POLITECNICO DI TORINO
Repository ISTITUZIONALE

Automated Knowledge Extraction from Legal Texts using ASKE

Original
Automated Knowledge Extraction from Legal Texts using ASKE / Castano, S., Ferrara, A., Montanelli, S., Picascia, S., Riva, D.. - 3741:(2024), pp. 446-455. (SEBD 2024: 32nd Symposium on Advanced Database Systems Villasimius (ITA) 23-26 June 2024).

Availability:
This version is available at: 11583/2992932 since: 2024-09-30T12:26:00Z

Publisher:
CEUR

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

04 November 2026

Automated Knowledge Extraction from Legal Texts using ASKE*

(Discussion Paper)

Silvana Castano[1], Alfio Ferrara[1], Stefano Montanelli[1], Sergio Picascia[1,**] and Davide Riva[1]

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

Abstract

In this paper, we present 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 to progressively extract concepts representative of the different meanings of terminology used in legal documents chunks. We show ASKE in action in a case study of legal knowledge extraction from a real corpus of case law decisions in the framework of the *NGUPP* project.

Keywords

Legal Knowledge Extraction, Natural Language Processing, Digital Justice.

1. Introduction

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 AI-based techniques for legal knowledge extraction. The availability of techniques 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 general public. Legal search through legal knowledge extraction is an extremely important instrument for legal practitioners in both common law [2] and civil law [3] systems. 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

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*This paper presents an extended abstract of [1].

**Corresponding author.

silvana.castano@unimi.it (S. Castano); alfio.ferrara@unimi.it (A. Ferrara); stefano.montanelli@unimi.it (S. Montanelli); sergio.picascia@unimi.it (S. Picascia); davide.riva1@unimi.it (D. Riva)

0000000238262407 (S. Castano); 0000-0002-4991-4984 (A. Ferrara); 0000-0002-6594-6644 (S. Montanelli); 0000-0001-6863-0082 (S. Picascia); 0009000396819423 (D. Riva)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3741/paper37.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

(called "rationale") behind the decision. During this process, great help may come from "context-aware" knowledge extraction systems, based on Natural Language Processing (NLP), Machine Learning (ML), and Artificial Intelligence (AI), to deal with challenging requirements posed by the legal documentation (e.g., language complexity, significant length of the legal texts; lack of sufficiently-large annotated corpora for model training).

In this paper, we present ASKE (*Automated System for Knowledge Extraction*), an approach to legal knowledge extraction with focus on abstract concept discovery using 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 the 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 that, at each cycle, progressively incorporates newly extracted legal knowledge into the *ASKE Conceptual Graph* - ACG, a graph-based data structure initially populated through a data preparation step.

After describing the ASKE process in a nutshell, we show ASKE in action in the context of two use cases of knowledge extraction considering a corpus of 50 Italian case law decisions in the framework of the *Next Generation UPP* (*NGUPP*) project. *NGUPP*, funded by the Italian Ministry of Justice, aims at providing artificial intelligence and advanced information management techniques for the digital transformation of Italian legal processes and digital justice in general.

2. ASKE in a nutshell

ASKE is conceived to build a conceptual view over a considered corpus of legal documents. A *data preparation* step is initially executed, and it is followed by an iterative three-step extraction process characterized by i) *document chunk classification*, ii) *terminological enrichment*, and iii) *concept derivation* (see Figure 1).

Data preparation consists in the application of conventional text processing techniques, that are tokenization, lemmatization, and embedding. Tokenization has the goal to separate a document d into chunks. A chunk k represents the text unit to consider for classification and it determines the granularity of the document that can be associated with a concept. A chunk consists of a few sentence/phrase detected in a document, up to a maximum size of 512 words[1]. After tokenization, the terms appearing in chunks are lemmatized and a vector-based representation of each chunk is finally built. To this end, a chunk k is associated with a set of terms W_k therein contained. Any term $w \in W_k$ is described as $w = (w_l, w_d, \overline{w})$, where w_l is the label of the term (i.e., the lemma), w_d is a description of the term taken from a reference dictionary/vocabulary (e.g., WordNet), and $\overline{w}$ is the corresponding vector-based representation

[1]The size of the document chunk is experimentally determined according to the features of the considered corpus. A chunk should be large enough, so that the context can be captured, but not too much extended to avoid segments that are long to read and potentially noisy due to the presence of multiple concepts.

