legal experts' analysis. In particular a middle-out approach has been designed by combining top-down and bottom-up strategies to build a modular ontology, that is composed by four components: upper, core, domain and domain-specific. The first two modules have been built by using top-down strategies based on *UFO* and legal core (*LKIF-core*) ontologies whilst a bottom-up approach has been designed for modeling the last two modules, that are defined on the domain of interest. Finally, *FRAMER* has been designed as a legal knowledge prototype, whose domain knowledge has been modeled through *PROLOG*,[28] for performing assessment tasks on the Dutch Unemployment Benefits Act.

In turn, *NM-L* [70] is a high-level extension of the legal domain that represents an intermediate view between the top level and the central level. In particular, it models both physical and conceptual entities, classified them into animate (i.e. person, organization or electronic agents) and inanimate categories, while the occurrences are dived into mental, environmental and social ones. Furthermore, *NM-L* ontology extends the *NM* one in terms of roles for further considering a legal person that represents another one in the context of an occurrence. Finally, Legal Requirements ontology [71] defines "*legal concepts used to specify the legal requirements for the organization's compliance requirements*", that are typically used for classifying legal statements using statement-level concepts. In particular, these concepts are defined as state that an entity can achieve (*Permission*) or is required to achieve (*Obligation*) as well as state that an entity is required to avoid (*Refrainment*) or is not permitted to achieve (*Exclusion*). Furthermore, a middle-out approach has been designed by El Ghosh et al. [72] for building an ontological model for criminal domain, also formalizing rules based on it. In particular, the authors aims to reuse foundational and legal core ontologies to build up the criminal ontology. Castano et al. [73,74] developed an approach, named *CRIKE*, relying on the LATO ontology, that formalize legal abstract terms, to identify terminology within case-law decision contents to characterize concrete abstract-term instances. In [75], the authors designed a service contract ontology to formalize contract language, also extending the *ArchiMate* language to reflect the ontology's elements.

Other ontologies have also been designed for specific purposes, although it compromises their reusability in other contexts, such as JurWordnet [76] and the taxonomy proposed by Ajani et al. [77]. The semantic integration [78] and the semantic retrieval of legal documents [79–81] complete the domain specific ontologies.

Different researches have been carried out for the representation and formalization of legal knowledge (see for instance [82]). Several ontologies have been developed to describe different legal contexts [83,84]. Let us consider two main initiatives, LKIF [85–87] and LegalRuleML [88,89], which are probably the major attempts to harmonize legal concepts.

LKIF [85–87] is composed by 200 classes for modeling different legal contents through rules, also providing concepts for events' background and consequences. In turn, LegalRuleML [88,89] adapted the *RuleML* language for legal domain, classifying statements into facts – that is composed by constitutive, prescriptive and penalty declarations – and legal rules (constitutive, technical and prescriptive), that can have multiple semantic annotations, each one associated with a different legal interpretation.

Table 3 provides a synthesis of the main ontology-based techniques with respect to used category and LIR task, also focusing on their limits. In particular, the analyzed approaches have been categorized into Top-down, bottom-up and hybrid based ones according to their methodology used to addressed two main LIR tasks (information retrieval and extraction).

Nevertheless, the availability of few domain-specific ontologies and the complexity of their managing and updating poses different challenges inherent in the applicability of the ontology in real domains. Furthermore, the dependencies of legal terms on the context in which they are used makes it difficult to model them and integrate different domain-specific ontologies. Finally, regulatory changes may also require the introduction of new concept or changing ontology entities and their relationships, which is a task done manually making it difficult to update ontologies.

3.3. Deep learning based techniques

The amount of legal information produced daily in digital format is enormously increasing in the legal domain. In the last years, machine learning and deep learning models has attracted wide attention for processing legal documents. Legal documents classification, translation, summary, contract revision, case forecasting and retrieval of information are the main research tasks.

Semantic search has also benefited from the use of Machine Learning techniques for context identification and words analysis. More in details, these approaches are based on the assumption that words used in similar contexts are very likely to be semantically similar. The first approaches based on this assumption (see

26 That is particular ontologies that is only composed by "AS-IS" relationships.

27 https://www.w3.org/Submission/SWRL/

28 https://www.swi-prolog.org/

Table 3
Synthesis of the main Ontology-based techniques.

Approach	Category	Model	Dataset	LIR Task	Cons
[58]	Bottom-up	*TERMINAE* construction method has been designed to build a legal ontology by merging domain-specific micro-ontologies.	European Union Council directives	Information Retrieval	The availability of few domain-specific ontologies and the complexity of their managing and updating poses different challenges inherent in the applicability of the ontology in real domains.
[59]	Bottom-up	The authors developed a Chinese legal consultation system based on legal ontologies, built by using a bottom-up approach to integrate statutes and judicial precedents.	Legal cases	Information Retrieval	
[57]	Top-down	An ontology has been defined on the basis of expert judgment, that can be aligned with other ones to capture similar or complementary knowledge.	Law multi–lingual corpus	Information Retrieval	
[67]	Bottom-up	The authors defined an ontology for modeling legal norms in order to support the legal reasoning.	Legal cases	Information Retrieval	
[68]	Top-down	The authors proposed a generic legal ontology from which statute-specific ontologies have been defined for building legal systems.	Legal cases	Information Retrieval	
[70]	Top-down	The authors extended the NM-L ontology in the legal domain for supporting human understanding and reasoning.	Legal Cases	Information Retrieval	
[71]	Top-down	The authors defined the legal requirements upper ontology for classifying statements in a legal text.	Health Insurance Portability and Accountability Act	Information Retrieval	
[76]	Top-down	The authors developed an extension of EuroWordNet for the legal domain (Jur-Wordnet).	Legislative text	Information Retrieval	
[77]	Bottom-up	European Legal Taxonomy Syllabus has been defined as a multi-lingual and multi-jurisdictional terminological vocabulary connected through semantic relations.	Legal text	Information Retrieval	
[80]	Bottom-up	The authors defined a system for information extraction from Vietnamese legal cases using ontologies.	Vietnamese legal cases	Information Retrieval	
[81]	Bottom-up	The *SCRO_II* algorithm has been defined for identifying a set of synonyms and relations.	Thai Succession Law and Bill of Exchange Law	Information Retrieval	
[72]	Both	The criminal ontology has been built by using a middle-out approach by reusing foundational and legal core ontologies.	Legal Norms	Information Retrieval	
[60]	Top-down	An ontology has been defined on the basis of top-down approach evaluated techniques, implementation and role of formal languages.	Legal cases	Information Extraction	Legal document can use an abstract, formal and legislative language designed or a judicial language that has a large narrative part, underlying their volume, complex structures and domain specific language.
[64]	Bottom-up	A bottom up methodologies has been designed to build an ontology for defining vehicles' categories in an unambiguous manner.	Brazilian legal codes	Information Extraction	
[65]	Bottom-up	A bottom up methodologies has been designed to build an ontology for the representation of crimes against property.	Brazilian penal codes	Information Extraction	
[78]	Top-down	The author defines an ontology, named *Legal-RDF* to jointly, model layout and content of legal-document and metadata.	Legal text	Information Extraction	
[79]	Top-down	The authors developed a method for information extraction from legal text using domain-specific ontologies for capturing their structure and language.	Spanish notary act	Information Extraction	
[75]	Bottom-up	A service contract ontology has been designed to formalize contract language.	Legal contract text	Information Extraction	
[73,74]	Bottom-up	*CRIKE* methodologies relies on *LATO* ontology has been designed to identify terminology within case-law decision contents to characterize concrete abstract-term instances.	Case-law decision text	Information Extraction	
[85–87]	Top-down	*LKIF* ontology has been defined for supporting the knowledge interchange between legal knowledge systems.	Legal text	Information Extraction	
[88,89]	Top-down	*LegalRuleML* has been designed for classifying statements into facts and rules.	Legal text	Information Extraction	

