AI-Assisted Legal Holding Extraction

Praveen Bushipaka, Daniele Licari*, Gabriele Marino, Giovanni Comandé and Tommaso Cucinotta

Scuola Superiore Sant'Anna, P.zza dei Martiri della Libertà, Pisa, 56100, Italy

Abstract

This paper provides an overview of the investigations being carried out at Scuola Superiore Sant'Anna on the use of Artificial Intelligence techniques for automated extraction of rhetorical roles and legal holdings from Italian case documents. These activities are framed within the "Giustizia Agile" project funded by the Ministry of Justice, aiming at improvements to the efficiency of the Italian justice system, making use of advanced information technology means, among others.

Keywords

Artificial Intelligence, BERT, Summarization, Legal Holding Extraction, Rhetorical Roles, Legal AI

1. Introduction

In every country, the efficiency of the judicial system has an impact on the social and economic life of citizens. Italy has been constantly trying to make its legal system more efficient and in line with other European countries. For example, in the 2022 EU Justice Scoreboard [1], Italy was reported as being among the countries with the least efficient judicial system, with more than 500 days needed for the first sentence, 800 days for the appeal and reaching up to 1300 days for final judgments by the Supreme Court [2]. One factor believed to bring a tremendous potential for improving the efficiency of public administration systems in general, including judicial systems, is the widespread adoption of Information and Communication Technologies (ICTs), supporting fully digitalized processes. Indeed, the CEPEJ report by the EU Council [3] includes a survey on the use of ICT in judicial systems, highlighting for example that Italy exhibits the lowest score among EU countries on the Criminal justice ICT index (but with a much better ICT index on Civil and Administrative justice).

In this context, we can understand the efforts being made in the "Giustizia Agile" project[1], funded by the Italian Ministry of Justice. This project is framed within a wider initiative to enhance the performance of judicial offices, aiming at significant reductions of the backlog, by investigating on finding the major bottlenecks and critical factors addressing adversely the duration of judicial processes; investigating the opportunities to add several innovations on the side of management and organization of the processes, as well as to embrace a wider adoption of digitalization of the processes through the use of ICTs.

This very last topic is the one where this paper fits, reporting on some key experimentation being done with the use of Artificial Intelligence tools, and specifically Large Language Models (LLMs), in the area of automated summarization and Rhetorical Role Classification of Italian case documents. We focused on the extraction of legal holdings from Italian administrative justice documents. This activity is carried out by the highest exponents of Italian justice and is crucial to facilitate access to justice, create a 'precedent' and ensure transparency in decisions. Furthermore, this is a delicate task because lawyers and judges rely on legal holdings to select case-relevant documents when searching for similar cases. Extracting this information from a judgment is a complex task that requires time-consuming efforts and specific skills. Here, we present an innovative approach that combines a rhetorical role classifier, text summarization, and a scalable search engine to accurately and efficiently retrieve and analyze legal holdings from Italian case documents. Identifying the rhetorical roles of the different text segments allows for a better understanding of the structure and content of the document, which can help guide the summarization process. Irrelevant information (e.g. introduction) can be filtered out in pre-processing allowing the summary model to focus exclusively on the most important information in the document.

Previous attempts at using rhetorical roles classification for text summarization in the legal field date back to 2004. For example, LetSum [4] assigns Rhetorical roles to the sentences and uses TF-IDF to rank them. A percent-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

praveen.bushipaka@santannapisa.it (P. Bushipaka); daniele.licari@santannapisa.it (D. Licari); gabriele.marino@santannapisa.it (G. Marino); giovanni.comande@santannapisa.it (G. Comandé); tommaso.cucinotta@santannapisa.it (T. Cucinotta)

0000-XXXX-XXXX-XXXX (P. Bushipaka); 0000-0002-2963-9233 (D. Licari); 0000-XXXX-XXXX-XXXX (G. Marino); 0000-0003-2012-7415 (G. Comandé); 0000-0002-0362-0657 (T. Cucinotta)

[1]More information is available at https://www.unitus.it/it/unitus/mappatura-della-ricerca/articolo/giustizia-agile.

age of sentences for each rhetorical role is then selected to be a part of the summary. The work [5] approached the identification of Rhetorical roles with Conditional Random fields, extending to Extractive text summarization using term distribution. There are different classes of text summarization methods. *Extractive summarization* involves identifying the most important sentences or passages from the original text and combining them to create the summary [6]. This method has been widely experimented with within the Legal area, and a few tools were developed specifically for the Legal Domain. On the other hand, *abstractive summarization* generates new text which is not present in the processed documents. This method has been explored in [7] and proved efficient.

