

Principles of law: approaching a functional extraction

Marianna Molinari¹[0009–0003–1832–8135], Chiara Bonfanti²[0009–0007–8015–7786],
and Ilaria Angela Amantea²[0000–0003–1329–1858]

¹ Law Department, University of Turin, Torino, Italy

² Computer Science Department, University of Turin, Torino, Italy
`{marianna.molinari, chiara.bonfanti, ilariaangela.amantea}@unito.it`

Abstract. The need to step up the digitization degree of justice, through the use of advanced tools of knowledge, is increasingly felt. That is both for the exercise of jurisdiction and for the adoption of informed choices by judges, which are condensed into their judgments and in the principles of law established and renewed (i.e. the generalization of the interpretation and application of rules to a concrete case). The underlying exigency is not only that of enhancing the computerization of the trial, but rather that of making the most of the opportunities of new technologies in the organization of judicial work, also with the aim of imparting a predictive capacity to the demand for justice in compliance with the European indications. If the Office for Trial constitutes, in the intentions of the Italian legislator, an essential tool to dispose of the backlog currently existing in the Italian judicial offices; then artificial intelligence could reasonably facilitate those activities designed to reduce judicial pending. From this perspective, our attempt to approach a functional extraction of principles of law, that would not only be up-to-date, but could also be used by legal practitioners in post-project phases.

Keywords: Principles of Law · Natural Language Processing · Regular Expressions.

1 Introduction

The diversity of civil law and common law families is traditionally identified on the basis of the role attributed to the judge, who in Continental law systems is required to apply the law, while in Anglo-American law systems produces the law through the rule of *stare decisis et quia non movere*. Thus, in the Italian legal system, there is no codification of the binding nature of the precedent as such (*stare decisis*), not even if it originates from a ruling of the Joint Sections of the Supreme Court of Cassation. Nevertheless, it undeniably constitutes a trend value within the system, in force of which one should not deviate from a consolidated interpretation of the Judge of Legitimacy, institutionally invested with the function of nomophilacy, without strong and appreciable reasons. Therefore, despite in our procedural system, the precedent implies a pure logical relationship, which maintains the rulings in the rank of decisions on individual cases,

it represents something to be enhanced. That is because the use of precedents allows legal certainty, an equal treatment of similar cases, and an expectation about the future decision of the judge, too, above all in lawless cases.

Within this perspective, the role of living law emerges with particular incisiveness in the construction of the normative fabric, regulating the decision in the concrete case: thus the necessity to go deep inside the precedent, reaching its beating heart, namely the principle of law stated [4]. Furthermore, this is a data to be reached in the shortest possible time, since the principles of law, condensed in precedents, affect judgments motivation and so the timing of the decision-making phase of trials: in this sense, the importance of the reasonable duration of the proceedings must not be forgotten and so guaranteed by judges and Office for Trial employees.

The present paper aims at providing a possible feasible solution to functionally and automatically extract principles of law. The work reports the intermediate results experimented in the context of the Project NEXT GENERATION UPP - NGUPP³, which has the purpose to improve the efficiency of the judicial system in north-western Italy. To reach this goal, new collaborative schemes between universities and judicial offices have been tested to provide Office for Trial employees with transversal skills for the disposal of the backlog. The Office for Trial is an organizational structure made up of court assistants, operating in the judicial offices with the aim of ensuring the reasonable length of the proceedings. It is provided for in Article 16-*octies* of Decree-Law No. 179/2012, which firstly highlighted a link between technological innovation, organization and quality of justice. It has recently been revalued as a stable organizational structure, thanks to the Italian latest justice reform, above all to help out judges in the drafting of judicial measures, as a way to reduce the pending backlog and make the duration of the trials more reasonable. To pursue these goals, our research led us to the experimentation of a tool aimed at automatically extracting principles of law, stated by the Supreme Court of Cassation, when reported in first or second instance judgments. Starting from the annotation of domain experts, we extracted through a naive approach some of the patterns that identify the syntax of a principle of law to tackle the complexities of the task, giving a semantic connotation, and metric of comparison to the *corpus*. The work of extraction of the identified patterns has been developed using *ad hoc* Regular Expressions [1], then validated through comparison with similarities and performance valuing metrics.

