Automatic inference of taxonomy relationships among legal documents

Irene Benedetto[1,2][0000−0001−7086−7898], Luca Cagliero[1][0000−0002−7185−5247], and Francesco Tarasconi[2]

[1] Politecnico di Torino, 10129 Torino, Italy
[2] H-Farm Innovation, 10121 Torino, Italy

Abstract. Exploring legal documents such as laws, judgments, and contracts is known to be a time-consuming task. To support domain experts in efficiently browsing their contents, legal documents in electronic form are commonly enriched with semantic annotations. They consist of a list of headwords indicating the main topics. Annotations are commonly organized in taxonomies, which comprise both a set of is-a hierarchies, expressing parent/child-sibling relationships, and more arbitrary related-to semantic links. This paper addresses the use of Deep Learning-based Natural Language Processing techniques to automatically extract unknown taxonomy relationships between pairs of legal documents. Exploring the document content is particularly useful for automatically classifying legal document pairs when topic-level relationships are partly out-of-date or missing, which is quite common for related-to links. The experimental results, collected on a real heterogeneous collection of Italian legal documents, show that word-level vector representations of text are particularly effective in leveraging the presence of domain-specific terms for classification and overcome the limitations of contextualized embeddings when there is a lack of annotated data.

Keywords: Legal judgments annotation · Legal text modeling · Natural language processing · deep learning.

1 Introduction

Legal databases contain a vast and heterogeneous collection of documents of different types among which legislative documents, regulations, judgments and maxims [7]. To support legal experts, such as lawyers and judges, in the navigation of these electronic data sources it is necessary to design efficient and effective retrieval systems. However, the inherent complexity of the legal terminology, the increasing number of resources available in electronic form, and the variability of the data sources across different countries make the problem of efficiently retrieving and exploring legal documents particularly challenging [27].

Content browsing and retrieval in legal databases is commonly driven by human-generated annotations organized in domain-specific taxonomies (e.g., [7]).

A legal taxonomy consists of a set of is-a hierarchies describing the parent/child-sibling relationships between topical headwords. Each legal document is potentially annotated with many topical headwords. For example, according to the taxonomy depicted in Figure 1 documents ranging over *Rights of exploitation* also belong to the parent category *Collective exploitation of property*, which, in turn, includes the documents ranging over the *Public domain* as well. Based on their domain knowledge, legal experts also exploit arbitrary *Related-to* taxonomy relationships such as those linking *Public domain* to *Immovable property* categories.

Domain experts can leverage taxonomy relationships to browse legal documents, jumping from one to another according to their specific needs. For example, exploring two complementary documents linked by a related-to relationship can be deemed as useful for covering different aspects of the same topic. However, annotation-based legal content retrieval is hindered by the following issues:

- Taxonomies evolve with legal systems [18] according to a *temporal concept drift*, which is typical of legal topic classification [7]. Therefore, existing taxonomy-based document relationships may become unreliable.
- Taxonomy relationships are often incomplete. Specifically, the *Related-to* relationships are unlikely to be available in several legal domains.
- Most electronic documents and the related annotations are written in English. Hence, there is a lack of solutions tailored to languages other than English.

To overcome the above issues, this paper proposes a classification system, based on Deep Natural Language Processing techniques, to automatically infer the taxonomy relationships holding between pairs of legal documents. Starting from a proprietary dataset collecting Italian legal judgments, mainly in the domain of private property, it learns supervised machine learning techniques to automatically predict the type of relationship holding between the document pair, i.e., *Parent-of/Child-of*, *Sibling-of*, or *Related-to*. The automatic annotation of document pairs is instrumental for improving the efficiency and effectiveness of legal content retrieval, especially when the taxonomy content is incomplete or partly out-of-date.

The preliminary results acquired in a real use case show that 1) traditional word-level vector representations of text, such as Doc2Vec [15], are particularly effective in the considered domain as capture syntactic properties between domain terms, such as "Patents" and "Trademarks", that are useful for classification purposes. 2) Contextualized embeddings [8] achieve performance than Doc2Vec on Italian documents due to the lack of domain-specific training data.

The rest of the paper is organized as follows: Section 2 overviews the prior works. Section 3 describes the classification system. Section 4 summarizes the main experimental results, whereas Section 5 draws the conclusions of the present work.

