



Enforcing legal information extraction through context-aware techniques: The ASKE approach

Silvana Castano ^a, Alfio Ferrara ^a, Emanuela Furiosi ^b, Stefano Montanelli ^a, Sergio Picascia ^{a,*},
Davide Riva ^a, Carolina Stefanetti ^c

^a Università degli Studi di Milano, Department of Computer Science, Via Celoria, 18 20133 Milano, Italy

^b IUSS Istituto Universitario di Studi Superiori di Pavia, Piazza della Vittoria, 15 27100 Pavia, Italy

^c Università degli Studi di Milano, Department of Italian and Supranational Public Law, via Festa del Perdono, 7 20122 Milano, Italy

ARTICLE INFO

Keywords:

Legal knowledge extraction
Natural Language Processing
Legal knowledge graph
Digital justice

ABSTRACT

To cope with the growing volume, complexity, and articulation of legal documents as well as to foster digital justice and digital law, increasing effort is being devoted to legal knowledge extraction and digital transformation processes. In this paper, we present the ASKE (*Automated System for Knowledge Extraction*) approach to legal knowledge extraction, based on a combination of context-aware embedding models and zero-shot learning techniques into a three-phase extraction cycle, which is executed a number of times (called generations) to progressively extract concepts representative of the different meanings of terminology used in legal documents chunks. A graph-based data structure called *ASKE Conceptual Graph* is initially populated through a data preparation step, and it is continuously enriched at each ASKE generation with results of document chunk classification, new extracted terminology, and newly derived concepts. A quantitative evaluation of ASKE knowledge extraction and document classification is provided by considering the EurLex dataset. Furthermore, we present the results of applying ASKE to a real case-study of Italian case law decisions with qualitative feedback from legal experts in the framework of an ongoing national research project.

1. Introduction

Law plays a crucial role in almost every aspect of our life, both public and private. Thousands of legal documents are constantly produced by institutional bodies, such as Parliaments and Courts, and constitute a prominent source of information and knowledge primarily for judges, lawyers, and other legal practitioners involved in legal decision-making, but also for general subjects like citizens or private and public organizations. Knowing how to navigate such a complex context both in structure and content is a primary need for several categories of users like *legal practitioners*, to support their professional activities, *administrators*, to enforce law procedures, and *generic users/citizens*, to enforce effective legal information exploration and exploitation [1]. The availability of a technology for extracting knowledge from legal documents is not only desirable but even necessary, and the benefits and concrete outcomes that could result from the diffusion of such technology are many and different for both legal practitioners (i.e., lawyers, judges and Courts), administrations, and final users. Legal search through legal knowledge extraction is an extremely important instrument for legal practitioners in both common law [2] and civil law [3] systems. Indeed, one of the milestone principles of law is the

certainty of law. For example, legal search over precedent case law may be useful for a lawyer to retrieve a decision rendered in a case similar to the case at hand, where the Court decided in a way that is favorable to its client position, or a decision rendered in a different case on the basis of a reasoning that, applied to the case at hand, leads to a favorable interpretation of its client position [4]. When conducting case law research, it is important to focus on both the decision of the case, but also the motivation and the reasoning (called “rationale”) behind the decision. During this process, great help may come from knowledge extraction systems, especially if they are “context-aware”. In the context of administrative decisions, knowledge extraction from legal documents could help public administrations to identify the legislation applicable to the specific case, being able to ensure in-depth and always up-to-date knowledge of any relevant legislation, including the most specific one. Such technology could also be used to automate, at least in part, some administrative processes, considering that in most European countries the principle of “digital only” has been diffused, thus enabling the use of legal search technologies as well [5–8]. From the point of view of generic users/citizens, the development of knowledge extraction tools could foster transparency, accessibility, and fairness within the legal

* Corresponding author.

E-mail address: sergio.picascia@unimi.it (S. Picascia).

system, by equipping citizens with valuable insights and resources. By granting easier access to legal documents such as laws, judicial decisions, and administrative proceedings, it would empower citizens with a better understanding of their rights and available opportunities. To support case analysis and legal decision-making in an effective way, advanced knowledge extraction solutions are thus demanded, based on Natural Language Processing (NLP), Machine Learning (ML), and Artificial Intelligence (AI), to deal with challenging requirements posed by the legal documentation, such as the language complexity, the significant length of the legal texts, the poor accessibility of legal datasets that makes large-scale downloads challenging or even impossible, as well as the lack of sufficiently-large annotated corpora for model training. Context-aware embedding techniques are being adopted to address language complexity [9]: the whole chunk of text in which a word appears is considered for the embedding construction, to better deal with word-sense disambiguation and high domain-specificity of the legal terminology.

In this paper, we present ASKE (*Automated System for Knowledge Extraction*), an approach to legal knowledge extraction with focus on abstract concept discovery a combination of context-aware embedding models and zero-shot learning techniques. ASKE takes a corpus of legal documents as input and it extracts a graph of concepts which are used to classify the given documents at a chunk-level (e.g., paragraph) granularity. Through context-aware embedding, document chunks and concept definitions are projected in the same semantic space, to appropriately capture and manage the meaning of legal terminology by taking into account the context in which terms are used. Through zero-shot learning, a multi-label classification process is performed in an unsupervised way, without relying on any pre-existing annotation of legal documents. The distinguishing feature of ASKE is the implementation of a cyclic extraction process by which, at each cycle: (i) embedded documents chunks are classified against the current concepts (*document chunk classification*); (ii) new terminology recognized to be relevant for/similar to a given concept is extracted and assigned to the concept (*terminology enrichment*); (iii) new concepts are possibly derived based on the results of similarity-based term clustering (*concept derivation*). A graph-based data structure called ACG (*ASKE Conceptual Graph*) is initially populated through a data preparation step, and it is continuously enriched at each ASKE generation with results of document chunk classification, new extracted terminology, and newly derived concepts. To evaluate the extraction process, we run ASKE on a split of the EurLex dataset, containing 45,000 EU legislative documents from the EUR-LEX portal annotated with concepts from the EuroVoc taxonomy, and we use *BERTopic* and *Zero-Shot TM* as baselines for the comparison. Then, we discuss the application of ASKE to a real knowledge extraction case study based on a corpus of 50 Italian case law decisions, in the framework of the *Next Generation UPP (NGUPP)* project, funded by the Italian Ministry of Justice, aiming at providing artificial intelligence and advanced information management techniques for the digital transformation of Italian legal processes and digital justice in general.

The paper is organized as follows. In Section 2, we present the ASKE approach to legal knowledge extraction, with focus on the ACG knowledge model and on classification/extraction techniques featuring each ASKE cycle. In Section 3, we describe the ASKE evaluation to assess the quality of the extraction and classification results against the EurLex legal dataset. Section 4 is devoted to the presentation of a real case study of using ASKE on a corpus of Italian case law decisions in the framework of the *NGUPP* project. Section 5 discusses the related work. Finally, in Section 6, we provide our concluding remarks.

2. The ASKE approach to legal knowledge extraction

The two main characteristics of the ASKE approach to legal knowledge extraction regard the exploitation of zero-shot learning techniques

and the adoption of context-aware embedding models. Zero-shot learning techniques allow to deal with situations in which labeled datasets are not available, which is the case of the legal field. Legal corpora are usually not annotated and, when some annotation is available, it is not suitable for fine-grained, sentence-level information retrieval tasks, as expected. Common benchmarks of legal information retrieval, such as the ones proposed for the TREC Legal Track [10] or the FIRE AILA Track [11], require an entire document to be used as input in order to retrieve precedents or similar cases. This is not the situation in which legal actors would like to work, in that they prefer to retrieve specific portions/sentences of a document, rather than going through an entire document in order to find the section of interest.

Dealing with domain-specific text-processing applications, our choice is for context-aware embedding models rather than non-contextual ones (e.g., Word2Vec). The motivation is that the language employed in the legal field is usually technical. Not only is the sentence structure peculiar, but some terms may also assume completely different meaning when used in different contexts/sentences of court decisions. Being able to capture contextual information, such as the part of a document where a certain idea is expressed, can be crucial when it comes to distinguishing if a certain matter is treated in the introduction or in the conclusion of a case law decision. Furthermore, we choose neural embedding models over non-neural retrieval models (e.g., BM25) for their ability to map concepts and documents in the same space.

After a data preparation step, knowledge extraction consists of the execution of a cycle composed of three phases (see Fig. 1): (i) document chunk classification, (ii) terminology enrichment, (iii) concept derivation.

The *data preparation* step performs the embedding of concepts and document chunks into the same vector space using a context-aware embedding model. In the *document chunk classification* phase, document chunks are assigned to concepts by exploiting the similarity σ between their respective vector representations. In the *terminology enrichment* phase, ASKE assigns to a concept c the most similar new terms occurring in document chunks classified under c in the previous phase. In the *concept derivation* phase, all the terms associated with each concept c are submitted to a clustering algorithm to (possibly) derive new concept(s) representative of a cluster of highly similar terms. Results of the knowledge extraction process populate the so-called *ASKE Conceptual Graph (ACG)*, organized according to the knowledge model described in the following.

2.1. The ASKE knowledge model

The ASKE knowledge model is organized as a graph-based data structure called ASKE Conceptual Graph (ACG) whose conceptual schema is shown in Fig. 2. ACG is based on three main entities, namely *document chunk*, *term*, and *concept*.

- A **document chunk** corresponds to a portion (e.g., sentence) of the original documents extracted through the application of tokenization techniques. A document chunk k has the form $k = (k_d, \bar{k})$, where k_d is the textual content of the chunk and \bar{k} is its vector representation.
- A **term** corresponds to an n-gram occurring in a document chunk. A term w has form $w = (w_l, w_d, \bar{w})$, where w_l is the label of the term, w_d is a description of the term meaning, and \bar{w} is its vector representation.
- A **concept** represents the meaning of a set (i.e., a cluster) of terms. A concept c has the form $c = (c_l, \bar{c})$, where c_l is the label assigned to the concept and \bar{c} is its vector representation. The label c_l abstracts the meaning of the concept, expressing it in a synthetic and human-understandable way. The label is selected among the terms in the c cluster.¹ A concept is active by

¹ In ASKE, the concept label corresponds to the term whose vector is closest to the mean of the vectors of all the terms associated with the concept.

