A Service Infrastructure for the Italian Digital Justice

A Service Infrastructure for the Italian Digital Justice

Valerio Bellandi[1][0000−0003−4473−6258], Silvana Castano[1][0000−0002−3826−2407], Stefano Montanelli[1][0000−0002−6594−6644], Davide Riva[1][0009−0003−9681−9423], and Stefano Siccardi[1][0000−0002−6477−3876]

Università degli Studi di Milano
DI - Via Celoria, 18 - 20135 Milano
{name.surname}@unimi.it

Abstract. The management of legal documents, especially court judgments, can be a daunting task due to the vast amounts of data involved. Traditional methods of managing legal documents are no longer sufficient, as the volume of data continues to increase, leading to the need for more advanced and efficient systems. The proposed infrastructure seeks to address this challenge by organizing a repository of textual documents and annotating them in a way that facilitates various downstream tasks. The framework is designed to be developed and maintained in a sustainable way, ensuring multiple services and uses of the annotated document repository while considering the limited availability of annotated data. This approach ensures that the output of the annotation algorithms aligns with the organizational processes used in Italian courts. The experiments conducted to demonstrate the feasibility of the solution employed different low-resource methods and solutions designed to combine these approaches in a meaningful way.

Keywords: Legal Document Annotation · Named Entity Recognition · Concept Extraction · Zero-Shot Learning

1 Introduction

In this paper, we present an infrastructure designed to facilitate the management of legal documents such as court rulings and orders. The primary objective of this infrastructure is to allow for flexible use, meaning that it can accommodate any service that may be required throughout the project's lifetime, without prescribing specific types of information management in advance. Legal documents are valuable sources of information for various purposes and stakeholders. For instance, judges may need to find court decisions in similar cases or lawsuits involving the same individuals or entities. The legal community may require statistics on general trends in areas such as maintenance granted and the economic conditions of partners in divorces. Justice department officials may need to assess court performance in terms of the time taken to complete lawsuits, the number of sentences confirmed or changed in appellate judgments, and other

metrics[26]. The infrastructure we propose aims to provide solutions to these various needs by allowing for the integration of additional services as required. This flexibility enables the infrastructure to accommodate different stakeholders and their varying needs. The ability to integrate additional services ensures that the infrastructure can evolve over time and continue to meet the needs of its users. The infrastructure design achieves several goals, such as storing text documents and relevant metadata, conducting standard searches on text and metadata by utilizing logical operators to locate occurrences of strings or numbers. The infrastructure includes modules and services to recognize entities occurring within documents and to classify them by using entity types taken from external, reference ontologies. It has the capability to disambiguate the recognized entities and to search for their occurrences in the documents, locate document sections, cluster documents, and perform advanced statistical analyses. Furthermore, it can enforce knowledge extraction based on a combination of *context-aware embedding* models and *zero-shot classification* techniques. The goal is to mine a concept network extracted from documents without relying on any external knowledge base. The network can be exploited to provide concept-driven services to users, like citizens and legal actors (e.g., layers, practitioners), interested in searching, exploring, and analyzing the underlying corpus of legal documents ingested by the system. The design must also prioritize stability and scalability, which are essential for ensuring that the system can handle large volumes of data and continue to deliver high-quality services[22]. By incorporating these features into the infrastructure design, we can ensure that it meets the diverse needs of its users and can evolve over time to keep up with changing requirements. The paper also provides two application examples of the proposed architecture to concrete case-studies in the framework of the Italian digital justice. Evaluation results are finally discussed to show the feasibility of the proposed solution in real situations. In conclusion, the main contribution of this work is the description of a semantically enriched document management system, to support daily operations of multiple users. Judges, clerks of the courts and of the central Justice offices, statisticians, to mention a few, can access the system to search for information suited to their specific needs.

2 Related Work

Several infrastructure have been described, to support document management in the legal domain, focusing on one or more associated tasks. For instance, a software architecture with a specific pipeline to extract knowledge from documents is described in [1]. The systems aims at assisting lawyers in resolving legal cases by automatically extracting key information and suggesting potential arguments. It uses a combination of rule based and statistical NLP techniques. [2] describes a document management system that combines NLP and ontologies in the legal domain. An architecture that analyzes and extracts a set of specific data from texts in natural language has been described in [4] and later improved in [5], implementing an architecture based on microservices and message brokers.