for instance [90]) defined multidimensional matrices to model the coexistence of words, represented as vectors, in a given context. In the last years, Mikolov et al. [91] have proposed two architectural models (*Feedforward Neural Net Language Model* (NNLM) and *Recurrent Neural Net Language Model*. (RNNLM)) for the efficient computation of words' vector representation within large document corpora. In particular, the proposed methodology aims to maximize the probability of words' co-occurrence in a given context window using neural networks architectures. Furthermore, deep learning techniques has been designed for different task: in [92] word2vec embedding has been used for improving Named Entity Recognition, Grbovic et al. [93] combined context and content information for performing query expansion to improve information retrieval task or legal document ranking has been discussed by Nalisnick et al. [94] through a Dual Embedding Space Model (DESM), that map query words and document into the input and output space respectively.

As shown in [95], some approaches about legal domain have been designed by incorporating relevant word representation. We classified the proposed approaches about embedding models into the following three classes:

- Pre-trained embedding models using *word2vec*, *glove* or *fastext*. The main limitation is that these models are trained on generic corpora (i.e. news articles, or web pages). Therefore, these models do not properly capture the semantics of legal

texts because they are trained on generic datasets, that are not related to the domain of interest, such as legislation, case law, and other legal documents.
- Domain specific embedding models. Several approaches have been trained on manually annotated data-sets or a larger collection of documents relevant for the domain of interest. Nevertheless, although these approaches seem to improve the results, these models may suffer from poor generalization because they are trained on controlled datasets avoiding the noise present in generic models.
- Hybrid embedding models. Different researchers have combined generic and domain-specific embedding models for extracting features that can be used by neural network-based architectures to improve understanding of documents' content. This is a common practice when the corpus of documents is small for guaranteeing a minimum level of quality representation of the words in the document.

Nevertheless, deep learning-based approaches is becoming more and more an emerging topic in the legal domain as shown in [96], in which the authors compare different traditional (i.e. TF–IDF and LDA) and embedding (i.e. BERT) methodologies for textual similarity task in the domain of legal information retrieval. Another analysis of deep learning models on the Ontario Court of Appeal has been discussed in [97] to predict employment notice, investigating different adaptations of pretrained BERT models. Bansal et al. [98] also provide a comprehensive study of different deep learning models (i.e. CNNs, RNNs, LSTM and GRU) for different tasks in legal domain. In [99], the authors developed a framework based on deep learning model, relying on a pre-trained BERT (specifically *BERT-base-uncased*) model on which a fine tuned strategy has been performed, for legal retrieval task. Another approach based on BERT (specifically BERT XL) deep learning model has been designed by Mokanov [100] in order to provide legal qualification to a set of facts about Canadian administrative tribunal decisions, predicting the most relevant source of law for each piece of content. Shao et al. [101] developed *BERT-PLI*, a pre-trained BERT model on which a fine tuning strategy has been performed on a sample of law entailment dataset, with the aim to consider the semantic relationships at the paragraph levels for inferring relevance between two cases. In [102], the authors analyzed BERT model for legal domain, also investigating pre-trained and fine tuning strategies although they do not always generalize well in this domain. Furthermore, the authors released *LEGAL-BERT*, composed by BERT-BASE with 12 layers, 768 hidden units and 12 attention heads (110M parameters), for supporting legal NLP research and technology applications.

Sugathadasaet et al. [103] developed a domain-specific LIR system based on three different models: i) Node2vec, ii) sentence similarity and iii) representing legal domain in a vector space. These models are, then, combined to incorporate domain specific semantic similarity measure, showing an improvement of accuracy in the information search phase with respect to traditional search systems, that find an exact match with a given input string. In [104], the authors developed a LIR system, built using word2vec, also analyzing semantic patterns for each search query. Furthermore, the authors stated that the use of NLP pipeline for pre-processing could improve the system performance. Furthermore, Vu et al. [105] designed a methodology, that combines lexical and latent features, for legal information retrieval by encoding documents in continuous vector space using deep neural networks. In [106], the authors designed a methodology for recommending similar legal cases using deep learning model based on word2vec with the aim to extract k keywords of the fact. In particular, a representative central word vector of each training case, built by analyzing the extracted keywords, has been used for retrieving similar legal cases having most similar vector to the test case one. In turn Do et al. [107] proposed a two stage framework with the aim to deal with legal information retrieval task, in which, firstly, the pair of query and article are ranked through a SVM ranking, whose output are used to train a Convolutional Neural Network for question answering task. Another important issue concerns the named-entity linking for supporting legal retrieval cases has been discussed in [108]. In particular, the authors designed a methodology by combining entity embedding and local model with neural attention to jointly consider semantic meaning of entities and contextual words.

