Legal Summarization: to each Court its own model

Flavia Achena[1], David Preti[1], Davide Venditti[2], Leonardo Ranaldi[2], Cristina Giannone[1], Fabio Massimo Zanzotto[2], Andrea Favalli[1] and Raniero Romagnoli[1]

[1] *Cognitive and AI R&D, Almawave S.p.A.*
[2] *Università di Roma Tor Vergata*

Abstract

In the Italian Civil Law System, easily accessing legal judgments through *massime* is crucial. In this work, we compare extractive summarization models to produce massime in two Italian courts: the *Constitutional Court* and the *Supreme Court*. The aim of our study is to assess the effectiveness and efficiency of these models in summarizing the decisions of the two courts. Through a comprehensive analysis of two large datasets, we evaluate the quality of the summaries generated by each model and their ability to capture the key legal principles and linguistic features present in the courts' decisions.

Keywords

Legal Text Analysis, BERT-based Summarization, Legal NLP,

1. Introduction

In civil and common law systems, accessing legal judgments to retrieve legal decisions is crucial when lawyers have to defend clients, prosecutors have to build cases, and judges have to draw decisions. To ensure widespread information on the decisions of the courts, in Italy, for this purpose, a specific body drawn up *massime*.

These *massime* present, in a short but detailed way, a legal principle present in judgments. Hence, justice professionals can read these *massime* instead of the complete legal decisions.

The process of analyzing judgments and extracting relevant sentences can be significantly simplified through the use of pre-trained models [1, 2], which serve as versatile universal sentence/text encoders, capable of addressing various downstream tasks, including summarization [3]. These models consistently outperform other approaches, especially after fine-tuning or domain-adaptation [4]. However, despite the success of pre-trained transformers in other summarization tasks, the task of producing *massime* is challenging for current extractive and abstractive summarization systems. Unlike standard summaries, *massime* must follow rigorous specifications in some courts. Extractive and abstractive summarization datasets and relative systems, in contrast, aim to reduce the size of a text while preserving its overall meaning. However, this approach differs from the specific requirements of creating *massime*.

Additionally, legal texts are often extensive, further increasing the summarization task's complexity. Identifying the portions of the text that contain the relevant information to be reported in the *massime* becomes challenging due to their length[5].

Legal document summarization has seen rapid progress in recent years, and several approaches[6, 7] have been proposed to manage this kind of data, ranging from fine-tuned Transformer models on legal domain, Reinforcement Learning to Generative Models[8].

In this paper, we compare extractive summarization models to produce *massime* in two different contexts: the Constitutional Court (Corte Costituzionale) and the Supreme Court (Corte di Cassazione). We discuss similarities and differences about *massime* and how the kind of court impacts the data in terms of their availability and privacy management. Then, we propose two models tailored to the specific type of courts, discussing the approaches we implemented to circumvent the issue related to lengthy documents. Results of the experiments confirm that producing *massime* is a real challenge even for dedicated systems. Hence, these systems should be designed as facilitators in a human-in-the-loop environment [9].

2. Different courts, Different Judgments, and Different *massime*

The data for the Italian legal domain have some peculiarities that require careful consideration. Different courts, such as the Constitutional and Supreme Court, produce different judgments, leading to different *massime*. Moreover, within the same court, there can be judgments with varying numbers of related *massime*, ranging from one to five or even more. In addition, the availability of data depends on the presence or absence of sensitive informa-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

[first-initial].[surname]@almawave.it (F. Achena); [name].[surname]@uniroma2.it (D. Venditti)

tion within the judgments, so the legal courts provide access to the data in different ways.

Both Constitutional and Supreme Court courts share that producing a *massima* requires a relevant cognitive task carried out by the "Massimario Office" body, as it involves identifying the principle of law present in the judgment and satisfying some precise criteria for its writing. The following are the characteristics of the two Italian Courts under consideration (Sec. 2.1), the nature of their judgments and *massime*, how a *massima* is structured (Sec. 2.3), and, finally, a comparative analysis of the two types of corpora that can be derived from these two courts (Sec. 3) is provided.

2.1. Two Italian Courts: Constitutional Court and Supreme Court

The Italian Constitutional Court has the primary responsibility to assess the constitutionality of the acts and laws of the State and the Regions. Among other functions, it assesses charges against the President of the Republic in accordance with constitutional provisions. The Court examines the admissibility of abrogative referendums. To ensure impartiality and independence, the Constitutional Court is composed of 15 lawyers, chosen from among judges, law professors, or lawyers with at least 20 years of experience.

The Supreme Court - also known as the "Corte di Cassazione" - is the highest authority in the Italian judicial system. It serves as the court of final appeal and has two main functions. Firstly, it resolves judicial conflicts to determine which judge has jurisdiction over a case. Secondly, it has a *nomophylactic* function, ensuring that the law is interpreted uniformly. Within the Court, there is the "Massimario Office", responsible for identifying nomophylactic judgments and producing concise summaries called *massime*. These summaries contain the legal principles from the Court's judgments, not just a summary of the cases themselves. The primary objective of the Massimario Office is to disseminate legal knowledge and facilitate comprehension of past court decisions. To accomplish this, the office updates its collection by incorporating new judgments, ensuring access to the most current precedents. This results in a large number of judgments and massime so in the vision of making the judicial system more efficient by digitising court proceedings, providing automatized support to the processes can reduce the time and effort required to analyze and summarize them.