It is based on an ontology containing information about the types of documents, their properties and sections, entities from each section and how their relations. From the point of view of the semantic text analisys several approaches propose NLP techniques to support knowledge extraction and integration in the broad legal domain. An overview can be found in [6], which emphasizes the roles of Named Entity Recognition (NER) techniques and Relation Extraction (RE). Several papers are related to the legal case retrieval task that consists in reading a new case Q, and extracting supporting cases S_1, S_2, ... S_n for the decision of Q. See for instance the Competition on Legal Information Extraction/Entailment (COLIEE) organized since 2017 [7] and the Artificial Intelligence for Legal Assistance (AILA) shared task [21]. For the present work, the Relation Extraction (RE) task is particularly relevant. In general, it is considered quite challenging, especially when arbitrary relations are of interest, see [8], sometimes, joint entity and relation extraction has been proposed. For instance [9] uses a pre-trained BERT model to find and classify text spans and relations given a set of predefined relation classes. For concept extraction, the lack of annotated data poses an additional issue to consider for the application of supervised techniques. Zero-shot classification (ZSC) techniques have been proposed as a solution to such an issue (e.g., [15]). The use of fine-tuned embedding models has been also proposed for legal documents in the English language [18], as well as in the Italian language [20]. Our proposed solution to knowledge extraction exploits ZSC techniques to enforce classification on unlabeled data instances without annotation. The proposed solution is also characterized by the use of a pre-trained model without any fine-tuning for a semantically-meaningful document representation rather than a token representation, by relying on a contextual, transformer-based embedding models (i.e., Sentence-BERT[16]).

3 Architecture

In this section the main features of the proposed infrastructure will be described in detail. In particular, Figure 1 illustrates the layers and their components.

The layer deputed to storage and elaborate the data are composed by i) **Ingested documents** texts and their metadata stored in a document database. It is the repository of the raw data coming from the legacy system; ii) Database to store **annotations** created by the system and an **index** system for full text, metadata and annotation search; iii) Graph database representing the **Entity Registry**, it is deputed to store a unique entry for every entity found in the documents. The ER exposes APIs to manage both the entity types (the ER metamodel) and entity instances as described in [25]; iv) **System logs** and other data are stored to monitor the system. v) a dynamic set of **NLP Services**, each performing specific analyses, to create annotations, or other operations: Named Entity Recognition, Linking and Concept Extraction; vi) a **Service Catalog** records available services, both for data analysis and to create new versions of the texts (e.g. data cleaning, data pre-processing, summarizing etc.). The

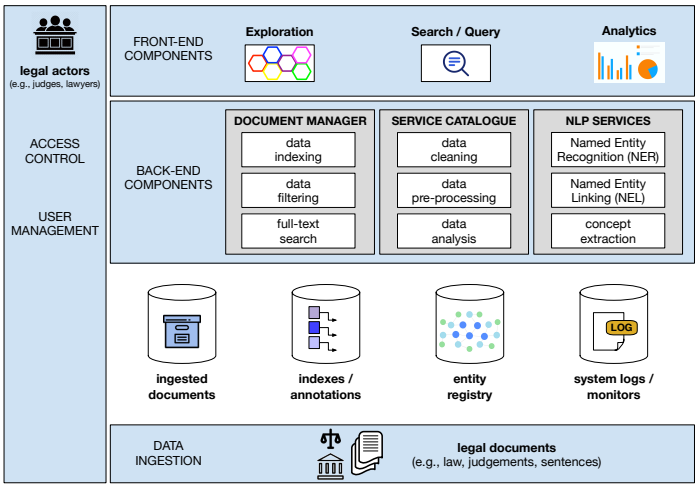

Fig. 1. Main Modules System Infrastructure.

component performs some service orchestration functions to manage workflows of services.