In particular, in Section 2 we introduce a background of the topic according both to legal and computer science perspective; while in Section 3 we will show some first experiments and we will provide some results, comments and future works.

³ <https://www.nextgenerationupp.unito.it>.

It is a research promoted by the Italian Ministry of Justice as a part of the PON Governance and Institutional Capacity 2014-2020 - Axis I – Action 1.4.1.

2 Background

2.1 A juridical review

What is a principle of law? In the Italian context, certain principles are directly identified as such by the regulatory authorities themselves. Instead, other principles are such by virtue of an evaluation by the interpreters. This means that the interpreters, and in particular the judges, identify certain provisions as principles, lacking any direct rule by the legislator. Therefore, the latter principles are the result, not of the meaning ascription to specific normative texts, but of an integration of the law by the interpreters. They steam from legal interpretation: from single rules, from more or less vast sets of rules, or from the legal system as a whole [7]. Thus, in the Italian legal system, the principle of law does not constitute a source of law, nor does it represent the codification of a detailed rule accompanying the interpreted one, but rather the generalization of the interpretation and application of the rule to a concrete case. Through its formulation, the individual decision is brought under a general rule intended to be applied not only to same cases, but also to similar or comparable ones: it is obviously not an abstract principle, but a principle that governs the case, that is the law applied to the case. In other words, it is the rule of judgment in which the decision criterion of the concrete case is condensed, a criterion extracted from the norm and capable of constituting the link to the solution of other analogous cases. Therefore, it becomes a principle of law, constantly enucleable from the chain of subsequent pronouncements [4].

The principle of law has the function of universalizing individual decisions. This is the heart of the uniforming function of the Supreme Court of Cassation. The indications that in this sense come from its decisions, consequently, constitute principles of law to all intents and purposes, since the function of the Court of Legitimacy is precisely that of identifying the application of the rules in a general way. In fact, the Court starts from the specific hypothesis of the case at object to gradually climb the steps of a more general statement [8]. Otherwise, if we remained anchored to the single case, formulating an abstract principle solely referable to it, the rulings would never be destined to be applied with certainty in the future. The formulation of the principle of law, translated into an hermeneutic directive for interpreters, assures the predictability of the decisions, as well as the coherence of the system and a desired deflationary effect. It is, indeed, destined to operate in an indefinite number of trials, and even earlier in the extra-trial reality [11].

What is its role within a decision? A decision, while not binding on the judges who will have to deal with identical or similar cases, has weight and importance so that the new judge cannot fail to take it into account. In other words, when deciding, the judge must interpret and apply the provisions of the law to the specific case, keeping into due consideration whether these provisions have already been interpreted and applied by other judges to identical or similar cases, especially whether by the Court of Legitimacy. That's the "persuasive efficacy"

of the principle of law stated in a precedent. Not by chance, in the decisions motivation the recall of the compliant precedents is requested. Article 118 of the implementing provisions of the code of civil procedure, in establishing the rules for the motivation of the judgments, requires the exposition of "the relevant facts of the case and the reasons for the decision, also with reference to the compliant precedents", as well as the indication of the "rules of law and principles of law applied". Thus, the choice of the legislator is aimed at obtaining greater conciseness in the motivation of the judgments, but also at urging the judge to give an account of the position of the new decision with respect to the previous ones. Therefore, the coherent response of the judicial system in identical or similar cases [2]. These requirements primarily respond to legal certainty, on the basis of which every person must be able to evaluate and foresee the legal consequences of their conduct. Secondly, to the substantial equality of citizens before the law, which requires equal treatment in equal legal situations, that means ensuring the same treatment for the same cases. Both aspects are exacerbated, especially in lawless cases, namely cases in which jurisprudence, with its principles of law, is called to make up for, even in the absence of legislation.