2 Related work

In recent years the interest in legal AI has constantly grown. It encompasses legal judgment prediction [5, 29], entity recognition [1, 22], legal document classification [2, 6, 7], legal question answering [11, 13], and legal summarization [12, 28]. To analyze legal documents' similarities prior works adopt rule-based approaches [9, 24], document-level text similarities [17], graph-based methods [25] and machine learning-based solutions [14, 21]. All of the aforesaid works focus their analyses on English-written documents. Conversely, this work applies machine learning-based strategies to analyze Italian document relationships. To the best of our knowledge, the only attempts to perform a similarity analysis between multilingual legal documents have been presented in [19, 20], where the authors analyze a multilingual corpus of 43 directives. Unlike [19, 20], this work focuses on inferring taxonomy relationships among topical headwords.

3 The classification system

3.1 Preliminaries

Let d_i be an arbitrary legal document, $\mathcal{L}$ be the set of labels present in a legal taxonomy $\mathcal{T}$, and $L_i \subseteq \mathcal{L}$ be the subset of labels used to annotate d_i.

The taxonomy $\mathcal{T}$ is a set of is-a hierarchies built over labels in $\mathcal{L}$. Each label consists of a set of headwords describing a document topic. A *Is-a* taxonomy relationship indicates the specular *Child-of* and *Parent-of* relationships holding between pairs of annotations $l_i, l_j \in \mathcal{L}$. Instead, the *Sibling-of* relationship holding between a pair of annotations indicates that l_i and l_j have a parent in common. Furthermore, an arbitrary *Related-to* relationship links a pair of label l_i and l_j that are semantically related one to another.

3.2 Problem statement

Given an arbitrary pair of annotated documents (d_i, d_j), we aim at inferring the pairwise taxonomy relationships between labels in L_i and L_j (and vice versa).

More specifically, given (d_i, d_j), the purpose is to define a classification function f that predicts the type of the relationships holding between labels in L_i and labels in L_j. Thus, the target of the prediction is the relationship type, which takes values *Sibling-of*, *Parent-of/Child-of*, *Related-to* or *Other* (if a relationship either does not hold or is not relevant to the domain under analysis). For the sake of simplicity, hereafter we will cast the problem under analysis to a single-label task, i.e., the relationship cannot belong to multiple types at the same time.

f is computed as the composition of two functions $(h \circ g)(d_i, d_j) = h(g(d_i, d_j))$. Specifically, $g(d_i, d_j)$ produces a high-dimensional vector representation $e_i, e_j \in \mathbb{R}^N$ of the given documents. On top of the generated text representations, we train a classification model to estimate the function $h(e_i, e_j)$ and detect the

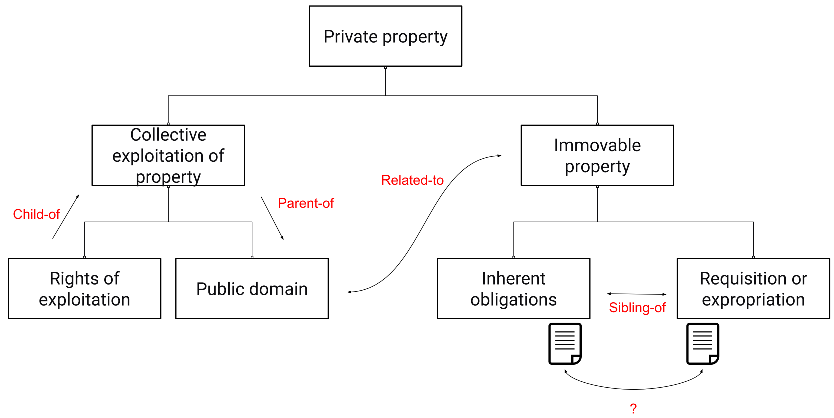

Fig. 1: Example of the taxonomy: in red we highlighted the relationship types we aim at identifying with our classification system.

corresponding relationship type r_{l_i,l_j}. At inference time, we predict the relationship type based on the document content, without any prior knowledge on the existing document annotations l_i and l_j.

3.3 Text representations

We consider the following established text representations of the input legal documents:

- Term Frequency-Inverse Document Frequency (TF-IDF, in short) [26]: a occurrence-based, word-level text representation.
- Doc2Vec [15]: a sentence-level embedding model based on a Word2Vec extension.
- Multilingual BERT [8]: a contextualized embedding model[3].

3.4 Classifiers

We integrate and test the following established classification approaches available in the scikit-learn library [23][4]: Support Vector Machines (SVMs), K-Nearest Neighbor (k-NN), Random Forest classifier (RF), and Logistic Regression (LR) [10].

[3] Due to the lack of a sufficient amount of domain-specific data in the Italian language, BERT cannot be retrained from scratch on the input data collection.

[4] Neural Network models are not suited to the prediction task under analysis due to the lack of a sufficient amount of training data.