In our approach, we focused on an extractive method, due to two main reasons: (i) it highlights the most relevant sentences in a given document, constituting an effective way to speed up a Judge's work, and (ii) summarizing long documents is extractive in nature [8], as this method takes advantage of the discourse structure [9] to generate factually consistent summaries, preserving the meaning of the original document [10]. However, previous efforts in this area were done only on English datasets.

In this paper, we propose a platform based on Italian Legal BERT models [11] to extract legal holdings from Italian administrative justice documents using rhetorical roles classification and extractive text summarization. We use a Hierarchical BERT model to identify only the most important sentences and apply an extractive summarization algorithm to improve the performance of the summarization. Later, we feed this information as metadata into an efficient information retrieval system. v

2. Legal Holding Research Platform Overview

The platform we are building consists of three stages, exemplified in Figure 1. In the first stage, we identify the most important rhetorical roles of each sentence present in a legal document using Hierarchical BERT. In the final stage, we ingest the documents, sentences, and holdings into Elasticsearch, letting the search engine index these additional meta-data, to ease later searches by users.

2.1. Rhetorical Roles Classification

In the first phase of the model, we predict Rhetorical roles for each sentence. We used a Hierarchical BERT model for this task. Each sentence is categorized into a single role. Overall, we categorized 5 different roles (INTRODUCTION, PARTIES, DEVELOPMENT, REASON, and CONCLUSION). The sentences of REASON are filtered which are those that contain the information on the legal holding.

2.2. Legal Holding Extraction

We used an Extractive summarization method to extract holdings from the legal documents. We used BERT with a regression head. The top 5 sentences are picked based on the scores and chosen as holdings.

2.3. Legal Search Engine

The final stage of the AI system includes a search engine for the efficient retrieval of a large corpus of Italian Legal documents. The documents collected, generated roles, and extracted summaries were given to the data store. A web app will be developed for easier usage.

3. Methodology

Our work is based on fine-tuning the Italian-Legal-BERT [11] model for both rhetorical role classification and holdings extraction. However, we used different approaches for these two tasks, as explained below.

3.1. Dataset Description

We used an ITA-CASEHOLD dataset, which consists of 1101 judgments and holding pairs between the years of 2019 and 2022 collected from the Italian Administrative Justice. The dataset consists of a wide range of issues, including public contracts, environmental protection, public services, immigration, taxes, and compensation for damages caused by the State. It also provides citizens with the opportunity to challenge administrative decisions in an independent and impartial trial. The dataset was further divided into 792 documents in the training set, 88 in the validation set, and 221 in the test set. A token-level compression ratio between a document and its holding shows that there is a high standard deviation across all the datasets w.r.t the length of documents and holdings. This is because the documents are quite long, whilst their holdings are much shorter.

A new dataset was created for the training and evaluation of the rhetorical role classifier model. We extracted and annotated (mainly using regular expressions) 152,368 sentences from 1,503 Italian civil cases. For each sentence of the dataset, we derived its rhetorical role between the following:

1. INTRODUCTION: an indication of the judge who pronounced it; an indication of the parties and their lawyers;
2. CONCLUSION OF THE PARTIES: the conclusions of the prosecutor (if any) and those of the parties;

