A Service Infrastructure for Management of Legal Documents

Valerio Bellandi[1,*,†], Silvana Castano[1], Alfio Ferrara[1], Stefano Montanelli[1],
Davide Riva[1] and Stefano Siccardi[1]

[1]*Università degli Studi di Milano*
DI - Via Celoria, 18 - 20135 Milano ,

Abstract

Managing legal documents, particularly court judgments, can pose a significant challenge due to the extensive amount of data involved. Traditional methods of document management are no longer adequate as the data volume continues to grow, necessitating more advanced and efficient systems. To tackle this issue, a proposed infrastructure aims to establish a structured repository of textual documents and enhance them with annotations to facilitate various subsequent tasks. The framework is designed with sustainability in mind, allowing for multiple services and applications of the annotated document repository while taking into account the limited availability of annotated data. By employing a combination of machine learning and syntactic rules, a set of Natural Language Processing (NLP) services pre-processes and iteratively annotates the documents. This approach ensures that the resulting annotations align with the organizational processes utilized in Italian courts. The solution's feasibility was demonstrated through experiments that employed different low-resource methods and solutions, effectively integrating these approaches in a meaningful manner.

Keywords

Legal Document Annotation, Named Entity Recognition, Concept Extraction, Zero-Shot Learning

1. Introduction

Court rulings and other legal document are rich in information that should be made available to different users categories, for instance: judges and lawyers to find cases similar to one at hand, staff of the justice department to evaluate courts' performance, the general public for statistical reports and so on. Obviously, users requirements can grow and change over time. Accordingly, any infrastructure aimed at managing legal documents should not prescribe in advance any specific types of information management. On the contrary, it should be able to accommodate new services for data preparation, extraction, manipulation as new requirements emerge.

In the solution we propose, this flexibility is achieved providing an environment where any additional services can be integrated, sharing a common data repository that is accessed through a set of APIs. The infrastructure design ensures scalability, so that it is stable even for

ITADATA2023: The 2nd Italian Conference on Big Data and Data Science, September 11–13, 2023, Naples, Italy

valerio.bellandi@unimi.it (V. Bellandi); silvana.castano@unimi.it (S. Castano); alfio.ferrara@unimi.it (A. Ferrara); stefano.montanelli@unimi.it (S. Montanelli); davide.riva@unimi.it (D. Riva)
0000-0003-4473-6258 (V. Bellandi); 0000-0002-3826-2407 (S. Castano); 0000-0002-4991-4984 (A. Ferrara); 0000-0002-6594-6644 (S. Montanelli); 0009-0003-9681-9423 (D. Riva); 0000-0002-6477-3876 (S. Siccardi)

CEUR Workshop Proceedings (CEUR-WS.org)

increasingly large volumes of data, an essential characteristic for ensuring that the system can continue to deliver high-quality services[?] and keep up with changing requirements[?].

Some specific design goals are:

1. store documents' texts and metadata
2. provide the usual searching capabilities on both texts and metadata
3. recognize and classify entities occurring within documents, using reference entity types or an entity taxonomy
4. disambiguate entities and searching for their occurrences
5. perform statistical analyses and cluster documents

A specific functionality aims at extracting a concept network from documents. This network can provide services to search, explore and analyze the legal documents, driven by concepts instead of keywords or entities. Two application examples to concrete case-studies in the framework of the Italian digital justice are described and evaluation results are finally discussed to show the feasibility of the proposed solution in real situations.

2. Related Work

The management of legal documents has been considered by several architecture designers, with the purpose of extracting knowledge, in addition to meeting the normal needs of text query.

The system described in [?] uses a combination of rule based and statistical NLP techniques to help lawyers by suggesting arguments and extracting relevant information from texts.

Ontologies have been considered both in [?] and [?]. The former system manages paper documents, automatically transforming them into RDF statements, the latter semi-automates the extraction of norms and populates legal ontologies. They use general-purpose NLP modules combined with pre- and post-processing based on rules. Reference [?] describes an implementation working on a real document management system and performing intensive processes; the system has been later improved (see [?]). As the cases quoted above, it uses ontologies to describe the documents' structures and the entities that can be found. It shares some characteristics with our design, as it is based on microservices and message brokers, however entities do not play a central role like in our system.

Considering more specifically knowledge extraction and integration in the legal domain, several NLP techniques have been proposed (see [?] for a review). The legal case retrieval task and the legal case entailment task are two typical examples of problems faced in this field. The first task consists in extracting supporting cases for the decision of a given case; the latter aims at identify a paragraph from existing cases that entails the decision of a new case. See for instance the Competition on Legal Information Extraction/Entailment (COLIEE) organized since 2017 [?] and the Artificial Intelligence for Legal Assistance (AILA) shared task [?].