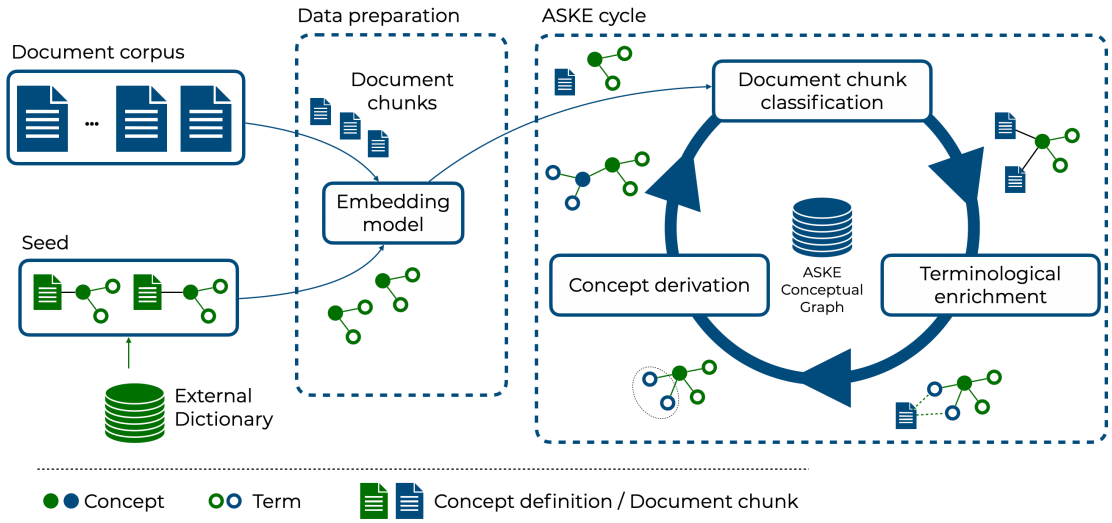


Figure 1: The ASKE approach to legal knowledge extraction.

according to Sentence-BERT [5], respectively. The use of embedding techniques to represent chunks allows to map the document contents on a semantic vector space where the similarity of two chunks can be measured by comparing the corresponding vector representations through a similarity metric (e.g., cosine similarity). Sentence-BERT has been chosen for ASKE since it is trained in such a way to ensure consistent representation of the meaning of entire paragraphs. This is particularly appropriate in the legal field, where the phrase structure can be highly articulated, and some common terms can have a precise technical meaning when used in a court (e.g., citation, clemency, designation). A chunk has the form $k = (k_d, \bar{k})$, where k_d is the original textual content of the chunk and $\bar{k}$ is the corresponding vector-based representation calculated as the mean of term vectors $\bar{w}$ with $w \in W_k$.

For triggering the knowledge extraction process, in data preparation, ASKE requires to specify a set of *seed concepts*. A seed concept can be expressed as a short text (e.g., one or two phrases) providing a gross-grained description of the target. In this case, as a common example, a seed concept can be specified by taking an excerpt from pertinent law/case law documentation. As an alternative, a seed concept can be defined as a list of keywords. As an example, for a seed concept about banking contract, a corresponding list of keywords could be the following: bank deposit, safe deposit box, bank credit opening, bank advance, bank account, bank discount.

Document chunk classification has the goal to annotate chunks with featuring concepts and *zero-shot learning techniques* are employed to this end. Zero-shot learning is an unsupervised classification technique, characterized by the ability to work without requiring any pre-existing annotation of the considered documents. Given a set of concepts (i.e., the seed concepts at the beginning of the process), a similarity measure σ, e.g., cosine similarity, is calculated over any pair of embeddings between chunks and concepts. A chunk k is classified with the concept c when the similarity value satisfies $\sigma(\mathbf{k}, \mathbf{c}) \geq \alpha$, with α defined as a similarity threshold

configured in the system. A concept c in ASKE is defined as a pair $c = (c_l, \bar{c})$, where c_l is a label featuring the meaning of the concept expressed in a synthetic and human-understandable way, and $\bar{c}$ is a vector-based concept representation. Each concept c is initially associated with the set of terms W_c extracted from the textual description of c. The vector concept $\bar{c}$ is built as the mean of the vectors of all the terms in W_c. The label c_l corresponds to the label w_l of the term $w \in W_c$, whose vector representation $\bar{w}$ is closest to the concept vector $\bar{c}$.

