The CRIKE Data-Science Process for Legal Knowledge Extraction

Discussion Paper

Silvana Castano[1], Mattia Falduti[1], Alfio Ferrara[1], and Stefano Montanelli[1]

Università degli Studi di Milano
DI - Via Celoria, 18 - 20135 Milano
{silvana.castano,mattia.falduti,alfio.ferrara,stefano.montanelli}@unimi.it

Abstract. In this paper, we present *CRIKE*, a data-science approach to automatically detect concrete applications of legal abstract terms in case-law decisions. To this purpose, CRIKE relies on the use of the *LATO ontology* where legal abstract terms are properly formalized as concepts and relations among concepts. Using LATO, CRIKE aims at discovering how and where legal abstract terms are applied by judges in their legal argumentation. Moreover, we detect the terminology used in the text of case-law decisions to characterize concrete abstract-term instances.

Keywords: legal ontology, legal-term extraction, case-law analysis

1 Introduction

Law is general and abstract by definition. On the opposite, court case law decisions are specific and concrete, in that they provide a peculiar interpretation of law applied to the considered single cases. Legal interpreters, such as for example judges and lawyers, are daily involved in analysis and evaluation of court case law with the aim to extract/derive possible suggestions for incoming case applications by relying on the experience of past applications that can be considered as a sort of consolidated *legal knowledge*.

According to the Italian law, the legal terminology can be distinguished into three main categories, that are i) *statutory terms*, i.e., terms directly or indirectly defined by law; examples of statutory terms are public officer, illicit drug, and consumer; ii) *descriptive terms*, i.e., terms featuring actions, human activities, and any real-life object; examples of descriptive terms are escape, car, and year; iii) *abstract terms*, i.e., terms featuring something indeterminate that requires a concrete application for being really defined; examples of abstract terms are good faith, long-term cohabitation, and dangerous driving. Consider the abstract schema of a legal action provided in Figure 1. When a new case law is received for

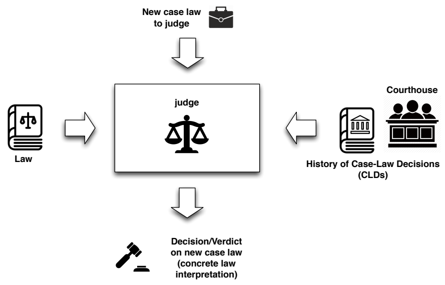


Fig. 1. The abstract schema of a legal action

judgement, the expected evaluation process has to take into account i) the law, for understanding the terms, either statutory, descriptive, or abstract, that can be relevant for the current case, and ii) the history of case-law decisions, for detecting possible relevant interpretations and concrete applications of abstract terms that can be useful to support the decision/verdict to eventually deliver.

In this paper, we present *CRIKE* (CRIme Knowledge Extraction), a data-science approach to detect concrete applications of legal abstract terms in large case-law decisions. To this purpose, CRIKE relies on the use of LATO (Legal Abstract Term Ontology) where legal terms are properly formalized as concepts and relations among concepts. Using LATO, CRIKE aims at discovering how and where legal abstract terms are applied by judges in their legal argumentation.

The paper is organized as follows. In Section 2, the CRIKE approach is introduced. The LATO ontology and the CRIKE techniques for legal knowledge extraction are discussed in Section 3 and 4, respectively. Related work are discussed in Section 5. Concluding remarks are provided in Section 6.

2 The CRIKE approach

The CRIKE approach (see Figure 2) is conceived to support extraction of legal knowledge from a (possibly large) dataset of Case-Law Decisions (CLDs) coming from different, official sources, such as for example First Grade and Court of Appeal judgements. CRIKE embeds the LATO ontology where relevant law concepts of a given domain of interest are properly formalized. To enforce knowledge extraction, CRIKE exploits a given dataset of CLDs in input by adopting a conventional data-science process where each CLD is indexed and stored in a digital format. In particular, the CLDs of our dataset are acquired from the Court and the Court of Appeal of Milan and they are usually provided in image format with highly heterogeneous quality. The indexing and storage activity exploits data cleaning and tokenization techniques to obtain a pure textual version of each CLD as well as a focused set of metadata. By exploiting the indexed CLDs metadata, knowledge extraction is enforced with the aim at classifying a CLD with respect to the LATO ontology knowledge. In particular, extraction