Wei et al. [109] described preliminary studies about the use of deep learning techniques, focusing on word embedding model as input of Convolutional Neural Network (CNN) classifier, with respect to SVM algorithm on four legal corpora for text classification in the revision of legal documents. Undavia et al. [110] developed a classification model (*Supreme Court Classifier (SCC)*) to classify legal texts in predefined document categories, also comparing machine learning algorithms with respect to NN-based systems on the Washington University School of Law Supreme Court Database (SCDB). Hammami et al. [111] investigated the use of Convolutional Neural Networks, whose input is a matrix generated by using Word2Vec (having fixed dimension equal to 200) on legal corpus, for French legal content classification. In [112], the authors developed a deep learning architecture based on *RoBERTa* [113] for multi-label legal document classification using label embedding and multi-task learning strategies. In particular, this approach has been evaluated on *POSTURE50k*, a legal extreme multi-label classification dataset, containing 50,000 legal opinions with the related labeled legal procedural postures. Another classification system for the Brazilian court document has been implemented by Da Silva et al. [114], where the authors implemented the CNN network on embedded text, which is computed on content extracted from legal documents through Optical Character Recognition (OCR) and NLP pipeline. However, this approach is mainly influenced by the performance of the OCR, whose results are typically affected by noise.

Finally, Table 4 summarizes deep learning-based methodologies applied to legal domain with respect to the used category and embedding model. In particular, the analyzed approaches have been categorized into Top-down, bottom-up and hybrid based ones according to their methodology used to addressed two main LIR tasks (information retrieval and document classification).

4. Challenges & open issues

In the last years, different challenges about information retrieval in legal domain are increasing on the basis of several tasks' definition. The Competition on Legal Information Extraction/Entailment (COLIIE-2019 [18]) is one of the main challenge, that proposes four tasks: i) identifying supporting cases w.r.t a new one; ii) identifying paragraph from previous cases for supporting new case decision; iii) extracting civil code articles to answer a user question and iv) identifying entailment relationship for relevant articles. Vu et al. [115] developed a summarization methodology combining latent features, extracted from document by encoding in continuous vector space, with lexical features from different part of input query to deal with first task in COLIEE-2019. A model based on deep learning, named *BERT-PI*, has been proposed by Shao et al. [116] to unveil semantic relationships at paragraph-level for inferring relevance between cases. Rabelo et al. [117] deal with four task by proposing a framework for unveiling entailment relationships between legal cases. More in details, it has been defined according to similarity measures and transformer-based technique in conjunction with post-processing strategy based on a priori probability, that

Table 4
Synthesis of the main Deep Learning-based methodologies for Legal Information Retrieval.

Approach	Category	Model	Dataset	LIR Task	Cons
[103]	Pre-trained embedding model	Word embedding+ Ensemble	Legal Corpora	Case Retrieval	The size of the dataset on which the model is trained and/or evaluated. The use of pre-trained embedding model that often do not capture heterogeneous semantic information and specific legal domain terms.
[104]	Pre-trained embedding model	Word embedding+ Semantic pattern analysis of search query	EU Data Protection Directive 94/46/EC	Case Retrieval	
[105]	Domain specific embedding model	Latent, generated by CNN, and lexical features, computed on the basis of skipgram and n-gram, and MLP	COLIEE-2019 dataset	Case Retrieval	
[108]	Domain specific embedding model	Methodology based on entity embedding and local model with neural attention for jointly considering semantic meaning of entities and contextual words.	European Union law	Named Entity Linking	
[107]	Domain specific embedding model	A two stage framework relying on SVM ranking and CNN for legal cases retrieval.	COLIEE-2016 dataset	Case Retrieval	
[106]	Domain specific embedding model	A methodology based on word embedding for recommending similar legal cases.	CAIL-2018 dataset	Case Retrieval	
[99]	Domain specific embedding model	A framework based on pre-trained BERT on which a fine tuning strategy has been performed.	COLIEE-2019 dataset	Case Retrieval	
[100]	Domain specific embedding model	Framework based on BERT deep learning model for predicting the most relevant source of law for each piece of content.	Canadian administrative tribunal decisions	Case Retrieval	
[101]	Domain specific embedding model	*BERT-PLI* aims to capture the semantic relationships at the paragraph levels for inferring relevance between two cases.	COLIEE 2019 dataset	Case Retrieval	
[102]	Domain specific embedding model	Legal BERT has been designed for supporting legal NLP research and technology applications.	EURLEX57K	Case Retrieval	
[109]	Pre-trained embedding model	Word embedding+CNN	Legal Corpora	Document Classification	The use of Deep learning standard model for this task and the size of the examined dataset.
[110]	Pre-trained embedding model	Word embedding+CNN	University School of Law Supreme Court Database (SCDB)	Document Classification	
[114]	Pre-trained embedding model	Word embedding+CNN	Brazilian Judicial Documents	Document Classification	
[112]	Pre-trained embedding model	Deep learning architecture based on RoBERTa using label embeddings and multi-task learning strategies for legal document classification.	POSTURE50k dataset	Document Classification	
[111]	Pre-trained embedding model	Word embedding+CNN	French Judicial Documents	Document Classification	

is computed on training samples' data distribution. Finally, an inter-paragraph entailment approach has been proposed by Kim et al. [118] to deal with Task 3 and 4 of COLIEE-2019 by analyzing law structure and patterns to predict entailments, also performing an heuristic attributes selection.

Another challenge, named *FIRE 2019 AILA* [19], is mainly focused on the framework design for Legal Information Retrieval by identifying most relevant cases (Task 1) and statutes (Task 2) on the basis of query input. More et al. [119] deals with first task by proposing a Named Entity Recognition approach to perform a pre-processing stage for documents and input query for improving relevant cases retrieval. The second task has been investigated by Lefoane et al. [120], that, firstly, compared different term weight models, whose best one is, then, used for unveiling how statutes' title and description affect retrieval effectiveness. For dealing with both tasks in *FIRE 2019 AILA*, Mandal et al. [121] designed an unsupervised methodology using cosine similarity between input query and target documents, encoded by pre-trained word embedding.