The components in the **Document Manager** subsystem, with their APIs, can be considered proxies for client programs to index, filter and fetch data stored in the system and to update them. Regarding the interaction with the users we extended our previous work [19] and this proposal provides functionalities that permits i) to *explore* documents: after reading a document the user wants to move to other documents sharing the same entities or concepts. ii) to *query* documents: the user wants to find documents containing specific entities or concepts, like in a search engine and iii) to *analyse* the document set, for instance number of entities or concepts in a document, and averages, or number of document for an entity or concept and averages; shortest path (through documents) between concepts or entities; entities and concepts centrality, and so on. In this instance, texts and metadata are stored in an ElasticSearch[11] instance and annotations in a SQL database as described in our previous work [19]. Moreover a communication queue is created when a NLP service must be invoked, with information needed to get the data. All services must expose standard APIs to be called by the system; they must read the queue and use the parameters found in it to obtain the input texts. At the end, they pass back their output to the Document Manager components, in order to store it. We stress that the system is equipped with configurable pre-processing services, to be used at ingestion time. The user may choose to store both the raw data (i.e., the incoming legal documents) as well as the pre-processed version, or just the latter. The pipeline and the involved services are defined using the Service Catalog component. As an example, a typical ingestion pipeline would consider: storage of the document *as is*, the raw data; creation and storage of a cleaned copy, where for instance extra blank lines or page breaks, dirty characters inserted by OCR, page headings and so on, are

stripped away; pre-processing of the cleaned text to find document sections and store their positions; full text and metadata indexing. Another important point is that the input of any services is a subset of documents, filtered by the user through the functions of the Front-End components.

4 Statistical Data Generation Pipeline Using Entity Extraction

Entity extraction is a critical step in NLP pipelines, and it involves identifying and extracting relevant entities from text, such as names of people, organizations, locations and so on. There are several motivations for applying entity extraction algorithms in NLP pipelines in the context of legal documents: i) in the *Information Retrieval* area Entity extraction can be used to improve information retrieval by enabling better search results. For instance, if a user searches for a company name, entity extraction can help identify all the mentions of that company in a corpus of text, ii) *Named Entity Recognition (NER)* is a subtask of entity extraction that involves identifying specific types of entities, such as names of people, organizations, and locations. NER can be useful in various applications, such as text classification, statistical report generation from text etc, iii) *Information Extraction*: Entity extraction can also be used to extract structured information from unstructured text. For instance, extracting entities and their relationships from court sentences can be used to create knowledge graphs that capture the underlying structure and meaning of the text. In its simplest form, a pipeline to extract entities from a set of documents may consist of running a single service that annotate some portion of text containing the entity. In our proposal, one of the main objectives is to provide a methodology that allows to define a pipeline for identifying entities from texts to generate statistical information. Usually the statistical reports released by the institutions are based on information contained in structured data or metadata, on the contrary, our infrastructure allows the users to extract the entities present in the documents and generate statistical data from them. This is an innovative tool in the context of legal documents.

The pipeline adopted in our infrastructure can be described as: 1. **Data selection**, that permits create a set to process using a query filters on metadata and/or the words in the text. 2. **Text partitioning**, this step identifies the three main part of Italian court decision: preamble, description of the case and decision, 3. **Named Entities Recognition**: this step receives as input the preambles of the documents and performs a Named Entity Recognition task. It permits to annotate persons, companies, fiscal codes and other information. 4. **Internal Linking**, that consists in relating to each other entities found in the previous step. For example it relates namely persons and companies to their fiscal codes, addresses, birth places and dates and role in the trial (plaintiff, defendant or lawyer). This is a complex task and it is the core of pipeline. 5. **Entity Registry Building** the entities, enriched with attributes deduced through the linking step, are added to the ER, avoiding duplicates and disambiguating homonyms

when possible. Text annotations are updated with identifiers to the entities 6. **Statistical Data Generation**, Once the Entity Registry is populated and the annotations are referred to the entities, it is possible to query the documents using entities instead of text strings; accordingly, any entity related statistics can be computed.