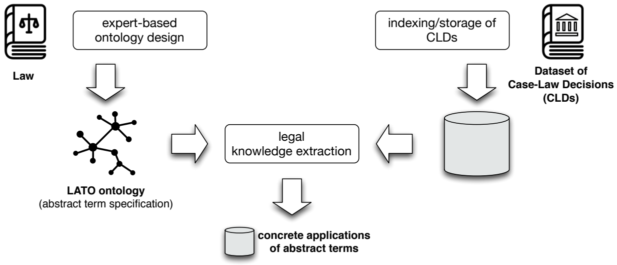

Law
expert-based ontology design
indexing/storage of CLDs
Dataset of Case-Law Decisions (CLDs)
LATO ontology
(abstract term specification)
legal knowledge extraction
concrete applications of abstract terms

Fig. 2. The CRIKE approach

is focused on detecting the concrete applications of legal abstract terms in the text of the considered CLDs. The crucial idea of CRIKE is that the detection of a given abstract term AT is not only concerned with the recognition of single terms featuring AT, but also with the recognition of terms associated with the ancillary concepts related to AT, that we call *abstract-term context*.

Motivating example. Consider the Italian law about drugs and related drug offenses, as reported in [11]. According to the Italian criminal order, "the Consolidated Law, adopted by Presidential Decree No 309 on 9 October 1990 and subsequently amended, provides the legal framework for trade, treatment and prevention, and prohibition and punishment of illegal activities in the field of drugs and psychoactive substances. Drug use in itself is not mentioned as an offense. [...] The threshold between personal possession and trafficking is determined by the circumstances of the specific case (e.g., the act, possession of tools for packaging, different types of drug possessed, number of doses in excess of average daily use, means of organization). The penalty for supply-related offenses, such as production, sale, transport, distribution or acquisition, depends on the type of drug. However, when the **offenses are considered minor** because of the means, modalities or circumstances, the terms of imprisonment are lower. Evaluating whether or not the offense is minor should take into account a set of "ancillary" elements such as the mode of action, possible criminal motives, quality and quantity of drug possessed, the character of the offender, conduct during or subsequent to the offense, and the family and social conditions of the offender". The notion of **minor offense** is an example of abstract term in the above law quotation. A precise definition of circumstances and related threshold quantities to associate with the notion of minor offense is not available/possible in the (abstract) law. Given a specific criminal charge of drug possession, the final decision/verdict is based on the specific interpretation of the abstract term "minor offense" where the specific circumstances and quantities of the considered case represent a concrete application of the corresponding abstract term.

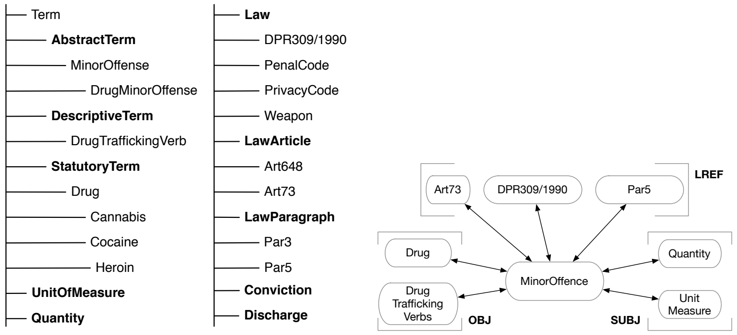
3 Legal knowledge representation

To formalize the knowledge related to abstract terms and their interpretation, we introduce LATO in CRIKE. LATO is a legal ontology where relevant law terms to exploit knowledge extraction in CLDs are defined; it contains concepts to represent general law terms, either abstract, statutory and descriptive terms.