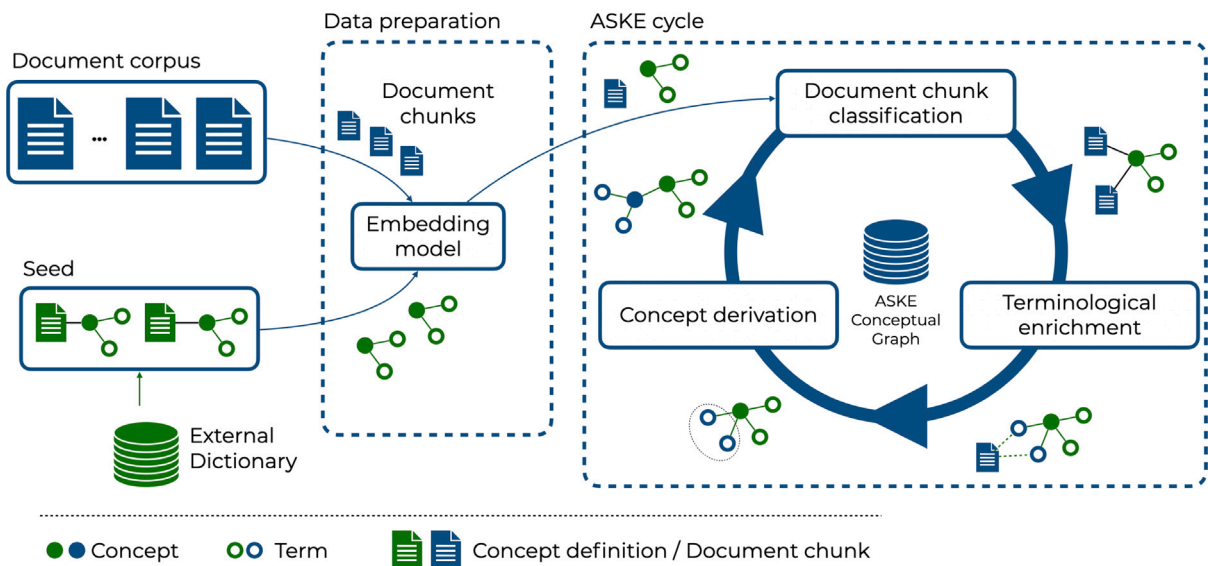


Fig. 1. The ASKE knowledge extraction approach.

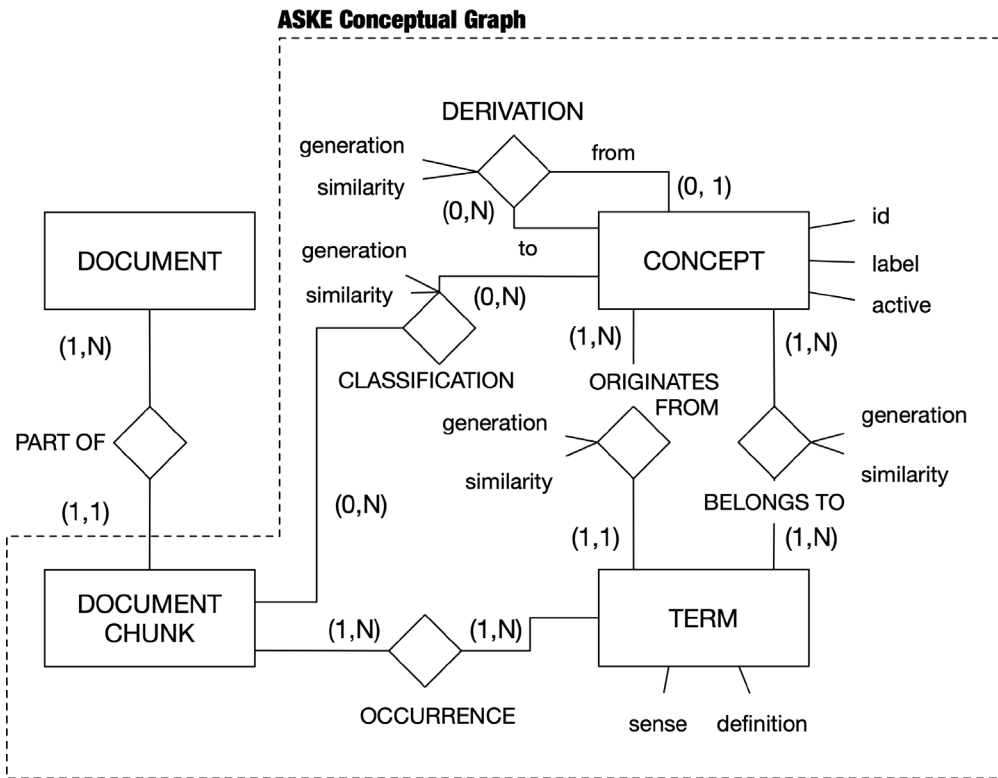


Fig. 2. The ASKE knowledge model.

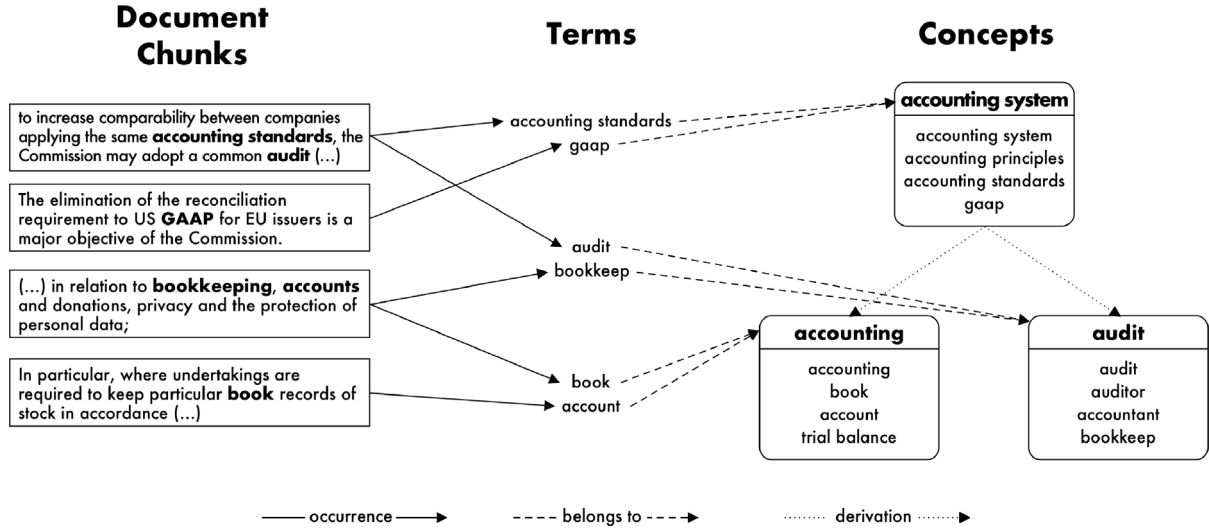


Fig. 3. Example of ASKE knowledge model instance.

default at its formation, meaning that it is used for classification and enrichment purposes in subsequent generations unless it is deactivated (i.e., ‘active’ attribute to *False*).²

The following relationships are defined in ACG, to capture the semantic relationships that hold between entities due to the ASKE extraction process.

- The **classification** relationship between document chunks and concepts represents the fact that a document chunk k is assigned to a concept c during the document chunk classification phase of the ASKE cycle. It is a “many-to-many” relationship in that a multi-label classification process is enforced, whereby a document chunk k can be associated with more than one concept c .
- The **derivation** relationship between a pair of concepts represents the fact that a concept c' is derived from a concept c during the concept derivation phase of the ASKE cycle. The ID of the ASKE generation at which the c' has been derived from c is maintained as well as the similarity value between them.
- The **originates from** relationship between terms and concepts represents the fact that a term w enters the first time the ACG being extracted from a document chunk classified under c .
- The **belongs to** relationship between terms and concepts represents the fact that a term w belongs to the cluster of terms associated with a concept c .
- The **occurrence** relationship between terms and document chunks represents the fact that a term w occurs in a document chunk k .

Example 2.1. According to the ASKE knowledge model, in Fig. 3, we show an example of document chunks, terms, and concepts extracted from the EurLex dataset. Starting from the left, the document chunks denote portions of text extracted from the legislative documents belonging to the dataset. In the center of the figure, we show terms that occur in these document chunks (highlighted in bold in the chunks) and that are also candidates for enriching the terminology of the concepts in the ACG. Finally, on the right, we show concepts and related terms. The derivation relationship is used to denote similarity relationships between concepts that are created in different ASKE cycles.

² A deactivated concept is no longer used for document chunk classification or knowledge enrichment purposes. Concept deactivation is explained in more detail in Section 2.2.

2.2. Knowledge extraction with ASKE

In this section, we describe the data preparation step and the techniques employed in each phase of the ASKE cycle for knowledge extraction.

2.2.1. Data preparation

The two main inputs of ASKE are a corpus $\mathcal{D} = d_1, \dots, d_n$ of legal documents and a *seed* S provided by the user to trigger the extraction process the first time. Data preparation consists in building the embeddings of both the corpus and the seed. Each document in the corpus is preprocessed using a tokenizer that splits the text in a set of document chunks \mathcal{K} , providing the lemmatized version of the terms \mathcal{W} therein contained. An issue to be addressed for data preparation is related to the definition of the logical document unit that should correspond to a chunk (e.g., a single sentence or paragraph). A document chunk should thus correspond to a logical unit within the document whose length can provide enough information content for its practical usage in application contexts. On the other side, the dimension of document chunks should fit the constraints of the adopted embedding model which generally works with a fixed-size window (e.g., 512 tokens). Considering the writing style featuring legal documents, such as case law documents, where sentences are generally rich and articulated for a suitable argumentation, a single sentence is a good candidate to become a document chunk. The seed provides the description of an initial target concept to be used to trigger the first cycle of the ASKE knowledge extraction process. To comply with different legal users and skills, the initial description to provide as seed is kept “easy to formulate” according to the following options:

1. **query seed:** the user provides a textual description of the initial target concept in form of one or more textual queries (e.g., a query can be a sentence taken from law/case law documentation);
2. **keyword seed:** the user provides a textual description of the initial target concept in form of one or more keywords.