Referring to the infrastructure proposed above we can consider that the data selection step involves the Front-End exploration and search components to enter query parameters and to present results, then Data Filtering component of the Document Manager subsystem to fetch the data. When the user executes the statistical data generation pipeline, the Service Catalog composes the tasks pipeline and for that reasons it checks the analysis services that must be run: text partitioning (from the data pre-processing set services), then Named Entity Recognition service to search in the preambles for the required entity types, and Named Entity Linking. The Entity Registry is updated in the final step. Parameters needed by each service are stored, then services are run in the proper sequence. Each service receives parameters to fetch the documents, and calls the proper document manager components to retrieve the data. Depending on the users' choice, the services may check if their tasks have been already performed and results are not obsolete on a document, to avoid unnecessary processing. For instance, if document segmentation has already been done at ingestion time, it may be skipped. All involved services call again the document manager functions to store the annotations they compute. The last service in the pipeline calls the Entity Registry interface to store the entities (persons and companies) with their identifying attributes; moreover, it calls the document manager to update the annotations.

5 The knowledge extraction pipeline

The knowledge extraction pipeline exploits the ingested documents to mine a set of featuring concepts that provide a topic-oriented description of their textual contents. The concepts extracted from the documents are organized in a graph, where a pair of similar concepts is linked by an edge. Each concept is also connected to the document portions from which the concept emerged, meaning that we can explore the pertinent document segments where a certain concept somehow occurs. Our solution exploits Natural Language Processing (NLP) techniques based on zero-shot learning and context-aware embedding models to enforce concept extraction. A detailed description of the proposed zero-shot learning approach to classification of legal documents is provided in [24]. In the following, we discuss how such an approach to knowledge extraction has been integrated as a pipeline in the infrastructure of Figure 1.

5.1 Data pre-processing

For knowledge extraction, the data pre-processing stage is based on a tokenization step, where the text of each ingested document d is split into a set of chunks.

A *document 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. We stress that the size of the document 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. In this paper, we choose to tokenize documents by defining a chunk for few sentence/phrase detected in a document, up to a maximum size of 512 words. This is particularly appropriate for legal actors (e.g., lawyers, practitioners) that are typically interested in retrieving precise document excerpts in which a given concept of interest appears and can be rapidly read/assimilated.

As a further pre-processing step, the terms appearing in document chunks are lemmatized and a vector-based representation of each document chunk is finally built. 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). For embedding construction, Sentence-BERT [16], a modification of the original BERT model based on siamese and triplets networks, is employed to derive a semantically meaningful embedding for a given sentence/phrase. As such, a document chunk is associated with a set of terms W_k therein contained. Any term is described as $w = (w_l, w_d, \bar{w})$, where w_l is the label of the term (i.e., the lemma), w_d is a description of the term meaning taken from a reference dictionary/vocabulary (e.g., WordNet), and $\bar{w}$ is the corresponding vector-based representation according to Sentence-BERT, respectively. A document chunk k 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 $w \in W_k$. Embedding models have the capability to represent and compare the meaning of entire text blocks like document chunks. On such a target, context-aware embedding models fine-tuned on document similarity tasks, like Sentence-BERT, are appropriate. In the legal field, 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). Sentence-BERT can handle such a kind of situations, which may strongly deviate with respect to everyday conversations.

5.2 Concept extraction

The document chunks are exploited by zero-shot learning techniques to enforce a multi-label classification process with the aim at detecting a set of featuring concepts. Zero-shot learning is an unsupervised classification technique, characterized by the capability to enforce classification without requiring any pre-existing annotation of the considered documents.

Initially, a *seed knowledge* is defined as a set of textual descriptions, each one featuring a concept of interest, namely a *seed concept*, to consider for classification. Typically, for a seed concept, a basic, gross-grained description is provided as a short text (e.g., one or two phrases) or a list of keywords. As an example, for

a seed concept about banking contract, a corresponding textual description used for embedding is bank deposit, safe deposit box, bank credit opening, bank advance, bank account, bank discount. Further concepts are derived from seed ones during the extraction process, and they usually provide a more fine-grained description of the concept instances occurring in the document chunks. A concept c, either seed or derived, 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. Finally, 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 $\bar{c}$. Concept extraction is defined as a progressive, iterative process articulated in the following three steps:

Zero-shot classification. Given a set of concepts (i.e., the seed concepts at the beginning of the process), the document chunks are classified through zero-shot learning. A similarity measure σ, e.g. cosine similarity, is calculated over any pair of embeddings between chunks and concepts. A document 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. The value of α is empirically determined according to experimental results. In this paper, the value $\alpha = 0.3$ is employed in the proposed case-studies and experiments.