4 Experimental results

4.1 Experimental design

Dataset The proprietary dataset used in the empirical validation contains Italian legal judgments and maxims, mainly in the context of property law. Each legal document is annotated with one or more labels, corresponding to the legal principle of reference. Pairs of these legal principles are partitioned into groups according to the type of relationship between them (i.e., *Sibling-of*, *Parent-of/Child-of*, *Related-to* or *Other*).

The training dataset consists of a set of triples (d_i, d_j, r_{l_i,l_j}). Each triple represents a given pair of documents (d_i, d_j) annotated with the corresponding legal topics' relationship.

The complete dataset consists of more than 1300 examples for each relationship type and the relationship types are roughly equally distributed. For testing purposes, we apply a holdout strategy (80% train, 10% validation, 10% test).

Metrics To evaluate classifiers' performance we compute the Precision, Recall and F1-score scores [10]. The goal is to evaluate the ability of the classification system to correctly identify positive examples separately for each class. Their definitions follow.

- *Precision (Pr) of class* c: it indicates the fraction of document pairs correctly classified as c among all the pairs labeled as c.
- *Recall (Rc) of class* c: it indicates the fraction of document pairs classified as c that have been retrieved over the total number of pairs labeled as c.
- *F1-score (F1) of class* c: it is the harmonic mean of precision and recall of class c.

To compute the similarity between a pair of legal documents in the vector space we use the established cosine similarity [10]. Specifically, let $e_i = f(d_i)$ and $e_j = f(d_j)$ be the encodings of d_i and d_j, respectively. The document similarity is defined by

$$sim_{d_i,d_j} = sim(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\|\|e_j\|} \tag{1}$$

4.2 Results discussion

Analysis of the text representations We compute the pairwise similarity between each document pair, group the results by relationship type, and test the difference in mean between the per-type similarity values' distributions. The main goal is to understand whether the text encoding phase is able to clearly separate the groups related to different relationship types.

To verify the initial hypothesis of having a clear separation among relationship types for each pair we compute the two-sided t-test [16] for the difference between their means (with a type I error of 0.05). The outcomes of the significance

Table 1: Significance test for the difference of two means computed on documents' similarities groups. α=0.05.

Method	Type 1	Type 2	P-value
BERT	Sibling-of	Parent-of/Child-of	0.073
BERT	Related-to	Parent-of/Child-of	<0.001
BERT	Related-to	Sibling-of	<0.001
BERT	Other	Parent-of/Child-of	0.848
BERT	Other	Sibling-of	0.047
BERT	Other	Related-to	<0.001
Doc2Vec	Sibling-of	Parent-of/Child-of	0.308
Doc2Vec	Related-to	Parent-of/Child-of	<0.001
Doc2Vec	Related-to	Sibling-of	<0.001
Doc2Vec	Other	Parent-of/Child-of	<0.001
Doc2Vec	Other	Sibling-of	<0.001
Doc2Vec	Other	Related-to	<0.001
TFIDF	Sibling-of	Parent-of/Child-of	0.839
TFIDF	Related-to	Parent-of/Child-of	0.889
TFIDF	Related-to	Sibling-of	0.746
TFIDF	Other	Parent-of/Child-of	<0.001
TFIDF	Other	Sibling-of	<0.001
TFIDF	Other	Related-to	<0.001

Table 2: Classification results

Representation	Classifier	Precision	Recall	F1-Score
Doc2vec	Logistic Regression	0.740	0.744	0.740
Doc2vec	SVM	0.731	0.735	0.733
Doc2vec	Random Forest	0.721	0.724	0.722
TF-IDF	Random Forest	0.702	0.706	0.703
TF-IDF	SVM	0.690	0.695	0.685
TF-IDF	Logistic Regression	0.666	0.669	0.667
Doc2vec	KNN	0.660	0.664	0.662
BERT	Random Forest	0.650	0.660	0.648
BERT	SVM	0.639	0.647	0.642
BERT	Logistic Regression	0.610	0.620	0.614
TF-IDF	KNN	0.583	0.591	0.584
BERT	KNN	0.429	0.460	0.433

tests are summarized in Table 1. As expected, the differences in mean between *Parent-of/Child-of* and *Sibling-of* are never significant. Among all the candidate representations, Doc2Vec achieves the best performance. Unlike Doc2Vec, contextualized embedding models do not achieve performance superior to the others mainly due to the lack of in-domain training data.