In the first case, the input query directly provides a textual definition of the target concept, to be used for embedding the seed. In the second case, the textual definition of the target concept is reconstructed by ASKE exploiting the keyword(s) with the help of an external dictionary. Current version of ASKE relies on WordNet in order to be able to work with the majority of online legal datasets. In particular, for each keyword term w appearing in the seed, ASKE retrieves from WordNet

the definition associated with each sense of w ; all retrieved definitions are then exploited for seed embedding.

The last step of data preprocessing consists in transforming document chunks and the seed concept in their corresponding vector representation, projecting them in the same semantic space. This is achieved using a pre-trained version of Sentence-BERT [12], a modification of the original BERT model, which exploits siamese and triplets networks, being able to derive semantically meaningful sentence embeddings. Since the initial seed can be associated with multiple definitions, we define its position in the embedding space as the centroid \bar{c}_i of $\bar{w}_{i1}, \dots, \bar{w}_{ij}$. As a result of the data preparation step, document chunks and the set of terms in the initial seed will populate the first version of the ACG.

2.2.2. Document chunk classification

In this phase, ASKE performs the zero-shot multi-label classification of document chunks. Document chunks are assigned to zero, one, or multiple concepts, $f : \mathcal{X} \rightarrow \mathcal{C}$, without having the model exposed to training examples. This is possible due to the coexistence of the embeddings of concepts and document chunks in the same vector space.

A similarity measure σ , i.e., cosine similarity, is computed between the embedding vector \bar{k} of each document chunk and the embedding vector \bar{c} of each concept respectively. An association relationship is eventually defined between them if their similarity is higher than or equal to a predefined threshold α :

$$f(k, C) = \{c_i \in C : \sigma(\bar{k}, \bar{c}_i) \geq \alpha\} \quad (1)$$

The value of the threshold α is crucial since it may remarkably affect the classification output. Choosing a high value of α will result in a high value of precision of the classification on one side, while, on the other side, in a small set of document chunks for each concept. On the other hand, picking a low value of α will result in a higher recall but also lots of false positives. Appropriate ranges and threshold values have been determined on an experimental basis (see Section 3).

In the first execution of the ASKE cycle, the initial version of the ACG is considered and document chunks are classified against the initial seed concept. In subsequent executions, the number of concepts increases due to the terminology enrichment and concept derivation phases, and document chunks can be assigned to more than one concept, depending on the computed similarity measures.

Example 2.2. Considering as document corpus the EurLex dataset, which contains documents from the Official Journal of the European Union and will be discussed in detail in Section 3, and “ACCOUNTING SYSTEM” as keyword seed, the following document chunk is classified under the “ACCOUNTING SYSTEM” concept after the document chunk classification phase:

“Notwithstanding the difficulties in quantifying the exposure to operational risk, Directive 2006/48/EC of the European Parliament and of the Council of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions is the relevant benchmark for the purpose of establishing the capital requirement for CCPs. Consistently with Directive 2006/48/EC, the definition of operational risk should include legal risk in respect of technical standards on capital requirements for central counterparties. Directive 2006/48/EC and Directive 2006/49/EC of the European Parliament and of the Council of 14 June 2006 on the capital adequacy of investment firms and credit institutions are an appropriate benchmark for the purpose of establishing capital requirements to cover credit, counterparty and market risks non covered by specific financial resources, since they are similar to those carried out by credit institutions or investment firms. A CCP does not have to hold capital for trade exposures and default fund contributions which arise under an interoperability arrangement where the requirements of Articles 52 and 53 of Regulation (EU) No 648/2012 are fulfilled. However, where these requirements are not fulfilled, links between CCPs might expose them to additional risk if the collateral posted by them is not fully protected and bankruptcy remote or if the default fund contributions are at risk in case a clearing member of the receiving CCP defaults. Therefore, in such

cases capital charges should apply to default fund contributions and to trade exposures with other CCPs. In order to avoid contagion effects, the treatment regarding default fund contributions to other CCPs should in general be more conservative than the treatment of credit institution exposures to CCPs. The own resources of a CCP used to contribute to the default fund of another CCP should not be taken into account for the purposes of Article 16(2) of Regulation (EU) No 648/2012 as they are not invested in accordance with its investment policy. They should also not be double-counted for the purpose of calculating risk weighted exposures stemming from these contributions”.

If a concept does not classify any document chunk, it is deactivated. Concept deactivation can occur as the extraction process progresses, such as when a newly derived concept results more appropriate to classify document chunks at subsequent generation, thus leaving an older concept useless, or when a newly defined concept at current generation is not effectively used for classification purposes at the subsequent cycle.

2.2.3. Terminology enrichment

For each concept c_i , ASKE retrieves the set of terms \mathcal{W}_i appearing in the document chunks \mathcal{X}_i which are classified with c_i in the ACG. Then, these terms are placed in the same embedding space of concepts and document chunks, by computing the vector representation of their definition(s) w_d retrieved from the external dictionary. The terms that are absent from such a dictionary are ignored. For this reason, the employed dictionary must be characterized by the following features: (1) being sufficiently general, and (2) being sufficiently large, to encompass as many terms from the corpus as possible, as well as to avoid a priori restrictions of the search space. In the case of polysemous terms, only the definition whose embedding is the closest to concept embedding \bar{c}_i is considered. This characteristic allows ASKE to consider each term with the most appropriate sense in that context. For example, terms that have a specific meaning in the legal field will be taken into consideration by considering their domain-specific sense, rather than a general one.

For each term, the similarity σ between its embedding vector and the vectors of the concept c_i and the vectors of its associated document chunks \mathcal{X}_i is computed. The terms whose similarity is greater than or equal to threshold β are selected and become candidates for enriching the terminology of c_i .

$$g(c_i, \mathcal{X}_i, \mathcal{W}) = \{w_h \in \mathcal{W}_i : \sigma(\bar{w}_h, \bar{c}_i) + \sigma(\bar{w}_h, \bar{\mathcal{X}}_i) \geq \beta\} \quad (2)$$

where $\bar{\mathcal{X}}_i$ is the centroid of the embedding vectors of document chunks in \mathcal{X}_i .

The set of candidate terms is then sorted in descending order according to their similarity σ . In addition, a learning rate γ can also be defined, representing the maximum number of terms that will be associated with a certain concept at each generation. Applying an upper bound γ and lower bound β ensures that, at each cycle, the process of terminology enrichment will include only a small set of candidate terms that are supposed to be meaningful with respect to the concept at hand. During our tests, we found that the most satisfactory results are obtained by setting β and γ in the range $0.2 \leq \beta \leq 0.4$ and $3 \leq \gamma \leq 7$, respectively.

Example 2.3. Considering the document chunk reported in Section 2.2.2, “bankruptcy”, “financial”, “regulation” and “EC” are examples of new terms extracted in the terminology enrichment phase. After retrieving and embedding their definitions, we compute their similarity with concept “ACCOUNTING SYSTEM” according to Eq. (2). The terms “regulation” and “EC” are discarded as they don’t satisfy the lower bound β on the similarity with the concept.

Together with “bankruptcy” and “financial”, new terms extracted in the same way from other document chunks classified with concept “ACCOUNTING SYSTEM” include “accounting standard”, “balance sheet”, “review”, “audit”, “internal control”, “GAAP”, “accounting”, and “fiscal”.

2.2.4. Concept derivation

The last phase of the ASKE cycle is devoted to concept derivation whereby new concepts are (possibly) introduced in the ACG. To enforce concept derivation, for each concept c_i in the current ACG, the embedding vectors \bar{w} of the terms belonging to it are clustered using the Affinity Propagation algorithm [13]. Affinity propagation determines the number of resulting clusters without having the user explicitly specifying it, and allows for outliers, meaning that an observation may constitute a cluster on its own if it is found too dissimilar from all the others.

On the basis of the clusters obtained from a concept c_i in the current ACG, concept derivation consists in the following operations in the new ACG:

1. a new concept c_j is defined for each resulting cluster Cl_j ;
2. the label of c_j is chosen as the term closest to the centroid of Cl_j . The cluster c_j that contains the term w corresponding to the label of c_i is said to “conserve” c_i in the resulting ACG; for all other concepts, c_j represents a truly new concept in the resulting ACG, and a derivation relationship is defined between c_i and c_j ;
3. for each newly defined concept c_j , a deduplication check is performed: if all the terms associated with c_j already belong to an existing concept c in the current ACG, then c is maintained in the new ACG while c_j is discarded (that is, the older concept preempts the younger).

Example 2.4. When applied to terms extracted in the example of Section 2.2.3, the concept derivation phase produces concepts:

1. “AUDIT”, a new concept derived from “ACCOUNTING SYSTEM” to represent the cluster containing terms “audit” and “review”;
2. “ACCOUNTING”, a new concept derived from “ACCOUNTING SYSTEM” to represent the cluster containing terms (“accounting” and “balance sheet”);
3. “FINANCIAL”, a new concept derived from “ACCOUNTING SYSTEM” to represent the cluster containing (“financial” and “fiscal”);
4. “BANKRUPTCY”, a new concept derived from “ACCOUNTING SYSTEM” to represent the term “bankruptcy” alone (singleton cluster);
5. “ACCOUNTING SYSTEM”, which is the conservation of the original concept, associated with the cluster containing remaining terms, i.e. “accounting standard”, “internal control”, “GAAP” and the term corresponding to the seed concept “accounting system”.

As a consequence of the de-duplication check, concept “ACCOUNTING” is discarded as its term cluster is a subset of the cluster of concept “FINANCIAL STATEMENT”, previously derived from another seed concept.