Why is its automatic extraction useful with a view to optimizing trials? A higher specificity, in carrying over the principle of law, increases the possibility of giving life to a chain of subsequent application of it to an indefinite series of cases and situations. Given the above premises, it is possible to refer to principles of law, on the one hand, for the decisive ones, that are the principles around which the main question rotates. On the other hand, it is possible to refer to principles of law even for those that are strictly connected, or logically prejudicial to the former ones, as long as they can be independently isolated. The first ones are called "decisive" principles, since in their absence the main criterion could not be formulated or, in any case, would not assume any significant meaning. More specifically, they are *conditio sine qua non* with respect to the resolution of the devolved question. They pertain to the interpretation of the rules necessary to resolve the principal question. Quite the opposite, the argumentative sequences through which the *ratio decidendi* is articulated, or the *obiter dicta* possibly contained in the reasoning relating to the main question, or the formulations of general and abstract principles that completely go beyond the topic to which it refers, are not a principle of law of the question remitted. It should be noted that the identification of principles is not a solely general-theoretical or philosophical-juridical problem, devoid of practical outcomes. As mentioned, it crosses the need for precision, aimed at not stopping the application chain, together with a reduction in research times and in identifying the principle of law suited to the case at hand [7].

The Office for trial and the judgments drafting by its employees. The research and identification of the principles of law suited to the case at hand affects the drafting-motivation phase of rulings, that judges are called upon to write, as well as the Office for Trial employees. The «dual competence» attributed to the latter emerges from the list of their duties: a) on the one hand, they carry

out support for the judicial activity in the strict sense, together with the other members of the Office for Trial; b) on the other hand, they are included «fully among the ranks of administrative staff». The support activity to the judicial function, indicated under a), is expressed both on an organizational level and on a technical-legal level and is carried out under the supervision of the section president or another judge. In particular, on a technical-juridical level, the Office for Trial employees carry out study activities on the files (also preparing summary sheets for each proceedings, which can be used by the judge to illustrate the progress of the trial in the decision), that is to say, jurisprudential, doctrinal and regulatory reconstruction of context referable to the proposed cases. Above all, they can be used for drafting judgments⁴. Thus, the need for an implementation of motivation drafting methods that is in line, at the same time, with the parameters of the maximum economy imposed by the reasonable duration of trials and with the minimum constitutional law concerning the guarantees of a fair trial.

Possible support to the jurisprudential prediction. Predictive justice turns out to be a fascinating area, recalling the ability to exploit AI algorithms to predict the outcome of judgments. Certainly, the possibility of predicting a judicial measure depends mainly on the quantity and correctness of the informations available: the more informations are available, the higher the predictability degree of the outcome of a judgment will be. Jurisprudential precedents play a fundamental role in this sense: it is not a question of predicting the disposition of a sentence with punctual accuracy, but of identifying the orientation of the judge's reasoning. Indeed, the law was born to attribute certainty to human relationships, through a complex set of rights and duties. The theme of predictive justice takes its cue from this assumption and is based on a system whereby decisions are predicted thanks to the interpretation of the law and jurisprudential precedents. In short, the more a precedent, and its principles of law, is applied, then there will be a higher chance that in the future it will orient a new judgement.

2.2 Extraction

Why should computer science approach the extraction of principles of law? Whereas the extraction of principles of law can be a challenging topic, it can provide important results and contributions both to the legal and the computer science field. An important aspect is that the collection of the metadata represented by principles of law can give a metric of comparison, with a semantic connotation, that can provide aid both to tasks such as classification or similarity on different legal *corpora* [3]. Moreover, tracking the evolution of principles of law can add a new dimension to this newly found metric, time. Furthermore, the relationship and proprieties between entities, explicitated in the text found

⁴ MINISTRY OF JUSTICE, Circular 21th december 2021 - "Recruitment, tasks, training and working methods of the first 8,250 Office for Trial employees hired pursuant to decree-law n. 80 of 2021", www.giustizia.it.

in the motivations of a judgment, can be of interest as to model with more in depth the real world scenario represented by each judgment.