LATO is manually defined by domain experts and implemented according to the SKOS formalism. In particular, the concept hierarchy is based on a root concept Term with three main subconcepts, namely AbstractTerm, DescriptiveTerm, and StatutoryTerm (see Figure 3(a)). In addition to general law terms, the LATO ontology contains concepts that represent the Italian legislative structure, such as for example the concepts Law, LawArticle, and LawParagraph. Furthermore, the concepts Conviction and Discharge are also specified in LATO to represent the possible Court decisions (i.e., the verdict) of a given case law. In particular, the concept Conviction denotes a verdict in which the Court judges the defendant guilty, while the concept Discharge denotes a verdict in which the facts have a penalty relevance, but no punishment is finally delivered. Finally, the concepts Quantity and UnitOfMeasure are defined in LATO for allowing to represent the quantitative estimation of substances that can appear in legal documents.

AbstractTerm is the core concept of the LATO ontology since it represents the target of the knowledge extraction functionalities of CRIKE. The related construct of SKOS is exploited to enrich the specification of an abstract term AT by formalizing the ontology relationships between AT and the other concepts of the LATO ontology composing its context. In particular, given a considered abstract term AT, related is used to connect AT to ancillary concepts of LATO representing i) an *objective judgment element* OBJ usually expressed through the connection of AT with a descriptive/statutory concept; ii) a *subjective quantitative evaluation* $SUBJ$ usually expressed through a relationship between AT and Quantity/UnitOfMeasure concepts; and iii) a *legislative reference* $LREF$ usually denoted with a connection of AT with a specific law or regulation (i.e., Law, LawArticle, and LawParagraph concepts). According to SKOS, each LATO concept is associated with a *preferred label* (prefLabel) as well as with one or more *alternative labels* (altLabel) and *hidden labels* (hiddenLabel) to enrich the concept definition with a label-set of literal descriptions that is very useful for subsequent knowledge extraction, to capture possible synonyms, acronyms, and abbreviations in the text of CLDs.

Example. An example of SKOS definition for the abstract term $AT =$ MinorOffense is shown in Figure 3(b) according to the Italian drug-trafficking law. MinorOffense is related to the concepts Drug and DrugTraffickingVerb that represent the OBJ relationships since they are subconcepts of StatutoryTerm and DescriptiveTerm, respectively. The relationships with the concepts Quantity and UnitOfMeasure represent the subjective judge evaluations $SUBJ$. The concepts Par5, Art73, and DPR309/1990 are subconcepts of the LawParagraph, LawArticle, and Law, respec-



(a) (b)

Fig. 3. (a) Excerpt of the LATO concept hierarchy; (b) Example of concept definition for the abstract term MinorOffense

tively, and they express the legal references $LREF$ of MinorOffense in the Italian criminal code where the drug trafficking crime is defined.

4 Knowledge extraction in CRIKE

Knowledge extraction in CRIKE is based on the idea to exploit text analysis techniques for detecting the concrete applications of legal abstract terms belonging to LATO throughout the stored/indexed case-law decisions CLDs. To this end, for a given abstract term AT, we introduce the notion of *abstract-term context* Ctx_{AT} containing, besides the AT term, all the concepts of LATO that are ancillary to AT, namely OBJ, $SUBJ$, or $LREF$ concepts:

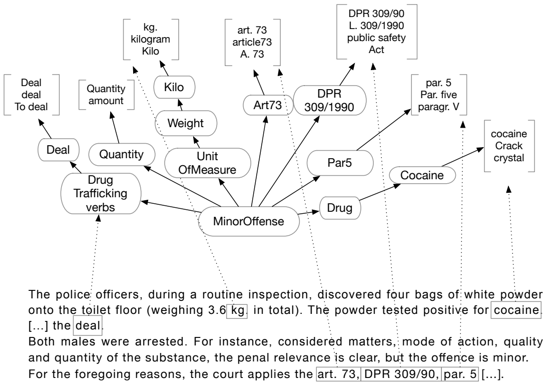
$$Ctx_{AT} = \{C_i \mid r(AT, C_i)\}$$

where $r(AT, C_i)$ denotes a SKOS related relationship between the abstract term AT and the concept C_i.