Named Entity Recognition (NER) can be considered a basic task, and more refined techniques can be based on its results. In particular, the Relation Extraction (RE) task is particularly relevant for the present work, as it allows to connect entities to their attributes (e.g. persons with their birth data) so that they can be uniquely identified. RE is a challenging task, and

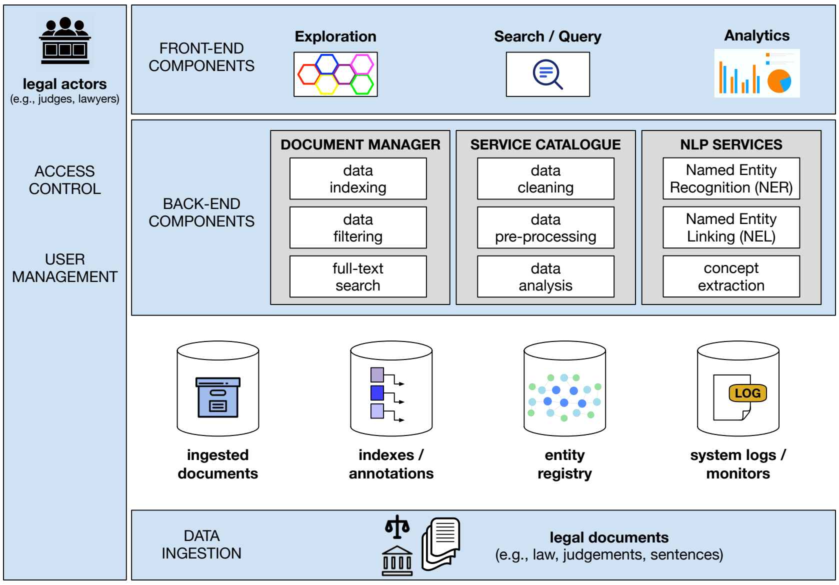

Figure 1: The System Infrastructure.

several techniques have been considered, for instance joint entity and relation extraction, sets of pre-defined relation classes, combinations of a statistical methods and rule based techniques (see e.g. [?], [?] and [?]). The system described in [?] has been applied to the Indian Supreme Court Judgements to extract entities and relations. An ontology described relation types and triples were the final output of the process. A gold standard of five manually annotated documents was used to evaluate the results.

The lack of annotated data is in general an issue for supervised techniques and especially for the concept extraction task. Fine-tuned embedding models have been proposed both for English and Italian language legal documents (see [?] and [?] respectively), as well as zero-shot classification (ZSC) (e.g., [?]). We use a pre-trained model without fine-tuning, relying on a contextual, transformer-based embedding models (i.e., Sentence-BERT[?]) to obtain a semantically-meaningful document representation. ZSC techniques are used to classify unlabeled data instances without annotation.

3. Architecture

The architecture is illustrated in Figure 1. The storage layer contains ingested documents, as raw data, and their metadata coming from the operational systems, in a document database. The expected format for raw data is plain text, if the documents are scanned pdf files or images, it is expected that an OCR tool has been used to extract text. Annotations created by the systems and pre processed versions of the documents are stored in a repository. An index system allows searching all the above. In our architecture, texts and metadata are stored in an ElasticSearch[?] instance, while annotations are stored in a SQL database as described in our previous work [?]. Entities are stored in an *Entity Registry (ER)*, that is implemented as a graph database.

The ER contains an entry with a unique ID for each entity occurring in the documents. It is based on a description of the entity types and of the attributes to uniquely identify them (the ER metamodel) and is accessed through a suitable set of APIs (see [?] for details of the ER logic).

The system is equipped with several Front-End components for specific user needs, like querying the data for entities or concepts, managing single documents and browsing similar ones, requesting statistical analyses, exploring the ER and so on. This is actually an extension of our previous work [?].

Our architecture considers a specific NLP Service for each required task, like NER, Entity Linking, Concept Extraction. Also ancillary tasks, that may create new versions of the documents, as data cleaning, pre-processing and summarization, are performed by dedicated services. Pipelines of services are managed by an orchestrator, based on a service catalog. For instance, an ingestion pipeline could include storing the document without any modification and its metadata as received, creating a cleaned copy (with stripped headings, blank lines, page numbers, etc.), storing start and end positions of the document sections, and adding a set of important annotations and indexing both the text, the metadata and the annotations.

Each service acts on a set of documents chosen by the user through the front end. It receives through a communication queue the information needed to fetch the data; in this way multiple instances for highly demanded services can be seamlessly created. Client programs, including both services and front end components, that need to access or modify data, use APIs of the Document Manager component, instead of interacting directly with the underlying databases.