2.2.5. ASKE endpoint

Concept derivation phase concludes the extraction cycle and the current ASKE generation. The resulting ACG, with newly defined/maintained concepts, associated labels, terms and relationships constitute the input for activating a new generation. The ASKE extraction process on a given corpus may continue for a number of generations. Decision to run a new ASKE generation is taken by the user on the basis of the extraction results in the current generation. The user can decide to stop running ASKE after the current cycle if extracted knowledge is considered satisfactory or if poor new terminology has been extracted with respect to the previous generation. There is, however, a formal condition for the termination of the ASKE extraction process on a given document corpus, that is, when no new terminology is extracted in the current generation. In this case, the user is notified that a new cycle is not triggered (i.e., ASKE reaches its endpoint). If the corpus is extended with new documents, the extraction process can be started again, using the last ACG version as the input seed for the new extended corpus.

Table 1

EurLex dataset statistics.

	Train split	Entire dataset
N. Documents	45,000	57,000
N. Labels from EuroVoc	4,108	4,271
Max. Labels per Document	26	26
Avg. Labels per Document	~5	~5
Avg. Words per Document	~729	~727

3. Evaluating ASKE

A quantitative evaluation of the ASKE approach was conducted on a split of the EurLex dataset [14] containing 45,000 EU legislative documents in English from the EUR-LEX portal.³ Documents include regulations, directives and decisions from the European Commission and the Council of the EU, each of which is annotated by the Publication Office of the EU with one or more labels from the EuroVoc thesaurus⁴ EuroVoc is a taxonomical and multilingual thesaurus of terms pertaining to the activities of the EU in several domains. Labels at the high, more general levels of the taxonomy tend to correspond to general topics or abstract concepts, while labels at lower levels represent specific entities (e.g. formal institutions, animal species involved in regulations, etc.). Since EuroVoc does not satisfy our two requirements for a suitable external dictionary (see Section 2.2.3), we only use EuroVoc for evaluation, and we adopt WordNet as an external dictionary for ASKE. While the dataset has originally been presented for Large-scale Multi-label Text Classification (LMTTC) task, we use it in a Knowledge Extraction perspective exploiting EuroVoc concepts and entities as ground truth. Dataset statistics are summarized in Table 1.

The evaluation objective is twofold:

1. to evaluate the quality of the Knowledge Extraction process in terms of the resulting knowledge model, by assessing the capability of ASKE to reconstruct the EuroVoc labels as extracted concepts, and
2. to evaluate the quality of the Document Classification process, by assessing the correctness of ASKE concepts assigned to each document against the ground truth labels.

Given the ground truth dataset, we will first describe the evaluation setting, and then present the experimental results.

3.1. Evaluation setting

While most evaluation methodologies in the field of Knowledge Extraction are based on exact or partial matching, the interest towards semantic-aware evaluation techniques has been rising in recent years in related fields, such as Machine Translation [15–17]. Such metrics aim at solving the problem of assessing the correctness of the meaning of the translated sentence, rather than its string form, which may differ due to the use of synonyms, paraphrases or circumlocutions. In the same way, in all our evaluation objectives, we want to evaluate concepts extracted by ASKE based on their semantic content and independently from their syntactical form (i.e., identifying term), which is only defined to provide a human-readable representation.

Inspired by YiSi [16], a metric originally proposed to evaluate Machine Translation from a semantic perspective, and FDTC [18], a metric to assess the quality of document-term co-clusters using word embeddings, we adopted an evaluation technique based on global (non-contextual) embeddings. Having trained a global embedding model on the whole corpus and the ground truth labels, the extracted concepts can be evaluated against the ground truth (EuroVoc or a subset of

³ <https://eur-lex.europa.eu/homepage.html>.

⁴ <https://op.europa.eu/s/yTaY>.

it) by computing the similarity between the embedding vectors of the two instead of the exact matching. As a global embedding model, we adopted FastText for its capacity to model out-of-vocabulary words by exploiting sub-word information while performing at the same level as previous models on word similarity and word analogy tasks [19].

Given a set C of extracted concepts, the set G of ground truth labels, and the embedding model F , then a pseudo-precision metric can be defined as:

$$\hat{P} = \frac{1}{|C|} \sum_{c \in C} \max_{g \in G} \sigma(F(c), F(g)) \quad (3)$$

where σ is the cosine similarity.

Conversely, a pseudo-recall metric can be defined as:

$$\hat{R} = \frac{1}{|G|} \sum_{g \in G} \max_{c \in C} \sigma(F(c), F(g)) \quad (4)$$

Indeed, false positives will have no high-similarity correspondence in G , thus reducing \hat{P} , while false negatives will have no high-similarity correspondence in C , thus reducing \hat{R} . The resulting metrics are analogous to YiSi metrics, except for the absence of word frequency-related weights.

In evaluation objective (1), extracted concepts are compared against all EuroVoc labels that are present in the EurLex dataset. Since the ASKE approach considers concepts as clusters of terms, the embedding vector of a concept c is computed as the center of the embedding vectors of the terms that constitute it (i.e. an *average linkage* approach is adopted). EuroVoc labels, instead, can be straightforwardly embedded as individual terms.

As for objective (2), we compute \hat{P} and \hat{R} and average the results at the document level. Then we compute the mean and standard deviation of the document-level metrics.

No benchmark is available for our task in the legal domain, and topic models are suitable competitors in the concept extraction and document classification tasks. In particular, neural topic models based on contextual embeddings offer an ideal baseline, exploiting techniques that are similar to ours. Therefore, for all objectives, ASKE is compared against:

- BERTopic [20], a state-of-the-art unsupervised approach to topic modeling based on transformers that, analogously to ASKE, doesn't require a predefined number of topics;
- Zero-Shot TM (ZSTM) [21], a topic model which exploits document contextual embeddings to extract a predefined number of topics and perform zero-shot classification of documents.

We experiment with 4 configurations for ZSTM and 2 for BERTopic due to the need to set the number of topics in the former. In particular, the number of topics extracted by ZSTM is set to 1,000 and 2,000 to match the order of magnitude of the number of concepts extracted by ASKE. In each case, we retain 5 terms per topic as it approximates the average term-per-topic ratio of our approach, plus we try with 10 terms per topic. As for BERTopic, we retain again 10 terms per topic, plus we evaluate a configuration in which terms within the lowest 5% quantile of TF-IDF score are filtered out so to reduce the occurrences of stop words in the results. Thus the final number of baseline models amounts to 6.

3.2. Experimental results

Experiments are run under various configurations of hyperparameters α and β (which impact document chunk classification and term enrichment phases, respectively), and number of ASKE generations (up to 21), using allmpnet-base-v2⁵ as embedding model. As seed keywords, we select 23 concepts from the EuroVoc thesaurus which are

associated with an explicit definition and which don't have any broader term in the taxonomy structure, so that they can be considered as "root labels". These labels are: *accounting system, adaptation to climate change, administrative check, administrative sanction, code of conduct, credit guarantee, emission allowance, emission trading, entrepreneurship, ESC opinion, EU body, EU financing, European Central Bank opinion, European Union, financing of the EU budget, futures market, gender equality, geochemistry, household, referendum, soil conditioning, teleworking, vocational training*.

The number of concepts and terms that ASKE extracts from the EurLex corpus with this setting is shown in Fig. 4, which represents the configurations that produce better results in our experiments, i.e. $\alpha, \beta \in [0.2, 0.4]$ with $\gamma = 5$.

We notice that the number of concepts extracted by ASKE in its most prolific configuration ($\alpha = \beta = 0.3$) surpasses 1,000 at generation 15 and the number of topics produced by BERTopic at generation 21. It is crucial to evaluate whether the growth in quantity of extracted concepts corresponds to an improvement in their quality.

In terms of execution time, the first generation of ASKE is the most expensive, requiring ~ 14 seconds per concept, due to the additional work to embed initial concepts. Then, time per concept rapidly decreases towards a cost around 2 and 3 seconds.⁶

3.2.1. Evaluation of concept extraction

In Fig. 5, we display the mean and standard deviation of \hat{P} and \hat{R} for the evaluation of objective (1), i.e. concept extraction.

While suffering a substantial gap in pseudo-recall at earlier generations, ASKE achieves state-of-the-art performance within 15 generations, with a growing trend resulting from the extraction of new concepts. Mean pseudo-precision never falls much below the best performance achieved by the baselines, proving that, even when the number of concepts gets higher, ASKE captures concepts that are closer to the ground truth ones. The pseudo-precision curve is not monotonous since concepts acquire new terms at every generation, thus moving in the vector space, but it presents a weakly decreasing trend with small changes in standard deviation after few generations. The precision-recall curve shows an improvement over the baseline models despite this trend. Finally, we notice that performance is almost unaffected by the choice of hyperparameters α and β in the feasible ranges. β has a weaker impact than α , and the impact of the latter is non-linear, $\alpha = 0.3$ yielding slightly better results when compared to $\alpha = 0.2$ or $\alpha = 0.4$.

In Table 2 we provide numerical results, considering generation 15 and 21 of the ASKE cycle in order to approximate the number of topics extracted, respectively, by ZSTM@1000 and BERTopic. A more detailed look at means and standard deviations of \hat{P} and \hat{R} reveals that our model always performs at least at the same level of the baselines, albeit presenting higher variability in \hat{P} with respect to ZSTM and in \hat{R} with respect to BERTopic. At generation 15, \hat{P} is substantially equal to that of ZSTM@1000, but ASKE performs better in terms of \hat{R} , at the expense of marginally higher variability. At generation 21, ASKE beats both baselines, with \hat{P} substantially at the same level as ZSTM@2000 but presenting a diminished variability in \hat{R} .

The most positive aspect of these results is the fact that the good performances of ASKE are obtained without the need to define in advance the expected number of concepts, which is in many cases impossible and hardly compatible with the need for an exploratory investigation of a textual corpus.

Performing a qualitative error analysis by looking at proxies (most similar concepts) to ground truth labels, we found that false negatives, i.e. labels that have no high-similarity correspondence in the extracted concepts, are mostly located at the lowest levels of the taxonomy,

⁵ <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>.

⁶ The computation time refers to the current Python implementation of ASKE, configured with $\alpha = \beta = 0.3$ and $\gamma = 5$, executed on a Ubuntu 20.04.4 LTS server with 8 CPUs, endowed with NVIDIA A100 80 GB PCIe GPU and CUDA 11.8.