Terminological enrichment is then enforced to enrich the term set W_c of a concept c by considering the terms W_k of any chunk k classified with c. The idea is that the initial description of the concept c can become more detailed if we add terminology taken from chunks that are pertinent (i.e., classified) with c. This is done by calculating the similarity between any pair of embeddings $\bar{w}$ and $\bar{c}$ in W_k and W_c. The most similar terms of W_k are inserted in W_c according to a system-defined β similarity threshold.

Concept derivation is finally executed to determine new and more fine-grained concepts that can emerge from existing ones after enrichment. Given a concept c, the Affinity Propagation (AP) algorithm is employed to cluster the embedding vectors $\bar{w}$ of terms in W_c. A new concept c' is created for each cluster returned by AP. A link is defined between a concept c' and c to denote that c' is derived from c and they are somehow similar/related in content. The concept c is then updated since the terms in W_c can be changed due to enrichment. As a consequence, c_l and $\bar{c}$ are re-calculated.

ASKE endpoint. The set of concepts obtained after derivation can trigger the execution of a new cycle based on the above three steps. Each cycle execution is called *ASKE generation*. The new concepts derived in a certain ASKE generation contribute to improve the classification of chunks in more fine-grained concepts. New concepts can also be discovered through a new execution of enrichment and derivation on the basis of a refined classification result. As such, concept extraction terminates when the number of new concepts created in the derivation step is lower than a predefined *termination threshold*. A final concept graph ACG is populated with all the concepts and corresponding derivation links extracted by ASKE.

3. ASKE in action

To show ASKE in action, we consider a case study of legal knowledge extraction from a legal corpus in the context of the *NGUPP* project. The case study dataset is composed by 50 court decisions from several Italian courts, spanning from year 2008 to year 2022. All decisions are first-degree verdicts selected by legal experts of the project for their relevance to the subject matter *unfair competition*. We illustrate knowledge extraction from the *unfair competition* dataset by considering two use cases, modeling the use of ASKE by two categories of users with different levels of expertise, namely, UC-A, related to a legal practitioner user (e.g., a lawyer) with high legal expertise, and UC-B, related to a general subject (e.g., a citizen) with limited legal expertise. To trigger the ASKE extraction process, user A provides a legal definition as a seed concept, i.e., the expression "*acts likely to cause confusion*", taken from art. 2598 of the Italian Civil Code.

In the second case, user B provides a seed in form of general keywords like "*bag, distinctive elements, imitation*". The seeds were in Italian, and translated here in English for readability. We asked two legal experts to evaluate the knowledge output extracted by ASKE in both use cases. In particular, we asked to qualitatively assess i) the pertinence of discovered chunks with respect to the initial seed concept, and ii) the appropriateness of the new ASKE concepts derived from the seed. In both cases, ASKE was running with hyperparameters $\alpha = \beta = 0.3$, number of generations equal to 21, `paraphrase-multilingual-MiniLM-L12-v2`[2] as the embedding model and Open Multilingual WordNet as external dictionary to retrieve term definitions[3].

3.1. UC-A: ASKE for legal practitioners

Figure 2 shows a portion of the concept graph produced by ASKE in case UC-A. Legal experts