Terminology enrichment. Given a document chunk k classified with the concept c, the terms in W_k are exploited for enriching the term set W_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. By enriching the term set W_c, it is possible that more fine-grained concepts emerge from c, and they can be generated as new concepts. The discovery of possible new concepts emerging from c is enforced by clustering the embedding vectors $\bar{w}$ of terms in W_c. The Affinity Propagation (AP) algorithm is adopted to this end, since it allows to detect the emergence of sub-groups of similar terms within W_c, without requiring to "a-priori define" the number of clusters to generate. A new concept c' is created for each cluster returned by AP on the terms W_c of a concept c. 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.
The set of concepts obtained after derivation can trigger the execution of a new cycle based on the above three steps. New derived concepts can contribute to improve the classification of chunks in more fine-grained concepts. Further new concepts can be also discovered through a new execution of enrichment and derivation on the basis of a refined classification result. As such, concept extrac-

tion is characterized by a predefined endpoint condition based on a *termination threshold*. When the number of new concepts created in the derivation step is lower than the threshold, the concept extraction process is concluded. A final concept graph providing a topic-based description of the underlying document corpus is stored in the entity registry for subsequent exploitation by the front-end services. An example of concept graph extracted from a case-study of Italian legal documents will be discussed in Section 6.

6 Application to the Italian context and evaluation

In the following, we discuss some application examples and evaluation results about entity and knowledge extraction. To this end, we consider a corpus of Italian court decisions collected in the framework of the *Next Generation UPP* (*NGUPP*) project, funded by the Italian Ministry of Justice. We will propose our results in the main areas of analysis exposed in the previous sections.

6.1 Statistical Data Generation Using Entity Extraction

Performance of NER algorithms with their attributes has been manually checked on a subset of 50 documents issued by 4 courts on 3 kinds of debates. the results are summarized in table 1, regarding the main entities annotation we consider True Positive (T.P.) only the value correctly found, False Positive (F.P.) correspond to text strings not related to any entities and False Negative (F.N.) are the entities not found by the algorithm. The percentages are considered with respect to total entities found of each type. True Negative do not make sense in this context. We also consider *Inaccurate* entities when either the string denoting the entity was not completely detected, or the entity was not assigned the correct meaning in the document (e.g. a lawyer was considered as a plaintiff). For example: 7.4 % of entity person was not be linked as a lawyers etc. Regarding the *Linked entities* are considered correctly found (True Positive) only when they are correct and linked to the proper main entity.

Main entities:	T.P.	F.P.	F.N.	Linked entities	T.P.	F.P.	F.N.
Plaintiffs (persons)	76.8	7.6	23.2	Sex	88.5	1.3	10.2
Plaintiffs (companies)	100.0	7.1	0.0	Fiscal code	81.8	0.0	18.2
Defendants (persons)	84.8	7.6	15.2	Birth date	78.0	0.0	22.0
Defendants (companies)	78.6	7.1	21.4	Birth place	65.9	7.3	26.8
Lawyers	81.9	7.0	10.7	Postal address	77.8	0.0	22.2

Table 1. Estimated performances of NER and linking: percentages of instances identified

After the steps of NER pipeline have been performed, we report some statistics about people involved in trials. As an example, we computed the percentage

of male and female plaintiff in divorce trials, in three districts (Milan, Rome and Palermo). The meaning of the statistics is of course which of the partners started the divorce. Results are shown in table 2. We considered only cases where parties have been identified with their gender.