Furthermore, different frameworks have been analyzed in Section 2, whose goal is to support legal practitioners in their activities. Nevertheless, these systems are focused on a specific

task (i.e. *Vaultedge* for Mortage automation), in which documents have a specific structure, or they support information search (i.e. *WestLaw*, *LexisNexis* and *Judicata*), mainly relying on the analysis of document text and/or citation network structure, which provide results or summaries that are, often, not satisfactory to law practitioners. In particular, it is still possible to identify different open challenges related to the understanding of legal case documents' structures, that improves the segmentation of information to support different tasks (i.e. document retrieval, similarity or summarization), or to the legal document summarization, that simplifies their fruition since legal texts are often long size with many legal terms and domain-related terminology, or to the legal document search and recommendation, that also considers domain knowledge inherent of a particular jurisdiction for supporting prior cases retrieval.

Nevertheless, the examined techniques suffer from different problems concerning the implementation of a Legal Information Retrieval system. One of the main problem concerns legal cases' variety, in terms of large amount of digital information in addition to provide different variants for each case, whose choice requires a thorough understanding of the case and the law; in fact, the *relevance* concept of a case or paragraph is strongly related to legal interpretation, as shown in [25]. Furthermore, laws are frequently modified, making some previous cases inappropriate, and requiring an additional effort of discernment to identify their updated version. Laws are also not independent, but it is necessary to manage the regulatory inter-relationships from which consequences and non-obvious constraints arise for the drafting of the acts.

Other issues concern needs about domain experts for knowledge modeling (as discussed in Section 3.2) and/or rules for extracting useful information (as analyzed in 3.1 and 3.3) from previous cases to support current analysis. LIR systems also need to formalize legal application domain for providing "semantic" analysis and not only "syntactic-grammatical". Finally, the variability of laws requires a further effort to update knowledge modeling that is often performed manually or semi-automatically through a significant contribution of human experts to validate the related formalization.

5. Conclusions

The large amount of electronically stored information (ESI), which is mostly are mostly heterogeneous and have long document size, has required the development of methodologies for document processing and for extracting useful information with the aim to improve relevant information retrieval.

In this paper we provided an analysis about the state-of-the-art on artificial intelligence on the legal domain, focusing on the automatic processing of the legal content using Natural Processing Language (NLP), Machine Learning (ML) and Knowledge based approaches, falling under the umbrella term *Legal Information Retrieval*. In particular, we firstly described the relevance factor of the legal information according to [25] – underlying their volume, complex structures and domain specific language – and we, successively, analyzed the main features of an Information Retrieval system, also providing an overview of the *Cross-Language Information Retrieval* in the legal domain. We, then, classified NLP-based LIR approaches into three different categories: the rule-based one, whose main limitation is the manual definition of the rules, the NLP-based approaches, dealing with the vocabulary size due to document length and the domain-related terminology, and the latter combining NLP and ontologies, whose main limitation is due to the availability of few domain-specific ontologies and the complexity of their managing and updating. Furthermore, ontology-based technique have been investigated in Section 2.2 by describing how they are constructed (top-down and bottom-up strategies) and focusing on main domain ontologies (*LKIF* and *LegalRuleML*). Finally, we analyzed deep learning-based techniques that, although they show promising results, typically use pre-trained embedding model due to few number of training samples and the difficulty of capturing the semantic information of documents and specific legal terms. We further analyzed challenges that are arising in the last years to deal with different Legal Information Retrieval task by discussing proposed approaches and the related open issues. In particular, different challenges are still opened about the understanding of legal case documents' structures, legal document summarization and legal document search and recommendation, also considering the specific structure of each document type (i.e. codes, case law and articles).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Bench-Capon, M. Araszkiewicz, K. Ashley, K. Atkinson, F. Bex, F. Borges, D. Bourcier, P. Bourgine, J.G. Conrad, E. Francesconi, et al., A history of AI and law in 50 papers: 25 years of the international conference on AI and law, Artif. Intell. Law 20 (3) (2012) 215–319, http://dx.doi.org/10.1007/s10506-012-9131-x.

[2] G. Governatori, A. Rotolo, R. Riveret, A deontic argumentation framework based on deontic defeasible logic, in: International Conference on Principles and Practice of Multi-Agent Systems, Springer, 2018, pp. 484–492, http://dx.doi.org/10.1007/978-3-030-03098-8_33.

[3] L.T. McCarty, A language for legal discourse I. Basic features, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, in: ICAIL '89, Association for Computing Machinery, New York, NY, USA, 1989, pp. 180–189, http://dx.doi.org/10.1145/74014.74037.

[4] K.D. Ashley, Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age, Cambridge University Press, 2017, http://dx.doi.org/10.1017/9781316761380.

[5] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nat. Mach. Intell. 1 (9) (2019) 389–399, http://dx.doi.org/10.1038/s42256-019-0088-2.

[6] P.M. Asaro, AI ethics in predictive policing: From models of threat to an ethics of care, IEEE Technol. Soc. Mag. 38 (2) (2019) 40–53, http://dx.doi.org/10.1109/MTS.2019.2915154.

[7] M. Palmirani, G. Governatori, Modelling legal knowledge for GDPR compliance checking, in: M. Palmirani (Ed.), Legal Knowledge and Information Systems - JURIX 2018: The Thirty-First Annual Conference, Groningen, the Netherlands, 12-14 December 2018, in: Frontiers in Artificial Intelligence and Applications, vol. 313, IOS Press, 2018, pp. 101–110, http://dx.doi.org/10.3233/978-1-61499-935-5-101.

[8] A. Kanapala, S. Pal, R. Pamula, Text summarization from legal documents: A survey, Artif. Intell. Rev. 51 (3) (2019) 371–402, http://dx.doi.org/10.1007/s10462-017-9566-2.