Analysis of classifiers-performance In Table 2 we report the values of the classifier performance metrics achieved on the test set. The joint use of the Doc2Vec

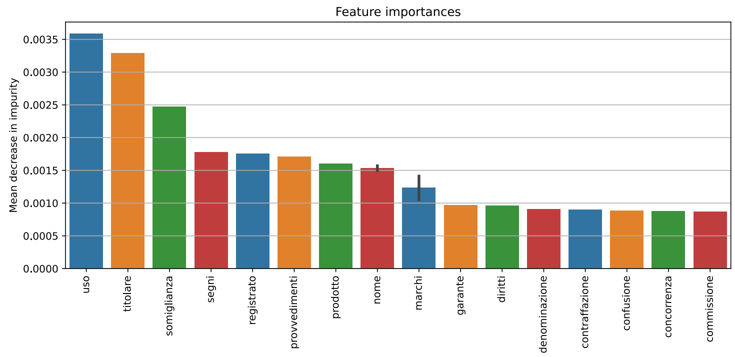

Fig. 2: Feature importance of Random forest classifier. This barplot shows that the most relevant terms for the classifier are mainly in-domain terms, such as “trademark” ("marchio"), “commission” (“commissioni”).

text representation and of the Logistic Regression classifier has shown to be highly beneficial, probably due to the inherent characteristics of the input data.

Model explainability To gain insights into the classification problem we leverage the characteristics of the decision tree models, which allow us to evaluate the influence of the input features on the output prediction.

Figure 2 shows that the headwords related to the legal domain, such as “Patents”, “Trademarks”, are actually very discriminating. This property is particularly helpful for modelling the input data, as less pertinent features can be early pruned.

Predictability of different relationship types Table 3 reports the per-class precision, recall and F1-score results achieved on the test set. The relationships of type *Related-ot* appear to be simpler to predict, whereas *Parent-of/Child-of* and *Sibling-of* turn out to be particularly challenging (see also the confusion matrix plot in Table 3.

Table 3: Predictability of different relationship types

	Parent-of/Child-of	Sibling-of	Related-to	Other
Precision	0.658	0.675	0.905	0.663
Recall	0.546	0.738	0.937	0.690
F1-Score	0.60	0.706	0.920	0.672

5 Takeaways and future directions

The present work focused on overcoming the main issues of annotation-based legal content retrieval systems by proposing a classification-based to document pair annotation. The main takeaways can be summarized below.

- *Concept drift*: The updates of the original document collection and the presence of a relevant drift in the covered topics triggers the periodic retraining of the entire classification model. Such an activity can be labour-intensive and time-consuming. To overcome the aforesaid issue, we recommend to first explore the graphical distributions of the pairwise vector similarities and set up the classification pipeline accordingly.
- *Missing Related-to relationships*: *Related-to* is the most unconventional relationship in legal taxonomies. Since it is not easy to map it to known concepts or to existing data structures its correct prediction is particularly challenging. The proposed classification system achieved 90% F1-score on class *Related-to* thus confirming its effectiveness and usability in real application contexts (see Table 3).
- *Portability to languages other then English*: The increasing availability of state-of-the-art multilingual pretrained models (e.g., MultiLingual BERT [8]) and the recent advances in cross-lingual approaches [3, 4] allow the direct processing of the raw data without the need to perform automatic machine translation. The preliminary results achieved on Italian legal documents confirm the feasibility of the multilingual extension.

As future works, we envision (1) The application of the proposed method to generate annotations in complex scenarios, e.g., zero-shot classification and active learning. (2) The application of eXplainable AI methods to increase model transparency. (3) The organization of a crowdsourcing validation process, based on domain experts, to study the applicability of the system in real scenarios (e.g., law firms, courthouse rooms). (4) The application of the proposed classification system to legal documents written in low-resource languages (e.g., Hindi, Vietnamese, Zulu).

Acknowledgements

The research leading to these results has been partly supported by the SmartData@PoliTO center for Big Data and Machine Learning technologies. The dataset has been provided by Giuffrè Francis Lefebvre S.p.A..

References

1. Angelidis, I., Chalkidis, I., Koubarakis, M.: Named entity recognition, linking and generation for greek legislation. In: JURIX (2018)