District	Trial n.	Male %	Female %
Milan	3195	55.9	44.1
Rome	4583	62.3	37.7
Palermo	1726	53.3	46.7

Table 2. Percentages of divorces started by males and females

6.2 Concept-driven data exploration

We consider a case-study about "unfair competition" as subject matter and we invoke our knowledge extraction pipeline with the aim to explore the concepts extracted from the corpus on such a subject. The user can enforce a preliminary filtering step over the document metadata to select the set of court decisions to consider for concept exploration. The example is based on a dataset of 34 documents resulting from the following filtering operations: first level of judgment, judicial district in North-Western Italy, year of decision from 2008 onwards, subject matter corresponding to 172011 or 172012, that are subject codes related to unfair competition in the Italian law. In Figure 2, we show the concept graph returned by the knowledge extraction pipeline for describing the filtered dataset on unfair competition. We note that most of the graph concepts pertain to the domain of trade justice (e.g., "consortium", "partnership", "transaction"), by also describing specific aspects concerned with unfair competition. Through links, it is possible to move from specific concepts (e.g., "sponsorship") to more general ones (e.g., "business"), and vice-versa. In the example, general concepts are usually associated with more chunks than specific concepts. We also note that some concept labels appear many times (e.g., "business", "sponsorship"), meaning that they refer to different senses of the concept label.

For evaluation of our concept extraction process, we consider EurLex57k [17], that is a dataset of 57,000 EU legislative documents annotated with labels representing entities, concepts, and topics from the EuroVoc thesaurus[1]. The goal of the evaluation is to assess whether our extracted concepts correspond with the labels of EuroVoc used for annotating the EurLex57k dataset. As a baseline, we consider BERTopic [23] since it is a topic modeling approach based on BERT and the mined topics can be straightforwardly compared to our extracted concepts. In Figure 3, we show the precision-recall curve obtained by our concept extraction pipeline when various values of α and β thresholds are employed. We note that our solution outperforms the BERTopic baseline: despite a 0.05

[1] `https://eur-lex.europa.eu/browse/eurovoc.html?locale=en`.

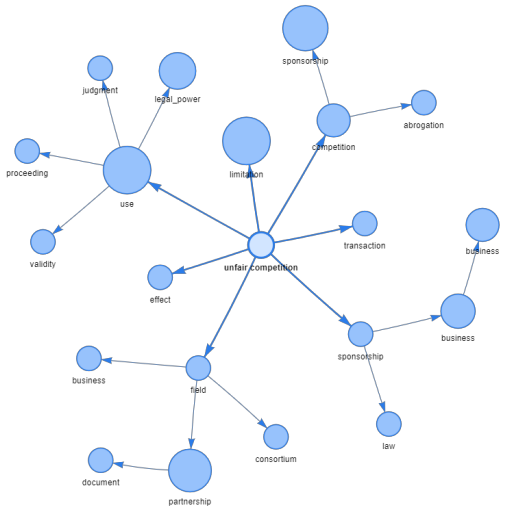

Fig. 2. Example of concept graph returned by the knowledge extraction pipeline for the case-study on "unfair competition". The size of a concept node is proportional to the number of document chunks classified with the concept. The original Italian labels of concepts have been translated into English for the sake of readability.

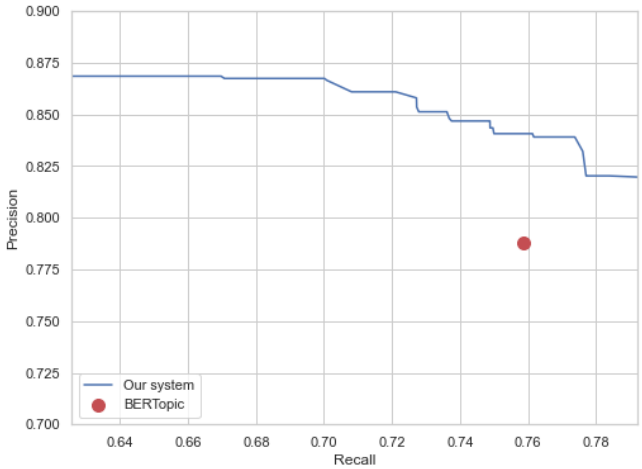

Fig. 3. Precision-Recall curve for concept extraction on the EurLex57k dataset.

decrease, precision remains higher than the baseline even when recall increases (i.e., when more concepts are extracted).