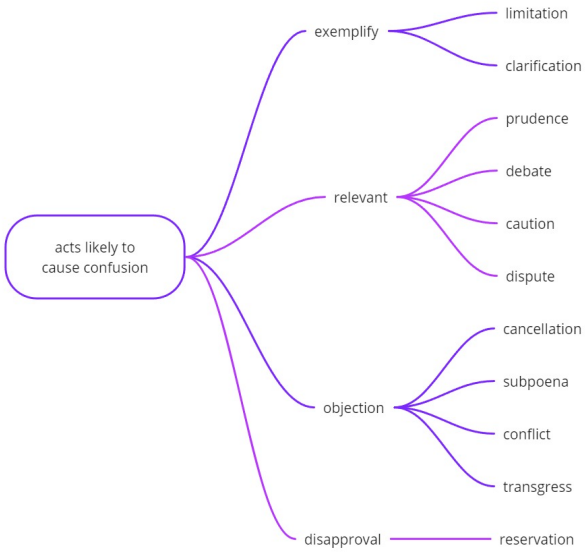


Figure 2: A portion of the ACG for the use-case UC-A

positively noted that concepts derived from the seed are coherent with the topic of the law provision. Indeed, concepts like *exemplify* and *relevant* are related to proof of the uniqueness of a product, while *hindrance* can be interpreted as a consequence of the acts producing confusion. Concept *clarification* is related to *exemplify*, and other derivations, such as that of *prudence*, *debate*, *caution* and *dispute* from *relevant*, though not strictly related from a semantic point of view, were judged as appropriate in this context, since relevance is a key feature in legal debate and disputes, while prudence and caution are exercised to prevent the introduction of misleading, irrelevant or prejudicial information and to evaluate the reliability of evidence.

[2]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[3]Further details about the ASKE configuration in this case study are provided in [1].

Next, legal experts analyzed the chunks with the highest similarity with respect to the seed concept. We report below the top-two similar chunks as an example:

> *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.*

Both chunks are definitions of the conditions that characterize the "*acts likely to cause confusion*". These and other similar chunks identified by ASKE can therefore be exploited by legal practitioners for interpretation of a new case. Indeed, their main aim is to verify that a certain fact fits or doesn't fit some conditions established by the law.

3.2. UC-B: ASKE for general subjects

Figure 3 shows a portion of the concept graph produced by ASKE in case UC-B. In this graph legal experts noticed the prevalence of general concepts over legal ones. Two concepts emerge among others, related to the two senses of the Italian word "*borsa*": "*bag*" and "*stock exchange*". Looking at concepts derived from these ones, it can be noticed that ASKE was able to perform a correct distinction between the two.

Looking at the document chunks with the highest similarity with respect to the seed concept, we highlight two chunks that refer to each of the derived concepts.

> *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*

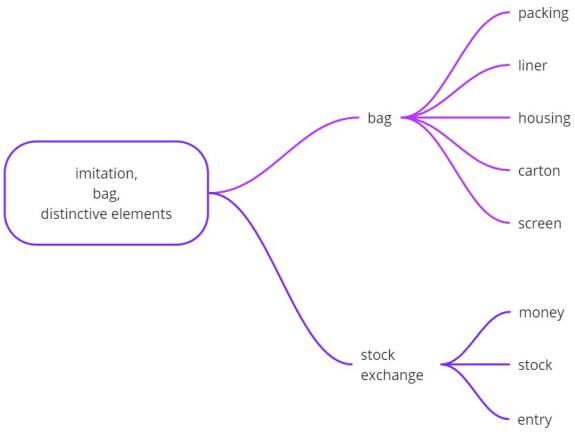


Figure 3: A portion of the ACG for the use-case UC-B

> *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.*

The retrieved chunks were evaluated positively by legal experts from the point of view of their pertinence to the seed. Also the relatedness of extracted concepts starting from initial seed has been evaluated as satisfactory. The application of ASKE for the exploration of a legal corpus proved promising, as the concept graph simplifies the navigation of the underlying corpus making it accessible to even non-expert users.

4. Related work

With the recent progress of digital transformation, increasing interest and efforts have been devoted to the development of advanced, AI-based approaches to process huge volumes of legal digital documents and extract knowledge from them. In [6], an approach to assist legal professionals in comparing relevant precedents is presented; in [7], 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. Semantic similarity is employed in [8], where documents are grouped into clusters, according to their content, and then regularities in the paragraphs are detected for each cluster. In [9], information extraction approach for named entity recognition has been presented, with focus on German legal documents. 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 [10] competition, where tasks for legal information extraction from case law and statute law are proposed.

It is worth noting that ASKE enforces document classification without the need to rely on pre-existing annotations, and without requiring to pre-define the number of target topics/concepts to discover. Thus, ASKE is particularly appropriate to satisfy exploratory information needs in those situations (e.g., the legal domain) where a-priori knowledge about the corpus is not available.