[9] A. Kanapala, S. Jannu, R. Pamula, Summarization of legal judgments using gravitational search algorithm, Neural Comput. Appl. 31 (12) (2019) 8631–8639, http://dx.doi.org/10.1007/s00521-019-04177-x.

[10] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European court of human rights, Artif. Intell. Law 28 (2) (2020) 237–266, http://dx.doi.org/10.1007/s10506-019-09255-y.

[11] K. Atkinson, T. Bench-Capon, Reasoning with legal cases: Analogy or rule application? in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 12–21, http://dx.doi.org/10.1145/3322640.3326695.

[12] A. Gangemi, V. Presutti, D.R. Recupero, A.G. Nuzzolese, F. Draicchio, M. Mongiovì, Semantic web machine reading with FRED, Semant. Web 8 (6) (2017) 873–893, http://dx.doi.org/10.3233/SW-160240.

[13] House of Lords Select Committee, et al., Ai in the uk: ready, willing and able, House Lords 36 (2018).

[14] ICO, Big data, artificial intelligence, machine learning and data protection, 2019, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf Last Accessed=4-11-2019.

[15] J. Olsson, Guide to uniform production of judgments, 2019, https://aija.org.au/wp-content/uploads/2017/10/Guide-to-Uniform-Production-of-Judgments-2nd-Ed-Olsson-1999.pdf Last Accessed=05-10-2019.

[16] S. Brüninghaus, K.D. Ashley, Improving the representation of legal case texts with information extraction methods, in: Proceedings of the 8th International Conference on Artificial Intelligence and Law, in: ICAIL '01, Association for Computing Machinery, New York, NY, USA, 2001, pp. 42–51, http://dx.doi.org/10.1145/383535.383540.
[17] A. Mandal, K. Ghosh, A. Pal, S. Ghosh, Automatic catchphrase identification from legal court case documents, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, in: CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 2187–2190, http://dx.doi.org/10.1145/3132847.3133102.
[18] Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019, ACM, 2019, http://dx.doi.org/10.1145/3322640.
[19] P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya, P. Majumder, FIRE 2019 AILA track: Artificial intelligence for legal assistance, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, in: FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 4–6, http://dx.doi.org/10.1145/3368567.3368587.
[20] C.M. de Oliveira Rodrigues, F.L.G. calves De Freitas, E.F.S. Barreiros, R.R. de Azevedo, A.T. de Almeida Filho, Legal ontologies over time: A systematic mapping study, Expert Syst. Appl. 130 (2019) 12–30, http://dx.doi.org/10.1016/j.eswa.2019.04.009.
[21] D. Wills, From the law librarian to legal information management, from bulletin to journal: A jubilee year, Leg. Inf. Manag. 20 (1) (2020) 4–16, http://dx.doi.org/10.1017/S1472669620000031.
[22] L. Zhang, X. Zhao, An overview of cross-language information retrieval, in: X. Sun, J. Wang, E. Bertino (Eds.), Artificial Intelligence and Security, Springer International Publishing, Cham, 2020, pp. 26–37.
[23] J.R. Baron, Law in the age of exabytes: Some further thoughts on 'information inflation'and current issues in e-discovery search, Richmond J. Law Technol. 17 (3) (2011) 9.
[24] D.W. Oard, J.R. Baron, B. Hedin, D.D. Lewis, S. Tomlinson, Evaluation of information retrieval for E-discovery, Artif. Intell. Law 18 (4) (2010) 347–386, http://dx.doi.org/10.1007/s10506-010-9093-9.
[25] M. Van Opijnen, C. Santos, On the concept of relevance in legal information retrieval, Artif. Intell. Law 25 (1) (2017) 65–87, http://dx.doi.org/10.1007/s10506-017-9195-8.
[26] G.J. Leckie, K.E. Pettigrew, C. Sylvain, Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers, Libr. Q.: Inf. Community Policy 66 (2) (1996) 161–193.
[27] B. Mitra, N. Craswell, An introduction to neural information retrieval, Found. Trends® Inf. Retr. 13 (1) (2018) 1–126, http://dx.doi.org/10.1561/1500000061.
[28] H. Bonab, S.M. Sarwar, J. Allan, Training effective neural CLIR by bridging the translation gap, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2020, pp. 9–18, http://dx.doi.org/10.1145/3397271.3401035.
[29] B. Li, P. Cheng, Learning neural representation for CLIR with adversarial framework, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1861–1870, http://dx.doi.org/10.18653/v1/D18-1212.
[30] P. Gupta, R.E. Banchs, P. Rosso, Continuous space models for CLIR, Inf. Process. Manage. 53 (2) (2017) 359–370, http://dx.doi.org/10.1016/j.ipm.2016.11.002.
[31] V. Zhebel, D. Zubarev, I. Sochenkov, Different approaches in cross-language similar documents retrieval in the legal domain, in: International Conference on Speech and Computer, Springer, 2020, pp. 679–686, http://dx.doi.org/10.1007/978-3-030-60276-5_65.
[32] J.F. Horty, Agency and Deontic Logic, Oxford University Press, 2001.
[33] W.N. Hohfeld, Some fundamental legal conceptions as applied in judicial reasoning, Yale Lj 23 (1913) 16.
[34] P. Giorgini, F. Massacci, J. Mylopoulos, N. Zannone, Modeling security requirements through ownership, permission and delegation, in: 13th IEEE International Conference on Requirements Engineering, RE'05, 2005, pp. 167–176, http://dx.doi.org/10.1109/RE.2005.43.