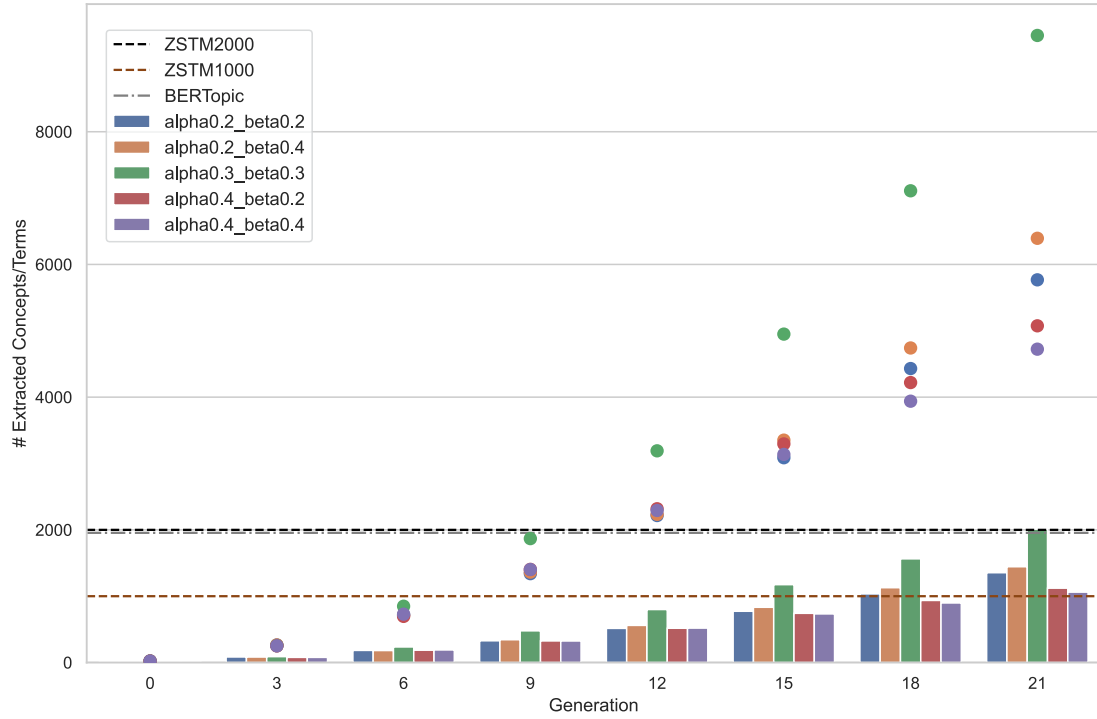


Fig. 4. Number of concepts (bars) and terms (points) extracted by ASKE under different hyperparameter configurations and through the 21 generations, compared with the baselines. The number of terms retrieved by the baseline models was not reported as it can be straightforwardly derived by multiplying the number of concepts by 5 or 10. The filtered version of BERTopic, instead, retains 18,574 terms.

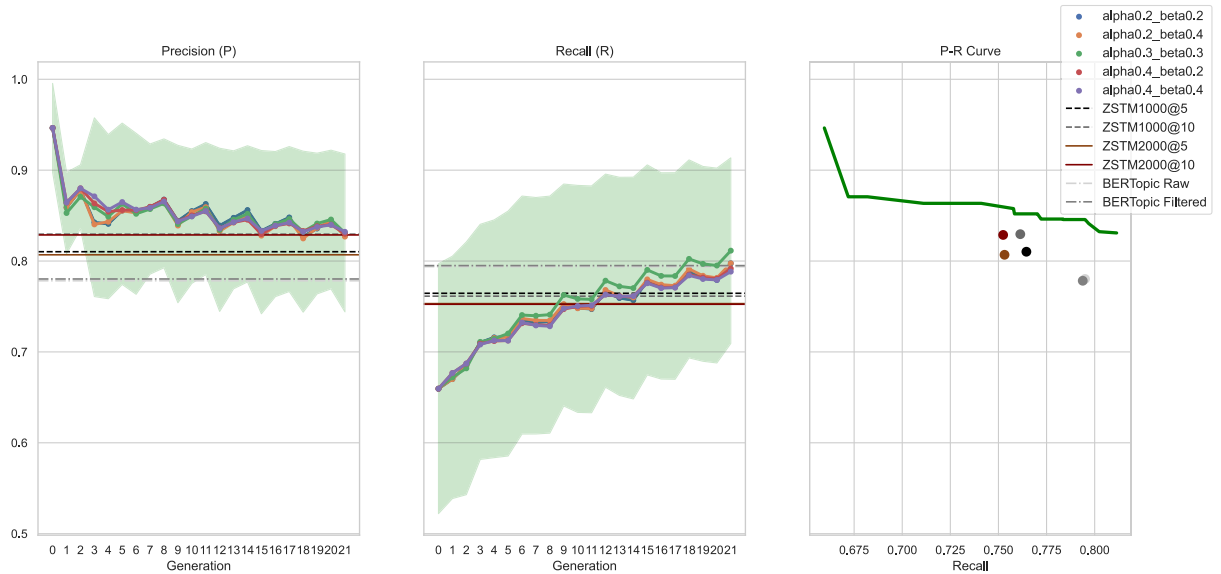


Fig. 5. Results for concept extraction task: mean \hat{P} , \hat{R} , and $\hat{P} - \hat{R}$ curve. For the sake of clarity the standard deviations of baseline models were omitted from this chart, while for ASKE only standard deviation of the model with $\alpha = \beta = 0.3$ is shown, differing little from other configurations.

Table 2

Means and standard deviations (sd) of \hat{P} and \hat{R} for the best configuration of ASKE and the baselines. According to a Welch t-test, difference between ASKE best results and baseline best results is found to be statistically significant for \hat{R} with p -value < 0.01 , and for \hat{P} with p -value < 0.1 .

	Mean \hat{P} (sd)	Mean \hat{R} (sd)	\hat{F}
ASKE ($\alpha = \beta = 0.3$, gen. 15)	0.832 (0.090)	0.790 (0.116)	0.810
ASKE ($\alpha = \beta = 0.3$, gen. 21)	0.831 (0.087)	0.811 (0.102)	0.821
ZSTM@1000 (T@5)	0.810 (0.065)	0.765 (0.102)	0.787
ZSTM@1000 (T@10)	0.830 (0.045)	0.761 (0.103)	0.794
ZSTM@2000 (T@5)	0.807 (0.071)	0.753 (0.112)	0.779
ZSTM@2000 (T@10)	0.829 (0.046)	0.752 (0.107)	0.789
BERTopic (raw)	0.778 (0.134)	0.794 (0.099)	0.786
BERTopic (filtered)	0.780 (0.135)	0.794 (0.099)	0.788

Table 3

Results for document classification (with standard deviation) of ASKE model with hyperparameters $\alpha = \beta = 0.3$ at generations 0, 15 and 21, compared with the best performing configuration of the baseline models. All differences between the best and second-best results were found to be statistically significant, with p -value < 0.01 , by performing a Welch t-test.

	Mean \hat{P} (sd)	Mean \hat{R} (sd)	\hat{F}
ASKE (gen. 0)	0.502 (0.091)	0.478 (0.101)	0.490
ASKE (gen. 15)	0.585 (0.065)	0.670 (0.074)	0.625
ASKE (gen. 21)	0.583 (0.070)	0.643 (0.080)	0.612
BERTopic (filtered)	0.538 (0.079)	0.559 (0.075)	0.548
ZSTM@1000 (T@10)	0.628 (0.079)	0.594 (0.082)	0.611
ZSTM@2000 (T@10)	0.630 (0.080)	0.597 (0.085)	0.613

corresponding to real-world entities (e.g. names of regions or provinces, specific institutions as *anti-dumping duty*, *EU Accession Treaty*, *World Intellectual Property Organization*, etc.) rather than abstract concepts, most of which were correctly retrieved (e.g. *public contract*, *healthcare*, *banking*, *aquaculture*, etc.). Conversely, false positives, i.e. concepts that have no high-similarity correspondence in the ground truth label set, reveal an occasional excess of generality, which is manifested in the retrieval of out-of-domain concepts such as *call* (in the sense of *phone call*), *field* (in the sense of *sector*), *claim*, *form*, etc.

3.2.2. Evaluation of document classification

Table 3 reports the evaluation results for document classification (objective (2)), considering the ASKE model with $\alpha = \beta = 0.3$, which is the hyperparameter configuration yielding the best results, and the baseline models.

From the table, it can be appreciated how concept extraction affects classification as well: after 15 generations, performance has already improved substantially, while it tends to stabilize at later generations.⁷ Indeed, ASKE is outperformed by both baselines at generation 0, but it achieves significantly higher and less dispersed \hat{P} and \hat{R} as the number of generations grows. Both at 1000 extracted concepts (generation 15 for ASKE) and 2000 (generation 21), obtained pseudo-precision still falls below ZSTM, while pseudo-recall is superior. This outcome, in the field of legal search, is particularly desirable to ensure an adequate coverage of the jurisprudence of interest. All in all, our model achieves state-of-the-art performance for the document classification task as well, also in this case with no need to pre-define the expected number of concepts.

4. ASKE at work in a real application

In this section, we describe the application of ASKE to a case study in the framework of the *Next Generation UPP (NGUPP)* project, funded

by the Italian Ministry of Justice, aiming at providing artificial intelligence and advanced information management techniques for digital transformation of Italian legal processes and digital justice in general.

The application is based on a corpus of 50 Italian case law decisions, retrieved from different legal data banks; all the considered documents concern first degree verdicts regarding the matter of unfair competition in the sale of commercial products, coming from different courts on the Italian territory. Unfortunately, such kind of documents does not always come in ready-to-use formats: text has to be extracted from PDF files having different structures depending on the court emitting it. Despite this heterogeneity in the structure of the documents, we do not employ any particular data cleaning process aimed at removing noise from the text. Operating on smaller portions of text, i.e. document chunks, instead of considering the document as a whole, allows us to ignore this issue given that the irrelevant chunks will not be taken into account in the knowledge extraction process.