For each concept $C \in Ctx_{AT}$, we define the *concept label set* L_C that contains the whole set of labels, either preferred, alternative, or hidden, associated with C. Furthermore, based on the notion of L_C, we define the *extended label set* $\mathcal{L}_C$ where the concept label set of C is enriched by including the concept label set of the concepts C_j subsumed by C:

$$\mathcal{L}_C = L_C \cup \left\{L_{C_j} \mid C_j \subseteq C\right\}$$

Consider the goal to detect the concrete applications of a certain abstract term AT in a dataset of case-law decisions $CLDs$. CRIKE knowledge extraction is enforced by exploiting the extended label sets $\mathcal{L}_C$ of the concepts in the context Ctx_{AT}. For each document $d \in CLDs$, we define a vector representation $\boldsymbol{d}$ where each element corresponds to a concept in the context Ctx_{AT}. The value $d[i] \in \boldsymbol{d}$ is set to 1 when a *label hit* is detected, meaning that at least one occurrence of a label in $\mathcal{L}_{C_i}$ is found in d for the concept $C_i \in Ctx_{AT}$, and 0 otherwise (i.e., *label miss*). A threshold based mechanism is defined to specify the minimum number of label hits required to consider that a concrete application of the abstract term AT is detected in the document d.



The police officers, during a routine inspection, discovered four bags of white powder onto the toilet floor (weighing 3.6 kg. in total). The powder tested positive for cocaine. [...] the deal.
Both males were arrested. For instance, considered matters, mode of action, quality and quantity of the substance, the penal relevance is clear, but the offence is minor.
For the foregoing reasons, the court applies the art. 73, DPR 309/90, par. 5 [...].

Fig. 4. Example of knowledge extraction for the abstract term MinorOffense

Example. Consider the abstract term AT = MinorOffense and the corresponding context $Ctx_{\mathsf{MinorOffense}} = \{Drug, DrugTraffickingVerb, DPR309/1990, Art73, Par5, Quantity, UnitOfMeasure\}$. Moreover, consider the extended label set $\mathcal{L}_{Drug} = L_{Drug} \cup \{L_{Cocaine}, L_{Heroin}, L_{Cannabis}\}$. In Figure 4, we show an example of knowledge extraction based on the concepts and corresponding extended label sets in the context $Ctx_{\mathsf{MinorOffense}}$. An example of vector-based document representation for the abstract term AT = MinorOffense is shown in Figure 5. If we consider a threshold of 80% of label hits, we have that a concrete application of MinorOffense is detected in document d_1 since 6 hits are found over the available 7 concepts in $Ctx_{\mathsf{MinorOffense}}$.

	Drug	DrugTraffickingVerb	DPR309/90	Art73	Par5	Quantity	UnitOfMeasure
d_1	1	0	1	1	1	1	1
d_2	1	1	1	0	0	1	0

Fig. 5. Example of label hits for the abstract term MinorOffense

5 Related work

Work related to the issues addressed in CRIKE regards legal argumentation mining and legal ontology design. Legal argumentation mining refers to the capability to automatically detect and classify the role of possible argumentative units within a considered legal case text [1]. In [10], authors propose to mine statutory texts by using natural language processing and supervised machine learning techniques. More recently, the LUIMA approach has been proposed to focus on extraction of evidential reasoning from a court decision dataset [5]. Moreover, a particularly relevant contribution is provided in [9] about extraction of case law sentences for argumentation of statutory terms.

A survey on legal ontology design is presented in [1], where a special focus is given to representation of legal concepts in type systems. In [4], the notion of mutual consensus is introduced to support the specification of concepts and relations about contract formation. An application example based on a corpus of Italian legal texts is presented in [7], where the results of exploiting a learning system are provided. A further specification of a legal ontology using ONTOLINGUA is presented in [13]. Furthermore, in [12], authors present the LOIS project (Lexical Ontologies for Legal Information Sharing), and discuss a methodology for building a multilingual semantic lexicon for law able to be used both as a source of semantic metadata and as an external tool for cross lingual retrieval. On that topic, in [8], a methodology to automatically create an OWL ontology from a set of legal documents is presented. In [3], an automated approach based on statistical analysis is described, for identification of core concepts and relations in a corpus of legal texts. Natural Language Processing (NLP) techniques are proposed in [6], to extract concepts and relations among legal concepts, with the aim to build an ontology for legal information retrieval.