2. Luz de Araujo, P.H., de Campos, T.E., Ataides Braz, F., Correia da Silva, N.: VICTOR: a dataset for Brazilian legal documents classification. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1449–1458. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.181
3. Cagliero, L., Quatra, M.L.: Inferring multilingual domain-specific word embeddings from large document corpora. IEEE Access **9**, 137309–137321 (2021). https://doi.org/10.1109/ACCESS.2021.3118093
4. Cagliero, L., Quatra, M.L., Garza, P., Baralis, E.: Cross-lingual timeline summarization. In: Fourth IEEE International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2021, Laguna Hills, CA, USA, December 1-3, 2021. pp. 45–53. IEEE (2021). https://doi.org/10.1109/AIKE52691.2021.00014, https://doi.org/10.1109/AIKE52691.2021.00014
5. Chalkidis, I., Androutsopoulos, I., Aletras, N.: Neural legal judgment prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4317–4323. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1424, https://aclanthology.org/P19-1424
6. 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. pp. 6314–6322. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1636, https://aclanthology.org/P19-1636
7. Chalkidis, I., Fergadiotis, M., Androutsopoulos, I.: Multieurlex - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: EMNLP (2021)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2019)
9. Geist, A.: Using citation analysis techniques for computer-assisted legal research in continental jurisdictions. SSRN Electronic Journal (05 2009). https://doi.org/10.2139/ssrn.1397674
10. Han, J., Kamber, M.: Data Mining. Concepts and Techniques. Morgan Kaufmann, 2nd ed. edn. (2006)
11. Hendrycks, D., Burns, C., Chen, A., Ball, S.: CUAD: an expert-annotated NLP dataset for legal contract review. CoRR **abs/2103.06268** (2021), https://arxiv.org/abs/2103.06268
12. Kanapala, A., Pal, S., Pamula, R.: Text summarization from legal documents: A survey. Artif. Intell. Rev. **51**(3), 371–402 (mar 2019). https://doi.org/10.1007/s10462-017-9566-2, https://doi.org/10.1007/s10462-017-9566-2
13. Kim, M.Y., Xu, Y., Goebel, R., Satoh, K.: Answering yes/no questions in legal bar exams. p. 199–213. Springer-Verlag, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-319-10061-6_14
14. Landthaler, J., Waltl, B., Holl, P., Matthes, F.: Extending full text search for legal document collections using word embeddings. In: JURIX (2016)
15. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR **abs/1405.4053** (2014), http://arxiv.org/abs/1405.4053
16. Limentani, G.B., Ringo, M.C., Ye, F., Bergquist, M.L., McSorley, E.O.: Beyond the t-test: statistical equivalence testing (2005)

17. Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., Ghosh, S.: Measuring similarity among legal court case documents. In: Proceedings of the 10th Annual ACM India Compute Conference. p. 1–9. Compute '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3140107.3140119, https://doi.org/10.1145/3140107.3140119
18. Mattei, U.: Three Patterns of Law: Taxonomy and Change in the World's Legal Systems. The American Journal of Comparative Law **45**(1), 5–44 (01 1997). https://doi.org/10.2307/840958, https://doi.org/10.2307/840958
19. Nanda, R., Caro, L.D., Boella, G.: A text similarity approach for automated transposition detection of european union directives. In: JURIX (2016)
20. Nanda, R., Siragusa, G., Di Caro, L., Boella, G., Grossio, L., Gerbaudo, M., Costamagna, F.: Unsupervised and supervised text similarity systems for automated identification of national implementing measures of european directives. Artificial Intelligence and Law **27** (06 2019). https://doi.org/10.1007/s10506-018-9236-y
21. Ostendorff, M., Ash, E., Ruas, T., Gipp, B., Moreno-Schneider, J., Rehm, G.: Evaluating Document Representations for Content-Based Legal Literature Recommendations, p. 109–118. Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3462757.3466073
22. Papaloukas, C., Chalkidis, I., Athinaios, K., Pantazi, D., Koubarakis, M.: Multi-granular legal topic classification on greek legislation. CoRR **abs/2109.15298** (2021), https://arxiv.org/abs/2109.15298
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
24. Raghav, K., Reddy, K., Reddy, V.B.: Analyzing the extraction of relevant legal judgments using paragraph-level and citation information (2016)
25. RS, W., D., A.: Legal document similarity: a multi-criteria decision-making perspective. PeerJ Computer Science **6:e262** (2020), https://doi.org/10.7717/peerj-cs.262
26. Sammut, C., Webb, G.I. (eds.): TF–IDF, pp. 986–987. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_832, https://doi.org/10.1007/978-0-387-30164-8_832
27. Van Opijnen, M., Santos, C.: On the concept of relevance in legal information retrieval. Artif. Intell. Law **25**(1), 65–87 (mar 2017). https://doi.org/10.1007/s10506-017-9195-8, https://doi.org/10.1007/s10506-017-9195-8
28. Wu, Y., Kuang, K., Zhang, Y., Liu, X., Sun, C., Xiao, J., Zhuang, Y., Si, L., Wu, F.: De-biased court's view generation with causality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 763–780. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.56, https://aclanthology.org/2020.emnlp-main.56
29. Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., Zhao, J.: Distinguish confusing law articles for legal judgment prediction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3086–3095. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.280, https://aclanthology.org/2020.acl-main.280