In the following, we consider two different use-cases of ASKE. The first use-case, denoted as UC-A, is about a legal practitioner with a given case law of interest whose goal is to exploit ASKE to obtain pertinent provisions and rules of law as well as scholar interpretations and Court decisions. As a further use-case, denoted as UC-B, we consider a general subject, like a citizen, a company, or Public Administration employee, interested in improving the general knowledge and understanding of the considered general subject (i.e., unfair competition in the sale of commercial products). Due to the different legal expertise of the two individuals, we define two different types of seed provided by the legal practitioner and the general subject respectively. For the legal practitioner in UC-A, we consider the following query seed: *acts likely to cause confusion*. For the general subject in UC-B, we consider the following keyword seed: *imitation, bag, distinctive elements*. The two seeds specified in the Italian language have been used to trigger ASKE and they are reported here in English for the sake of readability. We setup a configuration with 21 generations and hyperparameters $\alpha = \beta = 0.3$. To collect enough terminology for an effective clustering, we perform the concept derivation phase every three generations. The employed embedding model,⁸ namely paraphrase-multilingual-MiniLM-L12-v2, supports multiple languages and thus it can be used to manage Italian documents. As external dictionary, we relied on Open Multilingual WordNet [22], which provides access to WordNets in a variety of languages, including Italian.

In terms of execution time of both UC-A and UC-B, in the first generation, ASKE requires ~ 11 seconds per concept. The time per concept rapidly decreases towards a cost around 1 and 2 seconds.⁹

In Table 4, we report some statistics about the elements of ACG during the ASKE execution at each generation.

We note that the number of concepts C and terms \mathcal{W} are still growing after 21 generations, even though the rate of growth is much smaller compared to the initial ASKE generations. We also note that the ratio of terms-per-concept \mathcal{W}/C stabilizes around 4 from the generation #13. The “Maintained C ” are concepts of a generation that come from the previous generation, while the “New C ” are the new concepts derived in the current generation. “Deactivated C ” refers to the concepts having the *active* attribute set to False, meaning that they do not have associated document chunks. “Discarded C ” are concepts that have been dropped in the concept derivation phase.

⁸ Model available at <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>.

⁹ The computation time refers to the current Python implementation of ASKE, configured with $\alpha = \beta = 0.3$ and $\gamma = 5$, executed on a Ubuntu 20.04.4 LTS server with 8 CPUs, endowed with NVIDIA A100 80 GB PCIe GPU and CUDA 11.8.

⁷ This outcome holds true for any hyperparameter combination we experimented with.

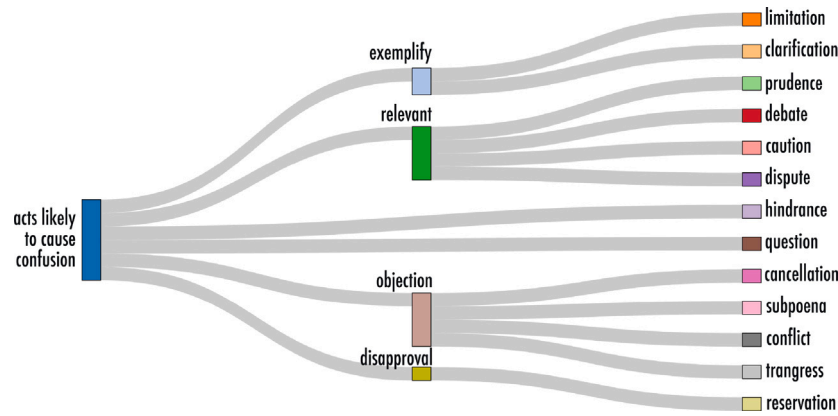


Fig. 6. A portion of the ACG related to the use-case UC-A.

Table 4

Summary statistics for each generation of the ASKE cycle run on the 50 Italian case law decisions.

Generation	C	W	W/C	Maintained C	New C	Deactivated C	Discarded C	K/C
0	2	11	5.5	0	2	0	0	503.5
1	2	21	10.5	2	0	0	0	445.5
2	2	31	15.5	2	0	0	0	508.0
3	10	41	4.1	2	8	0	0	692.8
4	10	76	7.6	10	0	0	0	691.6
5	10	115	11.5	10	0	0	0	697.4
6	42	154	3.67	10	34	0	2	711.57
7	41	266	6.49	42	0	0	1	712.71
8	41	344	8.39	41	0	0	0	721.85
9	110	411	3.74	41	91	3	19	704.13
10	104	523	5.03	110	0	0	6	752.25
11	103	595	5.78	104	0	0	1	777.67
12	182	663	3.64	103	121	21	21	701.92
13	176	766	4.35	182	0	0	6	743.94
14	172	822	4.78	176	0	0	4	758.32
15	239	881	3.69	172	134	47	20	675.42
16	234	1021	4.36	239	0	0	5	721.44
17	228	1081	4.74	234	0	0	6	756.74
18	281	1121	3.99	228	111	62	0	666.2
19	276	1225	4.44	281	0	0	1	722.8
20	270	1283	4.75	276	0	0	6	752.69
21	302	1317	4.36	270	98	54	12	633.21

4.1. UC-A: ASKE for legal practitioners

As mentioned before, the seed used for UC-A is *acts likely to cause confusion*. The choice of the seed is very appropriate from a practitioner's perspective because the seed is a portion of a provision of law on unfair competition (in the present case, art. 2598 of the Italian Civil Code). In Fig. 6, we show a portion of the ACG returned by ASKE.

As a general consideration, legal experts noted that the extracted concepts describe notions that are generally pertinent to the given seed. This is true for example for the link between the notion of *confusion* (included in the seed) and the concepts *exemplify*, *clarification* and *limitation* extracted by ASKE. Again, the likeliness of causing confusion expressed in the seed is pertinent to the notion of relevance (concept *relevant*) and to concepts *prudence*, *debate*, *caution*, and *dispute*, in the sense that a behavior that is relevant in potentially determining confusion might lead to a debate or a dispute and, hence, should be performed with prudence and caution. The same can be said for the link between the notion of confusion included in the seed and the related concepts of *question* and *hindrance*.

From the specific practitioner perspective, the text chunks associated with a given concept have been evaluated as useful if returned as answers to a legal research involving that concept. For example, hereafter we report the top-2 document chunks similar to the initial seed *acts likely to cause confusion*:

Suitability to cause confusion, therefore, consists of two elements: (1) the originality of the imitated product, endowed with distinctive capacity,

such as to become inherent, in the image with the consumer, of the product itself; (2) the absence of distinctive elements capable of showing that the origin of one product is different from that of the other.

[...] (b) conversely, infringement only exists where there is a likelihood of confusion for the public, consisting even in a mere danger of association between the distinctive elements. Prerequisites for the aforementioned discipline to operate, therefore, are: (i) the existence of substantial identity or similarity between the signs; (ii) their use for goods and services that belong to the same sector and are intended to satisfy the same market requirements; (iii) identity of characteristics in the eyes of the same average consumer or only relative affinity.

These chunks are explanations of the legal concept of likelihood to cause confusion, one of the key elements for assessing unfair competition. These chunks are thus pertinent for the purpose and goals of a legal research on *unfair competition*. A legal practitioner can explore these two chunks and all other text chunks associated with the concept *acts likely to cause confusion* to analyze examples of rules of law that may be applicable to the case currently at hand. Indeed, the main focus of practitioners' work is interpretation with the ultimate purpose of solving legal issues: practitioners, both lawyers and judges, exploit text chunks associated with a given concept for interpretation, in order to understand whether they fit a specific factual case and are suitable for solving it.

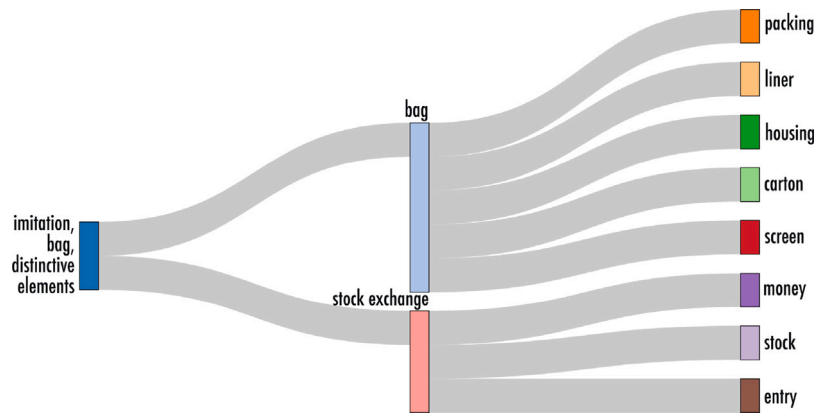


Fig. 7. A portion of the ACG related to the use-case UC-B.

4.2. UC-B: ASKE for general subjects

In Fig. 7, we show another portion of the ACG returned by ASKE that can be useful for UC-B. In this use case, the seed utilized to trigger the ASKE extraction process includes simple keywords like *imitation*, *bag*, *distinctive elements*. The selection of these keywords reproduces the background of a generic user who is interested in understanding the circumstances in which a bag can be considered a counterfeit of another, without specialized legal knowledge and not familiar with legal terminology and specific sentence formulation typical of case law on the matter.

In this case, we note that the nodes in the graph represent more general concepts, not necessarily related to the legal field. By using a generic term like *borsa* (bag) combined with words that are closer to legal terminology such as *imitazione* (imitation) and *elementi distintivi* (distinctive elements), the resulting ACG is characterized by two different branches whose concepts express different meanings. In Italian, the term *borsa* is polysemous and it can refer to a container used for carrying money and small personal items or accessories (the English word “bag”) or to an exchange where security trading is conducted by professional stockbrokers (the English expression “stock exchange”). It is worth noting that ASKE distinguishes these two meanings of “*borsa*” from the initial generations by creating two different concepts (*bag* and *stock exchange* in Fig. 7) which are both labeled *borsa* in the Italian graph. Furthermore, ASKE is capable to properly classify the document chunks according to the pertinent concept meaning between the two available. As an example, below we show two document chunks associated with the concepts *bag* and *stock exchange*, respectively:

As noted in the aforementioned judgment No. 5443/2017, “in the present case, such reproduction also applies to details such as, for example, the slightly rounded flap situated between the two handles and covering part of the zip fastener which, if they constitute an integral part of the shape of the bag model, nevertheless also appear to be elements in themselves capable of impressing themselves on the mind of the consumer who will be able to distinguish between products even legitimately having similar shapes, the one attributable to the source of production constituted by the present plaintiffs”.