Information extraction for juridical texts. The task of information extraction typically presents itself as the automatic extraction of structured information from unstructured or semi-structured data sources [5]. Regex extraction is a technique for extracting specific patterns or sequences of characters from unstructured text data and it can be considered a technique apt to perform information extraction [17]. In literature, regarding legal *corpora* analysis, it has a prominent contribution the ontology based approach [15], or the broad usage of argumentation [6]. On the specific task of information extraction from legal corpora, there have been different approaches: it can be used pre-trained language models for legal information extraction [18] or named entity recognition [9] and linking techniques in legal domains [19]. Although there are, as well, some experiments and precedents on the usage of regular expressions to extract entities from judgments [12].

3 Firsts experiments

Working in a team composed of experts both of the computer science and the legal field, we conducted some first naive experiments. The matter of the extraction of principles of law from a *corpus* made of judgments, of first and second instance, has been approached using data coming from Turin Court, Labour Section. We worked on three different sets of data, all regarding the same judgments but annotated with a different methodology, hereafter explained in details.

- **First Annotation:** the jurists have been asked to generally detect and highlight what they identify as principles of law. It has been found that there are two main types of principles: the implicit and the explicit. Each of them is reported in the text of the sentence in a different way. The implicit principles are so stated in the jurisprudential panorama that they are normally just mentioned. Instead, the explicit principles are usually correlated to citations of rulings of the Italian Supreme Court of Cassation.
- **Second Annotation:** since the explicit principles resulted easier to detect and even more meaningful in semantics, the experiment has been focused only on these ones. In fact, they are usually written in the form of " *keyword* id of the judgment" as per shared practice (i.e. *cfr. Cass. no. 32500 of 2018 cit.*, whereas "cfr. Cass." is the keyword, and "32500" is the id of the judgment issued in 2018).
- **Citation set:** this latter set of data has been created *ad hoc* based on the simple pattern expressed in the previous annotation, removing all noise and redundancy typical of this domain. We left in the annotation merely the identification numbers of the citations to the Supreme Court rulings.

During these early stages of annotation, it was found that there is no uniform pattern in expressing principles of law. However, the pattern related to Supreme

Throughout our code it was used broadly the scikit-learn free software machine learning library [10]. In regard to the preprocessing stage that was implemented in the conducted experiments, it consisted firstly of cleaning the *corpus* from stopwords⁵ mainly found in Italian, then it was decided to proceed as well with the tokenization and removal of all the special characters not included in our regular expressions. Consequently, it was decided to keep the upper and lower case difference in our tokens, as it was a useful information to recognize the patterns we identified.

A total of five regexs were used, some of which were more tailored to single-case scenarios (like *JPtype2* and *JPtype3*). The third set, onto which the experiments were carried out, is the result of an approach of noise reduction, as there was more precision in results using the regex *JPtype1*.

```
#The most poignant of the five regex used in our experiments.

JPtype1=    "[Cc]?(?:ass\.|assazione|orte[Cc]ost|Sez|Suprema).
             {0,20}\n?.{1,20}\d+(?:[\\\.n,:e\\s\\w-;°]+\d+)+"

JPtype2=    "Corte.*sentenza n\.. (\d{4}\\.\d{2})"

JPtype3=    "[\"']{1}[ a-zA-Z0-9?><;,\\.]{}[\\\_\\_+=!@#$%^&*|'ÀÈÌÒÙ
             àèìòù]+[\"']{1} *[\"']{1}[ a-zA-Z0-9?><;,\\.]{}[\\_
             \\_\\_+=!@#$%^&*|'ÀÈÌÒÙàèìòù ]*[Cass\\.]{4}
             [a-zA-Z0-9?><;,\\.]{}[\\\_\\_+=!@#$ %\\^&*|'ÀÈÌÒÙàèìòù]+
             [\\)]{1}"
```