[35] T.D. Breaux, M.W. Vail, A.I. Anton, Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations, in: 14th IEEE International Requirements Engineering Conference, RE'06, 2006, pp. 49–58, http://dx.doi.org/10.1109/RE.2006.68.
[36] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does NLP benefit legal system: A summary of legal artificial intelligence, 2020, arXiv preprint arXiv:2004.12158.
[37] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, C. Soria, Automatic semantics extraction in law documents, in: Proceedings of the 10th International Conference on Artificial Intelligence and Law, in: ICAIL '05, Association for Computing Machinery, New York, NY, USA, 2005, pp. 133–140, http://dx.doi.org/10.1145/1165485.1165506.
[38] M.J. Bommarito II, D.M. Katz, E.M. Detterman, Lexnlp: Natural language processing and information extraction for legal and regulatory texts, in: Research Handbook on Big Data Law, Edward Elgar Publishing, 2021.
[39] C. Soria, R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, Automatic extraction of semantics in law documents, in: Proceedings of the V Legislative XML Workshop, European Press Academic Publishing, 2007, pp. 253–266.
[40] 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, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 266–267, http://dx.doi.org/10.1145/3322640.3326737.
[41] T.M. van Engers, K. Sayah, A case study on automated norm extraction, Leg. Knowl. Inf. Syst. Jurix (2004) 49–58.
[42] E. Maat, R. Winkels, Suggesting model fragments for sentences in dutch laws, in: Proceedings of Legal Ontologies and Artificial Intelligence Techniques, 2010, pp. 19–28.
[43] R. Brighi, M. Palmirani, Legal text analysis of the modification provisions: A pattern oriented approach, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, in: ICAIL '09, ACM, New York, NY, USA, 2009, pp. 238–239, http://dx.doi.org/10.1145/1568234.1568272.
[44] N. Kiyavitskaya, N. Zeni, T.D. Breaux, A.I. Antón, J.R. Cordy, L. Mich, J. Mylopoulos, Automating the extraction of rights and obligations for regulatory compliance, in: International Conference on Conceptual Modeling, Springer, 2008, pp. 154–168, http://dx.doi.org/10.1007/978-3-540-87877-3_13.
[45] A.Z. Wyner, W. Peters, On rule extraction from regulations, in: JURIX, Vol. 11, 2011, pp. 113–122.
[46] G. Boella, L. Di Caro, L. Robaldo, Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines, in: International Workshop on Rules and Rule Markup Languages for the Semantic Web, Springer, 2013, pp. 218–225, http://dx.doi.org/10.1007/978-3-642-39617-5_20.
[47] A. Sleimi, N. Sannier, M. Sabetzadeh, L. Briand, M. Ceci, J. Dann, An automated framework for the extraction of semantic legal metadata from legal texts, Empir. Softw. Eng. 26 (3) (2021) 1–50, http://dx.doi.org/10.1007/s10664-020-09933-5.
[48] M. Dragoni, S. Villata, W. Rizzi, G. Governatori, Combining NLP approaches for rule extraction from legal documents, in: 1st Workshop on MIning and REasoning with Legal Texts (MIREL 2016), 2016, pp. 1–13.
[49] J. Pehcevski, A.-M. Vercoustre, J.A. Thom, Exploiting locality of wikipedia links in entity ranking, in: European Conference on Information Retrieval, Springer, 2008, pp. 258–269, http://dx.doi.org/10.1007/978-3-540-78646-7_25.
[50] R. Kaptein, P. Serdyukov, A. De Vries, J. Kamps, Entity ranking using wikipedia as a pivot, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, in: CIKM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 69–78, http://dx.doi.org/10.1145/1871437.1871451.
[51] M. Schuhmacher, L. Dietz, S. Paolo Ponzetto, Ranking entities for web queries through text and knowledge, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, in: CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1461–1470, http://dx.doi.org/10.1145/2806416.2806480.
[52] R. Prokofyev, A. Tonon, M. Luggen, L. Vouilloz, D.E. Difallah, P. Cudré-Mauroux, SANAPHOR: Ontology-based coreference resolution, in: International Semantic Web Conference, Springer, 2015, pp. 458–473, http://dx.doi.org/10.1007/978-3-319-25007-6_27.
[53] E. Filtz, Building and processing a knowledge-graph for legal data, in: E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, O. Hartig (Eds.), The Semantic Web, Springer International Publishing, Cham, 2017, pp. 184–194, http://dx.doi.org/10.1007/978-3-319-58451-5_13.
[54] J. Moreno-Schneider, G. Rehm, E. Montiel-Ponsoda, V. Rodriguez-Doncel, A. Revenko, S. Karampatakis, M. Khvalchik, C. Sageder, J. Gracia, F. Maganza, Orchestrating NLP services for the legal domain, 2020, arXiv preprint arXiv:2003.12900.
[55] M. Gifford, Lexridelaw: An argument based legal search engine, in: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, in: ICAIL '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 271–272, http://dx.doi.org/10.1145/3086512.3086548.
[56] R. Studer, V. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data Knowl. Eng. 25 (1) (1998) 161–197, http://dx.doi.org/10.1016/S0169-023X(97)00056-6.
[57] E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia, Integrating a bottom–up and top–down methodology for building semantic resources for the multilingual legal domain, in: Semantic Processing of Legal Texts, Springer, 2010, pp. 95–121.