Therefore, S.’s clients who intend to make investments of a financial nature first enter into a so-called ‘placement contract’ with S. itself, and then enter into the actual contracts relating to their investment (subscription of units of mutual investment funds, or shares in SICAVs, or conclusion of an insurance policy, or conclusion of a portfolio management contract) directly with the ‘product companies’ contracted with the plaintiff.

If we consider the first relevant chunk above, which pertains to the concept of *borsa* as an object, we can see that it provides valuable information to the user. For instance, it highlights that the resemblance between two objects becomes significant when this resemblance, characterized by specific distinctive elements, has the potential to confuse and mislead consumers who are considering a purchase. The knowledge extraction experiment has been evaluated positively in terms of its significance and applicability by non-expert users in the project. Understanding and navigating the intricate legal landscape, with its complex structure and content, is hard or even nearly impossible for final users and citizens, even considering small-size corpora of legal documents. The possibility to visualize and interact with the ACG representing the extracted knowledge has been considered a very important and valuable feature for a non-expert user. The potential benefits and practical outcomes that could arise from the widespread adoption of the ASKE approach have been evaluated as very positive especially in view of its application in-the-large, catering to the needs of a wide audience of citizens and generic users.

5. Related work

Work related to the issues discussed in this paper is about Legal Information Retrieval (LIR) approaches based on knowledge extraction and NLP techniques. LIR is the discipline that aims at extracting information from a corpus of legal documents, including case law decisions and legal codes, with the goal of supporting effective retrieval by legal practitioners and legal users to identify useful information for their job and needs. A very first system using text analysis techniques to overcome basic boolean search [23] in the legal domain has been proposed in [24], where a document representative was automatically extracted from text, containing index information necessary and sufficient to match the document with a query posed by a legal user. With the recent progress of digital transformation, increasing interest and efforts have been devoted to the development of advanced, AI-based LIR approaches to process huge volumes of legal digital documents and extract knowledge from them. In [25], an approach to assist legal professionals in comparing relevant precedents is presented; the approach extracts and classifies sentences in breach of contract court decisions according to predefined sentence types. A main obstacle encountered when analyzing legal documents is the lack of sufficient annotated data, especially in languages different from English; this aspect is crucial given the requirements of modern neural network-based language models. For example, in [26], a method for similarity case retrieval based on the legal facts is proposed, whose model combines topic distribution and legal entity facts to make the document representation vector more suitable for legal scenarios, with focus on text similarity

problem for Chinese. In [27], information extraction approach for named entity recognition has been presented, with focus on German legal documents. Other approaches to information extraction exploit external resources, such as ontologies combined with NLP techniques such as [28], where a tool for semantic analysis and annotation of legal documents is presented, based on an ontology of deontic concepts. In [29], an approach combining linguistic information provided by WordNet together with NLP techniques is proposed for the extraction of rules from legal documents, while [30] aims at extracting legal rules using hierarchical recurrent neural networks. The increasing interest towards the application of artificial intelligence techniques to the legal field brought to the proposal of several competitions related to the analysis of legal documents and related datasets. The most relevant for the purpose of this paper is the COLIEE [31] competition, where tasks for legal information extraction from case law and statute law are proposed.

Despite the increasing interest about the application of artificial intelligence techniques to the legal field, there is still a need for approaches suitable to deal with unlabeled datasets. For this reason, one of the topic of interest for our work has been the zero-shot learning (ZSL) approach. ZSL is a problem setup in the field of machine learning, where a classifier is required to predict labels of examples extracted from classes that were never observed in the training phase. First referred to as *dataless classification* in 2008 [32], ZSL has quickly become a subject of interest in the field of NLP. The great advantage of this approach consists in the resulting classifier being able to operate efficiently in a partially or totally unlabeled environment. It is possible to categorize ZSL techniques according to three different criteria, as explained in [33]: the learning setting, the semantic space and the method. Firstly, ZSL can be applied on a completely unlabeled dataset, as in the original paper [32], or on a partially labeled one, like in [34]; with this last approach, called generalized ZSL, the goal of the classifier shifts to distinguishing between observation from already seen classes, and examples from unseen ones. Secondly, one may discern an engineered semantic space from a learned semantic space: the former is designed by humans and can be constructed upon a set of attributes [35] or a collection of keywords [36], while the latter is built on top of the results of a machine learning model, as in the case of a text-embedding space [37]. Finally, ZSL methods can be divided in instance-based [38], whose focus is on obtaining examples for unseen classes, and classifier-based [39], which instead focus on directly building a classifier for unlabeled instances. With respect to the above solutions, the proposed ASKE approach enforces legal knowledge extraction in an unsupervised environment by operating in a text-embedding space, therefore eliminating the need for annotated data. The employed ZSL instance-based method goes under the category of projection methods, which consists in labeling instances (i.e. document chunks and terms in ASKE) by collocating these examples in the same semantic space with class prototypes (i.e., ASKE concepts).

A key component in ASKE is Sentence-BERT [12], a modification of BERT language model [40] that is specifically aimed at representing sentence meaning in a vector space. In the legal domain, LEGAL-BERT, a version of BERT pre-trained on legal corpora [9], has been proposed for the English language, and Italian Legal BERT [41] is under evaluation for the Italian language. Another proposal for Italian legal documents is LamBERTa [42], with a focus on law article retrieval. However, we eventually decided to adopt Sentence-BERT because it has been trained in such a way to ensure consistent representation of the meaning of entire sentences, which was a major requirement in designing ASKE and for dealing with legal language complexity. Indeed, appropriate sentence meaning representation is crucial for the quality of document chunk classification with ASKE concepts and for term sense disambiguation in the concept extraction process. With the consolidation of models that combine consistent sentence representation with in-domain pre-training, an extended version of ASKE based on them could be evaluated.

6. Concluding remarks

In the paper, we presented the ASKE approach to legal knowledge extraction, which is based on a combination of context-aware embedding models and zero-shot learning techniques. A featuring contribution of ASKE is the three-phase extraction cycle, which is executed a number of times (called generations) starting from even a poorly described seed concept to progressively extract new concepts which (i) are representative of the different meanings of the terminology used in legal document chunks and (ii) are used for fine-grained, multi-label classification of legal documents at the chunk level without relying on any document annotation. Another contribution is related to the definition of the ASKE knowledge model, by formalizing entities and relationships featuring the ACG initial population and subsequent evolution according to the progress of the extraction process. According to the obtained experimental results, we note that ASKE mostly outperforms a considered baseline (i.e., BERTopic), and it has comparable performance with respect to the other considered baseline (i.e., Zero-Shot TM). However, as a difference from Zero-Shot TM, it is worth noting that ASKE does not require to predefine the number of target topics to discover. Thus, ASKE results are particularly appropriate to satisfy exploratory information needs in those situations where a priori knowledge about the corpus is not available.

A further contribution of this work is related to the synergic collaboration within an interdisciplinary research team that involves experts from legal, linguistic, data science, and computer science fields. This multiplicity of expertise and background has been fundamental for results validation and tuning purposes. The evaluation feedback of ASKE results from legal and linguistic experts has been important for a semantic analysis of extracted concepts. Furthermore, it has been taken into account for ASKE setup and tuning, to define suitable configurations of the ASKE hyperparameters and to set the number of generations for a document corpus.

Ongoing and future work is in two main directions. First, we are working on further tuning the ASKE model to take into account the specificity of the Italian legal domain. Moreover, the development of a suite of application tools based on ASKE is currently in progress. On one side, the ASKE components for knowledge extraction are being integrated in a service-oriented infrastructure for Italian digital justice. In this context, the ASKE concept graph can be employed by citizens and general subjects for *concept-based query answering* on a given corpus of legal documents like judgments and sentences. Here, we assume that ASKE is triggered on a set of seeds predefined on main legal subject areas (e.g., banking, family, corporate law) according to the composition of the underlying corpus [43]. On the other side, ASKE can be employed as a support service to enforce *legal document building*, where a new case law document for a target case at hand can be composed starting from the most similar and prominent document chunks retrieved by ASKE. For example, by working with the document builder interface, a legal practitioner enters the specific background and the specific claims of the parties (in the form of a query seed). ASKE retrieves the document chunks describing decisions and related motivations extracted from similar precedents, to be used for the composition of the motivations and decision section of the case law document at hand [44].

CRedit authorship contribution statement

Silvana Castano: Conceptualization, Methodology, Writing - Review & Editing. **Alfio Ferrara:** Conceptualization, Methodology, Writing - Review & Editing. **Emanuela Furiosi:** Cooperated in drafting the Introduction and Section 4 and she is the sole authoress of Section 4.2. **Stefano Montanelli:** Conceptualization, Methodology, Writing - Review & Editing. **Sergio Picascia:** Conceptualization, Methodology, Writing - Original Draft, Software, Validation. **Davide Riva:** Conceptualization, Methodology, Writing - Original Draft, Software, Validation. **Carolina Stefanetti:** Cooperated in drafting the Introduction and Section 4 and she is the sole authoress of Section 4.1.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is partially supported by the Next Generation UPP project within the PON programme of the Italian Ministry of Justice and by project SERICS (PE00000014) under the MUR NRRP funded by the EU - NextGenerationEU.