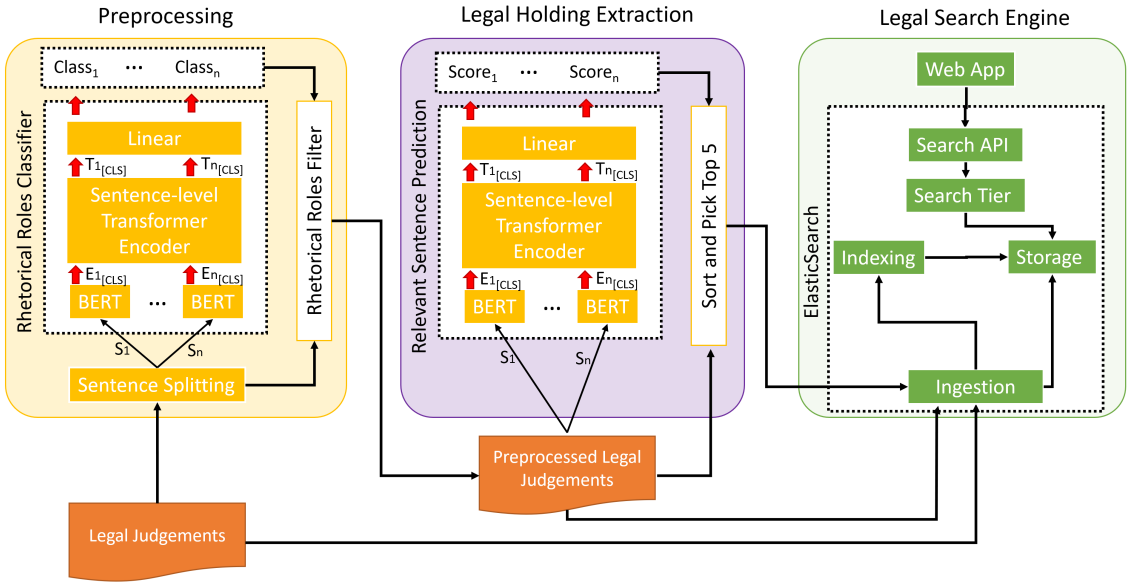

Figure 1: Legal Holding Research Platform Overview

3. DEVELOPMENT OF THE TRIAL: summary of the appealed judgment and reasons of appeal;
4. REASON: the concise statement of the factual and legal reasons for the decision (the statement of reasons);
5. CONCLUSION: the decisional content of the judgment.

Both datasets were split 80% for training models and 20% for model testing. The data were obtained through scientific collaboration agreements between some Italian courts and the Scuola Superiore Sant'Anna. The ITA-CASEHOLD dataset will be publicly released.

3.2. Rhetorical Roles Classification

The identification of the roles that different text segments play in a larger document has been done using a hierarchical BERT approach, in order to contextualize a single sentence based on the content of the document. This model is based on a layered architecture whose bottom layer is Italian-LEGAL-BERT and whose top layer is a 2 layers transformer encoder. The sentences to classify are tokenized and given as inputs to Italian-LEGAL-BERT. The CLS output tokens are then retrieved and fed to the transformer encoder, which extracts relevant features for each sentence. These features are then processed by a simple softmax-based classification layer to get the final predictions. We will provide precise details about the training and performance of this model in further work.

3.3. Italian-LEGAL-BERT Holding Extraction

We used a novel extractive method called Harmonic Mean-BERT. This approach involves fine-tuning the Italian-LEGAL-BERT model to predict a score for each sentence in a document. The scores for training and evaluation were given by the harmonic mean of Rouge R-1 and Rouge R-2 scores (generated by ITA-ROUGE a modified version of Rouge metric for the Italian Language) between the sentence and the corresponding document holding. We generate these scores only for the training and validation sets.

Since the sentences were already rhetorical role classified, based on the scores generated, we then only chose the sentences which have the highest importance. The higher the score of a sentence, the higher the similarity between the sentence and its corresponding holding. For our experiment, REASON was the most important and DEVELOPMENT OF THE TRIAL (DEVELOPMENT) was the second most important. We derived these two were the most important by their scores, 75th percentile of these sentences had score of more than 2.5 whereas other roles were near zero.

We made two datasets by removing sentences with other roles, (i) Only with REASON, (ii) with REASON, and DEVELOPMENT. After getting the scores and choosing only the important sentences of a document, we fine-tuned the Italian LEGAL BERT model with a regression head to predict these scores.

In more detail, the following steps have been followed:

1. R-1 and R-2 scores between each sentence and its respective document holding are computed for the training and validation sets.
2. To retrieve a single score out of the R-1 and R-2, we computed their harmonic mean for each sentence.
3. Based on these scores and the previously predicted roles from the Rhetorical roles classifier, we chose the most important roles. The higher the score, the higher the importance. Two datasets were created based on this.
4. Italian-LEGAL-BERT was fine-tuned in the regression task of predicting the score for a given sentence.
5. The validation dataset was used to determine the optimal number of top k sentences to compose the final holding. We tried k = 3,5,7 and found that k = 5 yielded the best results.

For testing, we followed the steps detailed below:

1. Two datasets were created based on roles similar to the training and validation sets. However, we don't calculate the scores here beforehand. Instead, we use the trained model to predict them.
2. The sentences were then grouped into documents based on their document id.
3. The trained model was used to compute the score of each sentence.
4. The sentences were sorted by predicted scores.
5. The top 5 sentences were selected and sorted according to their index position in the original document to compose the final holding.
6. The ROUGE scores were evaluated between the extracted and the original holdings.

Our software stack included PyTorch, Hugging Face transformers, and Py-Rouge. We used Italian-LEGAL-BERT as the encoder. This model has an embedding dimension of 768, an input token size of 512, 12 hidden layers with 12 attention heads, and an attention dropout of 0.1. A sequence regression head (i.e. a linear layer) was added to the pooled output. The training was carried out with an AdamW optimizer and a linear scheduler. We trained both datasets for 4 epochs, using a batch size of 16 and setting 256 as the maximum sequence length.

3.4. Legal Holding Search Engine

For the final stage, we adopted Elasticsearch for data storage and retrieval. It is built on the Apache Lucene [12] architecture, which uses inverted term frequency and Okapi BM25 [13] for ranking.