As a further experiment, we consider the results of the zero-shot classification and we evaluate the correspondence of our extracted concepts assigned to chunks w.r.t. the EuroVoc label assigned to documents. Results in terms of precision and recall are shown in Table 3 by providing mean and standard deviation at the document level. In the experiment, the following thresholds are set: $\alpha = \beta = 0.3$. We note that precision and recall of our concept extraction pipeline are not only

Model	Precision	Recall
Our system	**0.593 (0.061)**	**0.681 (0.078)**
BERTopic	0.455 (0.306)	0.422 (0.287)

Table 3. Mean (standard deviation) results for document classification.

higher, but also significantly less variable than the ones obtained by BERTopic according to the standard deviation.

7 Concluding remarks

In this paper, an infrastructure for managing legal documents and related metadata has been proposed. In particular, a service architecture was presented that provides the functionalities of ingestion, archiving and analysis of legal sentences. Some specific processing pipelines based on NLP and machine learning were described and tested. As far as the evaluation of the proposal is concerned, the experiments illustrated above have demonstrated how the proposed infrastructure and services make it possible to provide a new set of semantic functions which allow for the semi-automation of some needs of the Italian Ministry of Justice. In particular, both the excellent quality of the results of the single processing services and the sustainability of the infrastructural proposal were demonstrated. Since the proposed solution is part of a process of continuous development and evolution, various future activities have been planned. Specifically, the expansion of the set of knowledge extraction services and the introduction of a complex workflow management system are planned.

Acknowledgements

This work is partially supported by i) the Next Generation UPP project within the PON programme of the Italian Ministry of Justice, ii) the Università degli Studi di Milano within the program "Piano di sostegno alla ricerca", iii) the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience

Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strenghtening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D, and iv) the project SERICS (PE00000014) under the MUR NRRP funded by the EU - NextGenerationEU.

References

1. Breit, Anna and Waltersdorfer, Laura and Ekaputra, Fajar J. and Sabou, Marta, An Architecture for Extracting Key Elements from Legal Permits, in 2020 IEEE International Conference on Big Data (Big Data), pp. 2105-2110, doi 10.1109/BigData50022.2020.9378375, (2020)
2. Amato, Flora and Mazzeo, Antonino and Penta, Antonio and Picariello, Antonio, Using NLP and Ontologies for Notary Document Management Systems, in Database and Expert Systems Application, 2008. DEXA '08, pp.67-71, doi 10.1109/DEXA.2008.86 (2008)
3. Humphreys, Llio and Boella, Guido and van der Torre, Leon et al., Populating legal ontologies using semantic role labeling, Artificial Intelligence and Law, Vol. 29, pp. 171–211, doi 10.1007/s10506-020-09271-3 (2021)
4. Buey, Maria G. and Garrido, Angel Luis and Bobed, Carlos and Ilarri, Sergio, The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies, in Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016), pp. 438-445, doi 10.5220/0005757204380445 (2016)
5. Ruiz, Marcos and Roman, Cristian and Garrido,Angel Luis and Mena, Eduardo, uAIS: An Experience of Increasing Performance of NLP Information Extraction Tasks from Legal Documents in an Electronic Document Management System, in Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020), pp. 189-196, doi 10.5220/0009421201890196 (2020)
6. Haoxi Zhong and Chaojun Xiao and Cunchao Tu and Tianyang Zhang and Zhiyuan Liu and Maosong Sun, How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, arXiv,cs.2004.12158 (2020)
7. Rabelo, Juliano and Goebel, Randy and Kim, Mi-Young et al., Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021, The Review of Socionetwork Strategies, Vol. 16, pp. 111-133, doi 10.1007/s12626-022-00105-z (2022)
8. Yu, Dian and Huang, Lifu and Ji, Heng, Open Relation Extraction and Grounding, in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 854–864, url https://aclanthology.org/I17-1086 (2017)
9. Eberts, Markus and Ulges, Adrian, Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training, in Frontiers in Artificial Intelligence and Applications Vol. 325 ECAI 2020, Vol. 325, pp. 2006–2013 (2020)
10. Dragoni, Mauro and Villata, Serena and Rizzi, Williams and Governatori, Guido, Combining Natural Language Processing Approaches for Rule Extraction from Legal Documents, un AICOL 2017: AI Approaches to the Complexity of Legal Systems, Lecture Notes in Computer Science, Vol. 10791, pp. 287–300 (2018)
11. Gormley, Clinton, and Zachary Tong. Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. " O'Reilly Media, Inc.", 2015.