References

- [1] Surden H. Artificial intelligence and law: An overview. *Ga State Univ Law Rev* 2019;35:19–22. URL <https://ssrn.com/abstract=3411869>.
- [2] Waldron J. Stare decisis and the rule of law: A layered approach. *L Rev* 2012;1(11).
- [3] Tomasino R. Il valore del precedente: un'analisi critica. 2023. https://www.associazionemagistrati.it/media/79559/08_Tomasino.pdf. [Accessed 2023].
- [4] Montrose J. Distinguishing cases and the limits of ratio decidendi. *Mod Law Rev* 1956;19(5):525–30.
- [5] Galetta D-U, Pinotti G. Automation and algorithmic decision-making systems in the Italian Public Administration. CERIDAP 2023. <http://dx.doi.org/10.13130/2723-9195/2023-1-7>, URL <https://ceridap.eu/automation-and-algorithmic-decision-making-systems-in-the-italian-public-administration/>.
- [6] Schneider J-P, Enderlein F. Automated decision-making systems in German administrative law. CERIDAP 2023. <http://dx.doi.org/10.13130/2723-9195/2023-1-102>, URL <https://ceridap.eu/automated-decision-making-systems-in-german-administrative-law/>.
- [7] Gamero Casado E. Automated decision-making systems in Spanish administrative law. CERIDAP 2023. <http://dx.doi.org/10.13130/2723-9195/2023-1-119>, URL <https://ceridap.eu/automated-decision-making-systems-in-spanish-administrative-law/>.
- [8] Reichel J. Regulating automation of Swedish Public Administration. CERIDAP 2023. <https://dx.doi.org/10.13130/2723-9195/2023-1-112>, URL <https://ceridap.eu/regulating-automation-of-swedish-public-administration/>.
- [9] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics; 2020, p. 2898–904. <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.261>, Online. URL <https://aclanthology.org/2020.findings-emnlp.261>.
- [10] Baron JR, Lewis DD, Oard DW. TREC 2006 legal track overview. In: Voorhees EM, Buckland LP, editors. Proceedings of the 15th text REtrieval conference. NIST special publication, vol. 500–272, National Institute of Standards and Technology (NIST); 2006, URL <http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf>.
- [11] Bhattacharya P, Ghosh K, Ghosh S, Pal A, Mehta P, Bhattacharya A, et al. FIRE 2019 AILA track: Artificial intelligence for legal assistance. In: Proceedings of the 11th annual meeting of the forum for information retrieval evaluation. New York, NY, USA: Association for Computing Machinery; 2019, p. 4–6. <http://dx.doi.org/10.1145/3368567.3368587>.
- [12] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 2019, arXiv:1908.10084.
- [13] Frey BJ, Dueck D. Clustering by passing messages between data points. *Science* 2007;315(5814):972–6. <http://dx.doi.org/10.1126/science.1136800>, arXiv: <https://www.science.org/doi/pdf/10.1126/science.1136800>, URL <https://www.science.org/doi/abs/10.1126/science.1136800>.
- [14] Chalkidis I, Fergadiotis E, Malakasiotis P, Androutsopoulos I. Large-scale multi-label text classification on EU legislation. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019, p. 6314–22. <http://dx.doi.org/10.18653/v1/P19-1636>, URL <https://aclanthology.org/P19-1636>.
- [15] Lo C-k. MEANT 2.0: Accurate semantic MT evaluation for any output language. In: Proceedings of the 2nd conference on machine translation. Copenhagen, Denmark: Association for Computational Linguistics; 2017, p. 589–97. <http://dx.doi.org/10.18653/v1/W17-4767>, URL <https://aclanthology.org/W17-4767>.
- [16] Lo C-k. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In: Proceedings of the 4th conference on machine translation (Volume 2: Shared task papers, Day 1). Florence, Italy: Association for Computational Linguistics; 2019, p. 507–13. <http://dx.doi.org/10.18653/v1/W19-5358>, URL <https://aclanthology.org/W19-5358>.
- [17] Song Y, Zhao J, Specia L. SentSim: Crosslingual semantic evaluation of machine translation. In: Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies. Association for Computational Linguistics; 2021, p. 3143–56. <http://dx.doi.org/10.18653/v1/2021.naacl-main.252>, Online. URL <https://aclanthology.org/2021.naacl-main.252>.
- [18] Role F, Morbieu S, Nadif M. Unsupervised evaluation of text co-clustering algorithms using neural word embeddings. In: Proceedings of the 27th ACM international conference on information and knowledge management. New York, NY, USA: Association for Computing Machinery; 2018, p. 1827–30. <http://dx.doi.org/10.1145/3269206.3269282>.
- [19] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017;5:135–46. http://dx.doi.org/10.1162/tac1_a_00051, arXiv: https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1_a_00051/1567442/tac1_a_00051.pdf.
- [20] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022, arXiv:2203.05794.
- [21] Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E. Cross-lingual contextualized topic models with zero-shot learning. In: Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume. Association for Computational Linguistics; 2021, p. 1676–83. <http://dx.doi.org/10.18653/v1/2021.eacl-main.143>, Online. URL <https://aclanthology.org/2021.eacl-main.143>.
- [22] Bond F, Paik K. A survey of wordnets and their licenses. In: Proceedings of the 6th global WordNet conference. 2012, p. 64–71.
- [23] Blair DC, Maron ME. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun ACM* 1985;28(3):289–99. <http://dx.doi.org/10.1145/3166.3197>.
- [24] Gelbart D, Smith JC. Beyond Boolean search: FLEXICON, a legal tex-based intelligent system. In: Proceedings of the 3rd international conference on artificial intelligence and law. New York, NY, USA: Association for Computing Machinery; 1991, p. 225–34. <http://dx.doi.org/10.1145/112646.112674>.
- [25] Mok WY, Mok JR. Legal machine-learning analysis: First steps towards A.I. assisted legal research. In: Proceedings of the 17th international conference on artificial intelligence and law. New York, NY, USA: Association for Computing Machinery; 2019, p. 266–7. <http://dx.doi.org/10.1145/3322640.3326737>.
- [26] Hu W, Zhao S, Zhao Q, Sun H, Hu X, Guo R, et al. BERT_LF: A similar case retrieval method based on legal facts. *Wirel Commun Mob Comput* 2022;2022:1–9. <http://dx.doi.org/10.1155/2022/2511147>.
- [27] Leitner E, Rehm G, Moreno-Schneider J. Fine-grained named entity recognition in legal documents. In: Acosta M, Cudré-Mauroux P, Maleshkova M, Pellegrini T, Sack H, Sure-Vetter Y, editors. Semantic systems. The power of AI and knowledge graphs. Cham: Springer International Publishing; 2019, p. 272–87.
- [28] Zeni N, Kiyavitskaya N, Mich L, Cordy JR, Mylopoulos J. Gaiust: supporting the extraction of rights and obligations for regulatory compliance. *Requir Eng* 2015;20(1):1–22. <http://dx.doi.org/10.1007/s00766-013-0181-8>.
- [29] Dragoni M, Villata S, Rizzi W, Governatori G. Combining NLP approaches for rule extraction from legal documents. In: 1st workshop on Mining and Reasoning with legal texts. 2016.
- [30] Chalkidis I, Androutsopoulos I, Michos A. Obligation and prohibition extraction using hierarchical RNNs. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 2: Short papers). Melbourne, Australia: Association for Computational Linguistics; 2018, p. 254–9. <http://dx.doi.org/10.18653/v1/P18-2041>, URL <https://aclanthology.org/P18-2041>.
- [31] Rabelo J, Goebel R, Kim M-Y, Kano Y, Yoshioka M, Satoh K. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *Rev Socionetwork Strateg* 2022;16(1):111–33. <http://dx.doi.org/10.1007/s12626-022-00105->, URL https://ideas.repec.org/a/spr/trosos/v16y2022i1d10.1007_s12626-022-00105-z.html.
- [32] Chang M-W, Ratnikov L, Roth D, Srikumar V. Importance of semantic representation: Dataless classification. In: Proceedings of the 23rd national conference on artificial intelligence - Volume 2. AAAI Press; 2008, p. 830–5.
- [33] Wang W, Zheng VW, Yu H, Miao C. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans Intell Syst Technol* 2019;10(2). <http://dx.doi.org/10.1145/3293318>.
- [34] Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans Pattern Anal Mach Intell* 2019;41(09):2251–65. <http://dx.doi.org/10.1109/TPAMI.2018.2857768>.
- [35] Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. 2009, p. 951–8. <http://dx.doi.org/10.1109/CVPR.2009.5206594>.
- [36] Qiao R, Liu L, Shen C, Hengel AVD. Less is more: Zero-shot learning from online textual documents with noise suppression. In: 2016 IEEE conference on computer vision and pattern recognition. Los Alamitos, CA, USA: IEEE Computer Society; 2016, p. 2249–57. <http://dx.doi.org/10.1109/CVPR.2016.247>, URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.247>.

- [37] Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B. Latent embeddings for zero-shot classification. In: 2016 IEEE conference on computer vision and pattern recognition. Los Alamitos, CA, USA: IEEE Computer Society; 2016, p. 69–77. <http://dx.doi.org/10.1109/CVPR.2016.15>, URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.15>.
- [38] Xu X, Hospedales T, Gong S. Transductive zero-shot action recognition by word-vector embedding. *Int J Comput Vis* 2017;123(3):309–33.
- [39] Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato MA, et al. DeViSE: A deep visual-semantic embedding model. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K, editors. *Advances in neural information processing systems*, vol. 26. Curran Associates, Inc.; 2013, URL https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.
- [40] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019, [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [41] Licari D, Comandé G. ITALIAN-LEGAL-BERT: A pre-trained transformer language model for Italian law. In: *Companion proceedings of the 23rd international conference on knowledge engineering and knowledge management*. 2022, URL <https://ceur-ws.org/Vol-3256/km4law3.pdf>.
- [42] Tagarelli A, Simeri A. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artif Intell Law* 2021;30(3):417–73. <http://dx.doi.org/10.1007/s10506-021-09301-8>.
- [43] Bellandi V, Castano S, Montanelli S, Riva D, Siccardi S. A service infrastructure for the Italian digital justice. In: *Proc. of the 15th int. conference on management of digital EcoSystems*. 2023.
- [44] Castano S, Ferrara A, Montanelli S, Picascia S, Riva D. A knowledge-based service architecture for legal document building. In: *Proc. of the 2nd int. workshop on knowledge management and process mining for law*. 2023.