In Table 1 the results of the experiments, done on the data resulting from each of our annotation sets, are illustrated. As said, the findings do get better with each further annotation, peaking when it was chosen to focus only on the Supreme Court citations. The table represents the following values:

- ⁵ Stopwords library - <https://github.com/stopwords-iso/stopwords-it>

matches the data resulting from "*First Annotation*", the row "*second set*" is derived from the data of the "*Second Annotation*", whereas the row "*Citation set*" is defined by the data resulting from "*Second Annotation*" after a step of noise reduction.

- **Retrieved and Relevant set:** the regex pattern matching result has been stored in a list named *Retrieved*, representing a retrieved set of principles of law, whose cardinality is taken into consideration in our calculation. Meanwhile, the results of the annotations have been extracted from the .docx documents that the domain experts annotated. Subsequently, the results were stored in the *Relevant* list, whose cardinality similarly, as the aforementioned, was used into calculations of results and metrics.
- **Intersection and Threshold:** three approaches have been used to validate the presence of the *Retrieved* elements into the *Relevant* ones. Firstly, an as-is comparison method has been used to find perfect matches between the two lists. It didn't have relevant success in the first two annotations whereas the noise component in the *corpus* affected the comparison. Edit Distance⁶ has been used to furthermore maximize the cardinality of this *Intersection* set. Lastly, we generated vector embeddings through the libraries Torch⁷ and Transformers⁸. We processed them through Italian legal BERT⁹, and used them to calculate cosine similarity [14] to furthermore try to avoid the discrepancies between the *Retrieved* and *Intersection* set. It was decided to proceed to maximize the results whichever of the three similarities approaches had a better coverage of the intersection, regardless of the threshold onto which they were calculated.
- **Precision, Recall, F1-score:** of all the notations and sets that have been used in our experiments, *precision*, *recall* [16], and *f1-score* [13] have been calculated to further compare our findings. All of them, correlated to the threshold value, show that in the ideal dataset *citation set* there has been an improvement.

Table 1. Regex extractions and intersection based similarity results of Principles of Law (respectively "retrieved" and "intersection") and evaluation on the performances based on precision, recall, f1-score. The first and second sets in this table refer to the corresponding annotation made by domain experts.

dataset	retrieved	relevant	intersection	threshold	precision	recall	f1score
first set	7	22	6	10%	0.857	0.272	0.413
second set	7	14	5	10%	0.714	0.357	0.476
citations set	7	10	6	50%	0.857	0.6	0.705

⁶ Edit Distance - <https://www.nltk.org/api/nltk.metrics.distance.html>

⁷ Pytorch - <https://pytorch.org/>

⁸ Transformers- <https://huggingface.co/docs/transformers/index>

⁹ Italian legal BERT- <https://huggingface.co/dlicari/Italian-Legal-BERT>

It can be observed that Supreme Court related explicit citations (in *second set* and *third set*) can be captured by regex pattern matching with better results than the implicit ones. It is possible to validate more of the retrieved principles through a broader similarity coverage of the intersection between retrieved principles of law, and relevant ones, with a better threshold (50% compared to 10% of the *first set*). The recall and f1-score as well have better values in the *citations set* than the other sets.