[58] S. Despres, S. Szulman, Terminae method and integration process for legal ontology building, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2006, pp. 1014–1023, http://dx.doi.org/10.1007/11779568_108.
[59] N. Zhang, Y. Pu, S. Yang, J. Zhou, J. Gao, An ontological Chinese legal consultation system, IEEE Access 5 (2017) 18250–18261, http://dx.doi.org/10.1109/ACCESS.2017.2745208.
[60] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, Knowl. Eng. Rev. 11 (2) (1996) 93–136, http://dx.doi.org/10.1017/S0269888900007797.
[61] V. Leone, L. Di Caro, S. Villata, Taking stock of legal ontologies: A feature-based comparative analysis, Artif. Intell. Law 28 (2) (2020) 207–235, http://dx.doi.org/10.1007/s10506-019-09252-1.
[62] J. Hage, B. Verheij, The law as a dynamic interconnected system of states of affairs: A legal top ontology, Int. J. Hum.-Comput. Stud. 51 (6) (1999) 1043–1077, http://dx.doi.org/10.1006/ijhc.1999.0297.
[63] J. Lehmann, J. Breuker, B. Brouwer, Causation in AI and law, Artif. Intell. Law 12 (4) (2004) 279–315, http://dx.doi.org/10.1007/s10506-005-4157-y.
[64] F. Freitas, Z. Candeias Jr., H. Stuckenschmidt, Towards checking laws' consistency through ontology design: the case of Brazilian vehicles' laws, J. Theor. Appl. Electron. Commerce Res. 6 (1) (2011) 112–126.
[65] C.M. de Oliveira Rodrigues, F.L.G. De Freitas, R.R. De Azevedo, An ontology for property crime based on events from ufo-b foundational ontology, in: 2016 5th Brazilian Conference on Intelligent Systems, BRACIS, IEEE, 2016, pp. 331–336.
[66] C.M. de Oliveira Rodrigues, E. Palmeira, F. Freitas, I. Oliveira, I. Varzinczak, LEGIS: A proposal to handle legal normative exceptions and leverage inference proofs readability, J. Appl. Logics 2631 (5) (2019) 755.
[67] M. El Ghosh, H. Naja, H. Abdulrab, M. Khalil, Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules, Procedia Comput. Sci. 112 (2017) 632–642.
[68] R.W. Van Kralinger, P.R. Visser, T.J. Bench-Capon, H. Jaap Van Den Herik, A principled approach to developing legal knowledge systems, Int. J. Hum.-Comput. Stud. 51 (6) (1999) 1127–1154, http://dx.doi.org/10.1006/ijhc.1999.0300.
[69] G. Guizzardi, G. Wagner, J.a.P.A. Almeida, R.S. Guizzardi, Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story, Appl. Ontology 10 (3–4) (2015) 259–271.
[70] J. Shaheed, A. Yip, J. Cunningham, A top-level language-biased legal ontology, in: ICAIL Workshop on Legal Ontologies and Artificial Intelligence Techniques, LOAIT, Citeseer, 2005, pp. 13–24.
[71] T.D. Breaux, C. Powers, Early studies in acquiring evidentiary, reusable business process models for legal compliance, in: 2009 Sixth International Conference on Information Technology: New Generations, 2009, pp. 272–277, http://dx.doi.org/10.1109/ITNG.2009.72.
[72] M. El Ghosh, H. Abdulrab, H. Naja, M. Khalil, A criminal domain ontology for modelling legal norms, in: Conference of the Italian Association for Artificial Intelligence, Springer, 2017, pp. 282–294, http://dx.doi.org/10.1007/978-3-319-70169-1_21.
[73] 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, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 179–183, http://dx.doi.org/10.1145/3322640.3326730.
[74] S. Castano, M. Falduti, A. Ferrara, S. Montanelli, A knowledge-centered framework for exploration and retrieval of legal documents, Inf. Syst. (2021) 101842, http://dx.doi.org/10.1016/j.is.2021.101842.
[75] C. Griffo, J.a.P.A. Almeida, G. Guizzardi, J.C. Nardi, From an ontology of service contracts to contract modeling in enterprise architecture, in: 2017 IEEE 21st International Enterprise Distributed Object Computing Conference, EDOC, 2017, pp. 40–49, http://dx.doi.org/10.1109/EDOC.2017.15.
[76] A. Gangemi, M.-T. Sagri, D. Tiscornia, Metadata for content description in legal information, in: Procs. of LegOnt Workshop on Legal Ontologies, 2003.
[77] G. Ajani, G. Boella, L. Di Caro, L. Robaldo, L. Humphreys, S. Praduroux, P. Rossi, A. Violato, The european legal taxonomy syllabus: a multi-lingual, multi-level ontology framework to untangle the web of european legal terminology, Appl. Ontology 11 (4) (2016) 325–375.
[78] J. McClure, The legal-RDF ontology. A generic model for legal documents, in: LOAIT, 2007, pp. 25–42.
[79] M.G. Buey, A.L. Garrido, C. Bobed, S. Ilarri, The AIS project: Boosting information extraction from legal documents by using ontologies, in: ICAART, Vol. 2, 2016, pp. 438–445.
[80] T.D. Bui, S.T. Nguyen, Q.B. Ho, Towards a conceptual search for Vietnamese legal text, in: IFIP International Conference on Computer Information Systems and Industrial Management, Springer, 2015, pp. 175–185, http://dx.doi.org/10.1007/978-3-662-45237-0_18.
[81] T. Tantisripreecha, N. Soonthornphisaj, Supreme court sentences retrieval using thai law ontology, in: Intelligent Control and Computer Engineering, Springer, 2011, pp. 177–189, http://dx.doi.org/10.1007/978-94-007-0286-8_15.
[82] G. Boella, L. Di Caro, L. Humphreys, L. Robaldo, P. Rossi, L. van der Torre, Eunomos, A legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law, Artif. Intell. Law 24 (3) (2016) 245–283, http://dx.doi.org/10.1007/978-3-642-35731-2_9.
[83] W. Peters, M.-T. Sagri, D. Tiscornia, The structuring of legal knowledge in LOIS, Artif. Intell. Law 15 (2) (2007) 117–135, http://dx.doi.org/10.1007/s10506-007-9034-4.
[84] S.A. Board, Law, Governance and Technology Series, Springer, 2011.
[85] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer, The LKIF core ontology of basic legal concepts, Proc. LOAIT 07 (2007) 43.
[86] V. Kasper, Developing content for LKIF: Ontologies and frameworks for legal reasoning, in: Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference, Vol. 152, IOS Press, 2006, p. 169.
[87] A. Boer, R. Winkels, F. Vitali, Proposed XML standards for law: MetaLex and LKIF, in: Proceedings of the 2007 Conference on Legal Knowledge and Information Systems: JURIX 2007: The Twentieth Annual Conference, IOS Press, 2007, pp. 19–28.
[88] T. Athan, H. Boley, G. Governatori, M. Palmirani, A. Paschke, A.Z. Wyner, OASIS LegalRuleML, in: ICAIL, Vol. 13, 2013, pp. 3–12.
[89] H.-P. Lam, M. Hashmi, Enabling reasoning with LegalRuleML, Theory Pract. Log. Program. 19 (1) (2019) 1–26.
[90] K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence, Behav. Res. Methods Instrum. Comput. 28 (2) (1996) 203–208.
[91] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
[92] S.K. Sienčnik, Adapting word2vec to named entity recognition, in: Proceedings of the 20th Nordic Conference of Computational Linguistics, Nodalida 2015, May 11-13, 2015, Vilnius, Lithuania, (109) Linköping University Electronic Press, 2015, pp. 239–243.
[93] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, N. Bhamidipati, Context- and content-aware embeddings for query rewriting in sponsored search, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 383–392, http://dx.doi.org/10.1145/2766462.2767709.
[94] E. Nalisnick, B. Mitra, N. Craswell, R. Caruana, Improving document ranking with dual word embeddings, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 83–84.
[95] I. Chalkidis, D. Kampas, Deep learning in law: early adaptation and legal word embeddings trained on large corpora, Artif. Intell. Law 27 (2) (2019) 171–198, http://dx.doi.org/10.1007/s10506-018-9238-9.
[96] A. Mandal, K. Ghosh, S. Ghosh, S. Mandal, Unsupervised approaches for measuring textual similarity between legal court case reports, Artif. Intell. Law (2021) 1–35, http://dx.doi.org/10.1007/s10506-020-09280-2.
[97] J.T. Lam, D. Liang, S. Dahan, F.H. Zulkernine, The gap between deep learning and law: Predicting employment notice, in: NLLP@ KDD, 2020, pp. 52–56.
[98] N. Bansal, A. Sharma, R. Singh, A review on the application of deep learning in legal domain, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2019, pp. 374–381, http://dx.doi.org/10.1007/978-3-030-19823-7_31.
[99] H.-T. Nguyen, H.-Y.T. Vuong, P.M. Nguyen, B.T. Dang, Q.M. Bui, S.T. Vu, C.M. Nguyen, V. Tran, K. Satoh, M.L. Nguyen, JNLP team: Deep learning for legal processing in COLIEE 2020, 2020, arXiv preprint arXiv:2011.08071.
[100] I. Mokanov, D. Shane, B. Cerat, Facts2law: Using deep learning to provide a legal qualification to a set of facts, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 268–269, http://dx.doi.org/10.1145/3322640.3326694.
[101] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, S. Ma, BERT-PLI: Modeling paragraph-level interactions for legal case retrieval, in: IJCAI, 2020, pp. 3501–3507.
[102] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, 2020, arXiv preprint arXiv:2010.02559.
[103] K. Sugathadasa, B. Ayesha, N. de Silva, A.S. Perera, V. Jayawardana, D. Lakmal, M. Perera, Legal document retrieval using document vector embeddings and deep learning, in: Science and Information Conference, Springer, 2018, pp. 160–175, http://dx.doi.org/10.1007/978-3-030-01177-2_12.
[104] J. Landthaler, B. Waltl, P. Holl, F. Matthes, Extending full text search for legal document collections using word embeddings, in: JURIX, 2016, pp. 73–82.