Original contribution of the proposed CRIKE approach is related to the enforcement of a data-science process with the support of an expert-based law ontology to extract knowledge from CLDs. A further peculiar feature of CRIKE is related to the formalization of an abstract term as a legal ontology concept with a corresponding context of related concepts. Ontology concepts with associated contexts are used to drive the identification of concrete applications/interpretations of corresponding abstract terms in the text of CLDs.

6 CRIKE support to practices and concluding remarks

In this paper, we presented the CRIKE approach for legal knowledge extraction. We envisage the following main practices that can be supported by using CRIKE i) **knowledge-assisted verdict writing**, where the concrete terminology extracted for abstract terms can support the judge in the preparation of new case-law decisions; ii) **history-based verdict prediction**, where the knowledge extracted by CRIKE is used to train a machine learning mechanism with the aim to predict the possible decision on a new incoming case-law to judge; and iii) **legal analytics**, where the results of knowledge extraction are exploited to detect possible trends and common abstract-term interpretations.

A preliminary experimentation of CRIKE has been performed based on a dataset provided by the Courthouse of Milan, Italy, whose results are described in [2]. The goal of the experimentation was to analyze the effectiveness of CRIKE in recognizing the concrete applications of the abstract term MinorOffense.

Different research directions are currently being investigated. On the one side, we are working on a bootstrapping approach to enforce enrichment of the LATO ontology, so that the context of abstract terms can be progressively augmented with new relevant terms and literals as long as they are detected in CLDs during extraction. On the other side, machine learning techniques are being developed to enforce a supervised classification of CLDs based on abstract terms, by exploiting a training set of CLDs manually annotated by domain experts.

References

1. Ashley, K.D.: Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press (2017)
2. Castano, S., Falduti, M., Ferrara, A., Montanelli, S.: Crime Knowledge Extraction: An Ontology-Driven Approach for Detecting Abstract Terms in Case Law Decisions. In: Proc. of the 17th Int. Conf. on Artificial Intelligence and Law (2019)
3. Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D.: Integrating a BottomUp and TopDown Methodology for Building Semantic Resources for the Multilingual Legal Domain, vol. 6036, pp. 95–121. Springer (2010)
4. Gardner, A.: An Artificial Intelligence Approach to Legal Reasoning. MIT Press, Cambridge, MA, USA (1987)
5. Grabmair, M., Ashley, K.D., Chen, R., Sureshkumar, P., Wang, C., Nyberg, E., Walker, V.R.: Introducing LUIMA: an Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools. In: Proc. of the 15th Int. Conference on Artificial Intelligence and Law. pp. 69–78. ACM (2015)
6. Lame, G.: Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations, pp. 169–184. Springer Berlin Heidelberg (2005)
7. Lenci, A., Montemagni, S., Pirrelli, V., Venturi, G.: NLP-based Ontology Learning from Legal Texts. A Case Study. In: Proc. of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques. pp. 113–129. Citeseer (2007)
8. Saias, J., Quaresma, P.: A Methodology to Create Legal Ontologies in a Logic Programming Information Retrieval System, pp. 185–200. Springer (2005)
9. Savelka, J., Ashley, K.D.: Extracting Case Law Sentences for Argumentation about the Meaning of Statutory Terms. In: Proc. of the 3rd Int. Workshop on Argument Mining. pp. 50–59 (2016)
10. Savelka, J., Grabmair, M., Ashley, K.D.: Mining Information from Statutory Texts in Multi-Jurisdictional Settings. In: Proc. of the Int. Conference on Legal Knowledge and Information Systems. pp. 133–142. IOS Press (2014)
11. The European Monitoring Centre for Drugs and Drugs Addiction: Italy, Country Drug Report 2018. Tech. rep., The European Monitoring Centre for Drugs and Drugs Addiction (2018)
12. Tiscornia, D.: The LOIS project: Lexical Ontologies for Legal Information Sharing. In: Proc. of the V Legislative XML Workshop. pp. 189–204 (2006)
13. Visser, P., Bench-Capon, T.: The Formal Specification of a Legal Ontology. In: Proc. of the Int. Conference on Legal Knowledge and Information Systems (1996)