12. Andrew, Judith Jeyafreeda and Tannier, Xavier, Automatic Extraction of Entities and Relation from Legal Documents, in Proceedings of the Seventh Named Entities Workshop, pp. 1–8 (2018)
13. Sarika, Jain and Pooja, Harde and Nandana, Mihindukulasooriya and Sudipto, Ghosh and Abhinav, Dubey and Ankush, Bisht, Constructing a Knowledge Graph from Indian Legal Domain Corpus, in Text2KG 2022: International Workshop on Knowledge Graph Generation from Text, Co-located with the ESWC 2022, vol. 3184, pp. 80–93 (2022)
14. Marco Anisetti, Claudio A. Ardagna, Chiara Braghin, Ernesto Damiani, Antongiacomo Polimeno, and Alessandro Balestrucci. Dynamic and Scalable Enforcement of Access Control Policies for Big Data. In Proceedings of the 13th International Conference on Management of Digital EcoSystems (MEDES '21). Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/3444757.3485107 (2021)
15. Chang, Ming-Wei and Ratinov, Lev-Arie and Roth, Dan and Srikumar, Vivek. Importance of semantic representation: Dataless classification., in: Aaai, Vol. 2, 2008, pp. 830–835
16. Reimers, Nils and Gurevych, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
17. Chalkidis, Ilias and Fergadiotis, Manos and Malakasiotis, Prodromos and Androutsopoulos, Ion. "Large-scale multi-label text classification on EU legislation". arXiv preprint arXiv:1906.02192 (2019)
18. Chalkidis, Ilias, et al. "LEGAL-BERT: The muppets straight out of law school." arXiv preprint arXiv:2010.02559 (2020).
19. C. Batini, V. Bellandi, P. Ceravolo, F. Moiraghi, M. Palmonari and S. Siccardi, "Semantic Data Integration for Investigations: Lessons Learned and Open Challenges," 2021 IEEE International Conference on Smart Data Services (SMDS), Chicago, IL, USA, 2021, pp. 173-183, doi: 10.1109/SMDS53860.2021.00031.
20. Licari, Daniele, and Giovanni Comandè. "ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law.", CEUR WORKSHOP PROCEEDINGS. Vol. 3256. CEUR-WS, (2022)
21. Bhattacharya, Paheli, et al. "FIRE 2019 AILA track: Artificial intelligence for legal assistance." Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation. (2019)
22. Ardagna C.A., Bellandi V., Bezzi M., Ceravolo P., Damiani E., Hebert C. "Model-Based Big Data Analytics-as-a-Service: Take Big Data to the Next Level" (2021) IEEE Transactions on Services Computing, 14 (2), pp. 516 - 529. DOI: 10.1109/TSC.2018.2816941
23. Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022)
24. Bellandi, V, Castano, S., Ceravolo, P., Damiani, E., Ferrara, A., Montanelli, S., Picascia, S., Polimeno, A., and Riva, D.: "Knowledge-Based Legal Document Retrieval: A Case Study on Italian Civil Court Decisions", Proc. of the 1st Int. Knowledge Management for Law Workshop (KM4LAW), Bozen-Bolzano, Italy. CEUR-WS, (2022)
25. Bellandi, Valerio & Siccardi, Stefano. (2023). An Entity Registry: A Model for a Repository of Entities Found in a Document Set. 01-12. 10.5121/csit.2023.130301.
26. Amanda Carmignani & Silvia Giacomelli, 2010. "Too many lawyers? Litigation in Italian civil courts", Bank of Italy, Economic Research and International Relations Area. `https://ideas.repec.org/s/bdi/wptemi.html`