[105] V. Tran, M. Le Nguyen, S. Tojo, K. Satoh, Encoded summarization: summarizing documents into continuous vector space for legal case retrieval, Artif. Intell. Law (2020) 1–27, http://dx.doi.org/10.1007/s10506-020-09262-4.

[106] Z. Xu, T. He, H. Lian, J. Wan, H. Wang, Case facts analysis method based on deep learning, in: International Conference on Web Information Systems and Applications, Springer, 2019, pp. 92–97, http://dx.doi.org/10.1007/978-3-030-30952-7_11.

[107] P.-K. Do, H.-T. Nguyen, C.-X. Tran, M.-T. Nguyen, M.-L. Nguyen, Legal question answering using ranking SVM and deep convolutional neural network, 2017, arXiv preprint arXiv:1703.05320.

[108] A. Elnaggar, R. Otto, F. Matthes, Deep learning for named-entity linking with transfer learning for legal documents, in: Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference, in: AICCC '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 23–28, http://dx.doi.org/10.1145/3299819.3299846.

[109] F. Wei, H. Qin, S. Ye, H. Zhao, Empirical study of deep learning for text classification in legal document review, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3317–3320, http://dx.doi.org/10.1109/BigData.2018.8622157.

[110] S. Undavia, A. Meyers, J.E. Ortega, A comparative study of classifying legal documents with neural networks, in: 2018 Federated Conference on Computer Science and Information Systems, FedCSIS, IEEE, 2018, pp. 515–522.

[111] E. Hammami, I. Akermi, R. Faiz, M. Boughanem, Deep learning for french legal data categorization, in: International Conference on Model and Data Engineering, Springer, 2019, pp. 96–105, http://dx.doi.org/10.1007/978-3-030-32065-2_7.

[112] D. Song, A. Vold, K. Madan, F. Schilder, Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training, Inf. Syst. (2021) 101718, http://dx.doi.org/10.1016/j.is.2021.101718.

[113] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[114] N.C. Da Silva, F. Braz, et al., Document type classification for Brazil's supreme court using a convolutional neural network, in: The Tenth International Conference on Forensic ComputerScience and Cyber Law-ICoFCS, 2018, pp. 7–11.

[115] V. Tran, M.L. Nguyen, K. Satoh, Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 275–282, http://dx.doi.org/10.1145/3322640.3326740.

[116] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, S. Ma, BERT-PLI: Modeling paragraph-level interactions for legal case retrieval, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 3501–3507, http://dx.doi.org/10.24963/ijcai.2020/484, Main track.

[117] J. Rabelo, M.-Y. Kim, R. Goebel, Combining similarity and transformer methods for case law entailment, in: Proceedings of the Seventeenth International Conference on on Artificial Intelligence and Law, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 290–296, http://dx.doi.org/10.1145/3322640.3326741.

[118] M.-Y. Kim, J. Rabelo, R. Goebel, Statute law information retrieval and entailment, in: Proceedings of the Seventeenth International Conference on on Artificial Intelligence and Law, in: ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 283–289, http://dx.doi.org/10.1145/3322640.3326742.

[119] R. More, J. Patil, A. Palaskar, A. Pawde, Removing named entities to find precedent legal cases, in: FIRE (Working Notes), 2019, pp. 13–18.

[120] M. Lefoane, T. Koboyatshwene, G. Rammidi, V.L. Narasimham, Legal statutes retrieval: A comparative approach on performance of title and statutes descriptive text, in: FIRE (Working Notes), 2019, pp. 52–57.

[121] S. Mandal, S.D. Das, Unsupervised identification of relevant cases & statutes using word embeddings, in: FIRE (Working Notes), 2019, pp. 31–35.