4. Pipeline for Statistical Data Generation

In general, statistical reports, that the institutions have to produce on a regular basis, need to aggregate specific information that is not available in the documents metadata. For instance, age and gender of plaintiffs and defendants, correlations between outcomes of the first and second degree cases or economic value of the dispute can be found only in the text of the judgements. In order to extract such data, NER is the starting step, as it identifies specific types of entities, such as names of people and organizations, dates and locations, codes and digits representing amounts of money, and so on. A second step consists in linking together the entities, in order to obtain more detailed information, e.g. recognize that a date is the birth date of a person.

We describe a pipeline, that we used to create demographic statistics of plaintiffs and defendants, we stress however that it can be easily generalized:

1. Document filtering, that consists in creating a set of documents querying metadata and full texts
2. Identifying the main sections of each document in the set
3. Named Entity Recognition in each document section, so that entities are correlated to their locations in the text. In our example, NER was used to find and annotate persons and companies, fiscal codes, date, cities and addresses
4. Linking entities to each other, for instance: persons to their roles (plaintiff, defendant, lawyer), fiscal codes and birth data
5. Entry creation in the entity registry for each person: using names, birth data and fiscal codes it is possible to have unique entries, avoiding duplicates and disambiguating

homonyms when possible

6. Statistics report generation, where each mention in the document corpus is related to an entry in the ER, so that correct data about gendres, ages and roles can be obtained

We note that finding linking between entities (Internal Linking, IL) is at the core of our methodology and is the most complex task of the pipeline. For this reason, we consider that the IL services may provide an uncertainty score[**?**], expressing the degree of belief one can have in their results in the sense of e.g. [**?**]. Actually even many standard tools for NER, provide this type of scores. Uncertainty scores are then propagated to the statistical report generators and may be used to compute a kinf of confidence intervals for the results.

The pipeline is easily mapped on the proposed infrastructure, as the Front End components receive query parameters and show results, interacting with the Document Manager to retrieve the data. At execution time of the pipeline, the Service Catalog calls the needed services in the proper order: the pre processing service to perform text partitioning, the NER service, then the Named Entity Linking service and the Entity Registry to create the entries. In turn, individual services interact with the Document Manager to fetch the data they need. The user may choose to skip tasks that have already been performed (for instance, data partitioning might be executed once for all at ingestion time). The Document Manager is called again by the analytical services, when they need to store new annotations. The Entity Linking service calls the ER interface to store the entities and the Document Manager again to update the annotations with the entities IDs. As already stated, the platform allows users to easily define pipeline like the one described above.

5. Application to the Italian context and evaluation

A corpus of Italian court decisions was used to test the procedures and provides examples to illustrate the techniques described above for statistical data generation using entity extraction. The documents were collected in the framework of the *Next Generation UPP* (*NGUPP*) project, funded by the Italian Ministry of Justice.

First of all, we manually checked the performances of the NER algorithms on a document sample, in terms of the ability of both finding relevant entities and detecting correct relationships among them. The sample consisted of 50 judgements by 4 courts on 3 kinds of cases. For main entities, that is persons and companies, we considered as True Positive (T.P.) only cases where the value correctly found; False Positive (F.P.) are text strings not related to any entities; False Negative (F.N.) are the entities missed by the algorithm. True Negative do not make sense in this context, as any not spotted words could be considered as a true negative. Finally, we defined inaccurate entities cases where either the entity was not completely detected (e.g. the algorithm missed the second name of a person), or their roles were not correctly assigned (e.g. lawyer instead of plaintiff). Linked entities must be correctly detected and linked to the correct person to be counted as True Positive. Table 1 summarizes the results.

Our example statistics aims at describing which partner started divorces, comparing three Italian geographical districts (Milan, Rome and Palermo). For this, based on the NER popeline, we counted the numbers of male and female plaintiffs in divorce cases. Results are shown in table 2.

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

Main entities:	T.P.	F.P.	F.N.	Linked entities	T.P.	F.P.	F.N.
Plaintiffs (persons)	76.8	7.6	23.2	Gender	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 2
Percentages of male and female plaintiffs in divorce cases

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

6. Conclusion

This paper introduces a framework for effectively managing legal documents and associated metadata. It presents a service architecture that offers functions such as ingestion, archiving, and analysis of legal sentences. The paper also discusses specific processing pipelines that utilize NLP and machine learning techniques, which were described and tested.

Regarding the evaluation of the proposed solution, the aforementioned experiments demonstrate how the infrastructure and services provided enable the semi-automation of certain requirements of the Italian Ministry of Justice.

Since the solution is part of an ongoing development and evolution process, several future activities have been planned. These include expanding the range of knowledge extraction services and implementing a comprehensive workflow management system.

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