The documents, along with their generated rhetorical roles and extracted summaries, will be indexed into the Elasticsearch data store. Additional metadata available for each document will also be indexed along with the documents. A tokenization layer on top of the Elasticsearch data store will be added to tokenize the input text. The search engine will be developed with a web app for the judicial people to be able to use it. This efficient retrieval of documents and their holdings might fasten the process of searching through documents.

Apart from search, Elasticsearch can also be used for analyzing data. This will be explored alongside the main search engine functionality while developing the final system.

Table 1
Comparison on ROUGE scores.

Model	R-1	R-2	R-L	R-W
REASON	**54.91**	**36.44**	**31.37**	**11.97**
REASON + DEVELOPMENT	54.81	35.94	30.37	11.19

4. Preliminary Results

Our experiments showed that the hierarchical approach based on BERT and Transfomer improved the classification performance of rhetorical sentences by +12% in terms of Matthews Correlation Coefficient (from 0.81 to 0.91) compared to a model based only on BERT.

The experiments on holding extraction were on two datasets with different filters on the rhetorical roles: 1) only with the REASON and 2) with REASON + DEVELOPMENT OF THE TRIAL (REASON + DEVELOPMENT). Their performance was evaluated with ITA-ROUGE, a modified version of the ROUGE metrics for the Italian language. The experiments were carried out on an NVIDIA-DGX system equipped with a 32GB TeslaV100 GPU. REASON outperforms REASON + DEVELOPMENT proving that rhetorical roles and picking only the important sentences can yield better results.

5. Conclusions and Future Work

In this paper, we showed that the quality of extractive summarization can be increased by adding a Rhetorical Role layer and choosing only the most important parts of the document. This outperforms the HM-BERT model, which was trained on the same ITA-LEGAL-BERT without Rhetorical roles. Our future work involves in the future development of AI tools that can improve the performance of Judicial Offices. This includes information retrieval, summarization, classification, question answering, and others. For our immediate future work, we will explore the possibilities of using a search engine paired

with the summarization and role classification prototypes we already built.

References

[1] E. Commission, the 2022 EU Justice Scoreboard, https://europa.eu/!CJdXbP, 2022.

[2] D. Lettig, Italy, EU's least-efficient judicial system, https://www.euractiv.com/section/politics/short_news/italy-eus-least-efficient-judicial-system/, 2021.

[3] European judicial systems CEPEJ Evaluation Report – 2022 Evaluation Cycle (2020 Data) – Part 1 Tables, graphs and analyses, https://rm.coe.int/cepej-report-2020-22-e-web/1680a86279, 2022.

[4] A. Farzindar, G. Lapalme, Legal text summarization by exploration of the thematic structure and argumentative roles, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 27–34. URL: https://aclanthology.org/W04-1006.

[5] M. Saravanan, B. Ravindran, S. Raman, Automatic identification of rhetorical roles using conditional random fields for legal document summarization, in: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I, 2008. URL: https://aclanthology.org/I08-1063.

[6] J. Cheng, M. Lapata, Neural summarization by extracting sentences and words, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 484–494. URL: https://aclanthology.org/P16-1046. doi:10.18653/v1/P16-1046.

[7] P. Kalamkar, A. Tiwari, A. Agarwal, S. Karn, S. Gupta, V. Raghavan, A. Modi, Corpus for automatic structuring of legal documents, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4420–4429. URL: https://aclanthology.org/2022.lrec-1.470.

[8] H. Y. Koh, J. Ju, M. Liu, S. Pan, An empirical survey on long document summarization: Datasets, models, and metrics, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3545176. doi:10.1145/3545176.

[9] Y. Dong, A. Mircea, J. C. K. Cheung, Discourse-aware unsupervised summarization for long scientific documents, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1089–1102. URL: https://aclanthology.org/2021.eacl-main.93. doi:10.18653/v1/2021.eacl-main.93.

[10] P. Cui, L. Hu, Sliding selector network with dynamic memory for extractive summarization of long documents, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5881–5891. URL: https://aclanthology.org/2021.naacl-main.470. doi:10.18653/v1/2021.naacl-main.470.

[11] D. Licari, G. Comandé, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: CEUR Workshop Proceedings (Ed.), The Knowledge Management for Law Workshop (KM4LAW), 2022.

[12] A. Białecki, R. Muir, G. Ingersoll, Apache lucene 4, in: OSIR@SIGIR, 2012.

[13] G. Amati, BM25, Springer US, Boston, MA, 2009, pp. 257–260. URL: https://doi.org/10.1007/978-0-387-39940-9_921. doi:10.1007/978-0-387-39940-9_921.