3.3 Conclusions and Future Works

This paper illustrates what the principles of law are and their importance in the jurisprudential panorama, their weight in terms of legal certainty and substantive equality, and their impact on the drafting of judgments motivation and on the disposal of the backlog and on the optimization of bottlenecks. In this sense, some first experiments were conducted in order to obtain a computer based automatic extraction. This work may be used as foundations for further Natural Language Processing tasks like classification or similarity, as well as a baseline for this specific information extraction task. The next experiments regarding this information extraction task shall focus on the Supreme Court of Cassation quotes which in our experiments proved to be the most stable patterns to be recognized. During this interdisciplinary work with experts in different domains, emerged the great importance of tracking the semantic meaning and use of Supreme Court citations over time. In this sense, the changing of the principles of law over time may be valued, perhaps as a way to possibly support jurisprudential prediction (especially for the lawless cases) which in the Italian legal panorama often concern the rights of the individual.

Acknowledgements This research was conducted in the project "Next Generation UPP: New collaborative schemes between universities and judicial offices to improve the efficiency and performance of justice in North West Italy" in the context of PON Governance and Institutional Capacity 2014-2020.

References

- [1] Alfred V Aho. *Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity*. 1991.
- [2] Paolo Biavati et al. "Recensione a" La Cassazione civile. Lezioni dei magistrati della Corte suprema italiana", a cura di Maria Acierno, Pietro Curzio e Alberto Giusti". In: *QUESTIONE GIUSTIZIA* 27 (2020), pp. 1–3.
- [3] C. Bonfanti et al. "A pipeline for data management, knowledge extraction and semantic analysis of unstructured legal judgments". In: (*In press*) *Proceedings of Conference Ital-IA 2023*. 2023.
- [4] Gaetano De Amicis. "La formulazione del principio di diritto ei rapporti tra Sezioni semplici e Sezioni Unite penali della Corte di Cassazione". In: *Dir. pen. cont* 4 (2019).

- [5] Ralph Grishman. “Information extraction”. In: *Communications of the ACM* 40.8 (1997), pp. 80–91.
- [6] Giulia Grundler et al. “Detecting Arguments in CJEU Decisions on Fiscal State Aid”. In: *Proceedings of the 9th Workshop on Argument Mining*. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Oct. 2022, pp. 143–157. URL: <https://aclanthology.org/2022.argmining-1.14>.
- [7] Riccardo Guastini. “Principi di diritto e discrezionalità giudiziale”. In: *Diritto pubblico* 3 (1998), pp. 641–660.
- [8] Riccardo Guastini. “L’interpretazione dei documenti normativi”. In: *Giuffrè* (2004), 131 ss.
- [9] Bing Liu and Ian Zhang. “A survey of named entity recognition and classification”. In: *Journal of Zhejiang University-Science C (Computers & Electronics)* 20.1 (2019), pp. 1–23.
- [10] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [11] Andrea di Porto. *Tecniche di massimazione delle sentenze: con Prefazione di Andrea Di Porto*. Vol. 16. Sapienza Università Editrice, 2017.
- [12] Mariano Rico et al. “Extracting terminologies in the legal domain: a syntactic pattern-based approach for Spanish”. In: *Iberlegal workshop at JURIX conference*. 2019.
- [13] CV Rijsbergen. “Information retrieval 2nd ed Buttersworth”. In: *London [Google Scholar]* (1979), p. 115.
- [14] Gerard Salton and Michael J McGill. “Introduction to modern information retrieval”. In: (1986).
- [15] Anu Thomas and Sivanesan Sangeetha. “Semi-supervised, knowledge-integrated pattern learning approach for fact extraction from judicial text”. In: *Expert Systems* 38.3 (2021), e12656.
- [16] Kai Ming Ting. “Precision and Recall”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 781–781. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_652. URL: https://doi.org/10.1007/978-0-387-30164-8_652.
- [17] Jingjing Wang, Xiaohui Liu, and Yuzhong Guo. “A regex-based approach for entity extraction from scientific articles”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2277–2286.
- [18] Xiao Wang et al. “Exploring pre-trained language models for legal information extraction”. In: *Journal of Information Science* 45.6 (2019), pp. 817–836.
- [19] Bilge Yavuz, Oana Inel, and Martin Riedl. “Named Entity Recognition and Linking in Legal Domains”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 5869–5877.