We also note that a key component in ASKE is Sentence-BERT [5], a modification of BERT language model [11] that is specifically aimed at representing sentence meaning in a vector space. LEGAL-BERT, a version of BERT pre-trained on legal corpora [12], has been proposed for the English language, and Italian Legal BERT [13] is under evaluation for the Italian language. Another proposal for Italian legal documents is LamBERTa [14], with a focus on law article retrieval. 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. With the consolidation of these latter models that combine consistent sentence representation with in-domain pre-training, an extended version of ASKE based on them could be evaluated as future work.

5. 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 deep evaluation of ASKE has been performed considering the EurLex dataset [15] containing $45,000$ EU legislative documents in English, each of which is annotated by the Publication Office of the EU with one or more labels from the EuroVoc thesaurus[4]. The goal of the evaluation was twofold: i) to assess the quality of the knowledge extraction process, by assessing the capability of ASKE to reconstruct the EuroVoc labels as extracted concepts, and ii) 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. The results of the evaluation are positive on both sides (see [1]). Ongoing work is related to the inclusion of ASKE in a service architecture for legal knowledge extraction [16]. Furthermore, we are also working on the use of ASKE for enforcing *legal document building*, where a new case law document for a target case at hand can be interactively composed starting from the most similar and prominent document chunks extracted by ASKE.

Acknowledgments

This work was supported in part by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor the Italian MUR can be held responsible for them.

[4] https://op.europa.eu/s/yTaY.

References

[1] S. Castano, A. Ferrara, E. Furiosi, S. Montanelli, S. Picascia, D. Davide, C. Stefanetti, Enforcing Legal Information Extraction through Context-aware Techniques: the ASKE Approach, Computer Law & Security Review 52 (2024).

[2] J. Waldron, Stare decisis and the rule of law: A layered approach, L. Rev 1 (2012).

[3] R. Tomasino, Il valore del precedente: un'analisi critica, https://www.associazionemagistrati.it/media/79559/08_Tomasino.pdf, 2023. Accessed: 2023.

[4] J. Montrose, Distinguishing cases and the limits of ratio decidendi, The Modern Law Review 19 (1956) 525–530.

[5] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. arXiv:1908.10084.

[6] W. Y. Mok, J. R. Mok, Legal machine-learning analysis: first steps towards a.i. assisted legal research, in: Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 266–267. URL: http://doi.org/10.1145/3322640.3326737. doi:10.1145/3322640.3326737.

[7] W. Hu, S. Zhao, Q. Zhao, H. Sun, X. Hu, R. Guo, Y. Li, Y. Cui, L. Ma, BERT_LF: a similar case retrieval method based on legal facts, Wireless Communications and Mobile Computing 2022 (2022) 1–9. URL: http://doi.org/10.1155/2022/2511147. doi:10.1155/2022/2511147.

[8] G. De Martino, G. Pio, M. Ceci, Prilj: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments, Artificial Intelligence and Law 30 (2022) 359–390.

[9] E. Leitner, G. Rehm, J. Moreno-Schneider, Fine-grained named entity recognition in legal documents, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs, Springer International Publishing, Cham, 2019, pp. 272–287.

[10] J. Rabelo, R. Goebel, M.-Y. Kim, Y. Kano, M. Yoshioka, K. Satoh, Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021, The Review of Socionetwork Strategies 16 (2022) 111–133. URL: https://ideas.repec.org/a/spr/trosos/v16y2022i1d10.1007_s12626-022-00105-z.html. doi:10.1007/s12626-022-00105-.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[13] D. Licari, G. Comandé, 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, Bozen-Bolzano, Italy, September 26-29, 2022, 2022. URL: https://ceur-ws.org/Vol-3256/km4law3.pdf.

[14] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code, Artificial Intelligence and Law 30 (2021) 417–473. URL: https://doi.org/10.1007%2Fs10506-021-09301-8. doi:10.1007/

s10506-021-09301-8.

[15] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on EU legislation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6314–6322. URL: https://aclanthology.org/P19-1636. doi:10.18653/v1/P19-1636.

[16] V. Bellandi, S. Castano, S. Montanelli, D. Riva, et al., A service architecture for ai-based legal knowledge extraction, in: CEUR WORKSHOP PROCEEDINGS, volume 3478, CEUR Workshop Proceedings, 2023, pp. 110–119.