2.2. Availability of judgments and *massime* in the two Courts

A fundamental element that affects data availability concerns personal information and privacy. In cases where judgments contain sensitive personal information, access to such data is restricted due to privacy protection laws.

The Italian Constitutional Court, since it is central to the defense of the Constitution, prioritizes the availability of data on its proceedings and decisions. The Court must ensure the integrity and adherence to constitutional principles within the legal system. As a result, inquiries made to the Court generally focus on broad issues that do not involve specific individuals. Consequently, judgments do not contain any personal information and are not subject to privacy-related restrictions. Data of the Italian Constitutional Court are thus open and accessible through its portal[1].

On the other hand, the Supreme Court deals with cases that may involve specific physical or juridic people, which requires compliance with privacy regulations. Consequently, access to its data must be restricted. Information on the proceedings and decisions of the Supreme Court is only accessible through the Italgiure platform[2], which is exclusively available to professionals and legal practitioners. Data cannot be shared, and accesses are controlled and logged. Currently, the dataset selected for the Supreme Court cannot be made public because it would require an expensive anonymization process to ensure privacy.

2.3. The shape of massime

Each legal judgment (also called decision), despite its individuality in terms of case and subject matter, has a shared overall structure. This structure comprises the following key components:

- Heading/Epigrafe: It is the initial part containing the indication of the members of the court, the details of the initiating document, the reporting judge, and the attorneys heard by the Court.
- Statement of Facts: Summarizes the relevant facts of the case, often introduced by "considered in fact and in law.".
- Reasons: It is the section where the Court provides an explanation or argumentation for the conclusions reached in the judgment. This section typically presents the legal principles, factual analysis, and logical reasoning that support the Court's decision.
- Ratio Decidendi: Establishes the binding legal principle or rule derived from the court's decision.

[1] https://dati.cortecostituzionale.it

[2] https://www.italgiure.giustizia.it/

- Disposition: Concludes the decision with the final ruling and any related orders or remedies. It is often introduced by *"P.Q.M."*[3]: It contains the determination of the judges.

Similar to the decisions, the creation of summaries of legal principles, commonly known as *massime*, follows well-defined summarization criteria. As outlined in [10], these *massime* must contain explicit legal references and embody the fundamental principles of law. This detailed approach ensures the effective spread of legal knowledge. *Massime* must meet the following requirements:

- Faithfulness to the decision.
- Conciseness in stating the legal principle.
- Clarity and precision of the stated principle.

Hence, *massima* represents the expression of the legal principle and must not be considered a summary of the decision.

3. The datasets of judgments and *massime*

3.1. Analysis of the *massime* of the two Courts

To better understand how to develop a system for *massime* generation, we analyzed the correlation between the judgments and the *massime* of both courts as they have different roles and consequently deliver different judgments.

For the Supreme Court, we selected a subset of judgments, from 2010 to 2013, to build our dataset useful for the extractive summarization task. During our analysis, we noticed that some decisions may include a *massima* without any text or expressed with an abbreviation, such as *"CONFORME A CASSAZIONE ASN: ..."*. For these cases, we interpret them as references to previous *massime*, but decline to use these specific examples. We started by selecting only the judgments corresponding to at least one *massima*. Indeed, we observed that while most legal judgments of the Supreme Court are tied to a single *massima*, there are a sizable amount of cases in which multiple massime refers to the same judgment (see Tab. 1). Details about how we handled such cases are discussed in the next subsection.

In addition to analyzing judgments from the Supreme Court, we also conducted a systematic analysis of Italian Constitutional Court judgments from 1956 to 2021. We aligned sentences in massime with sentences in the judgments in order to understand how sentences in massime are different from those in the judgments (see Figure 1).

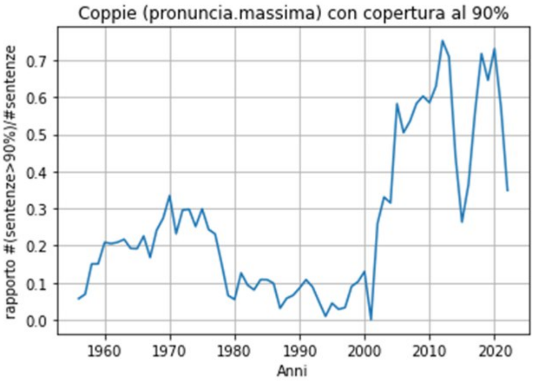

Figure 1: The plot illustrates an increase in "similarity" between the *massima* and pronunciation after year 2000 (with a similarity threshold of 90%) in the Constitutional Court.

According to our analysis, massime of the Constitutional Court became more extractive after 2000. Indeed, in that period, it seems that Constitutional Court Judges forced the "Massimario Office" to avoid changing the text extracted from judgments because even a small change of a single word could significantly alter the overall concept expressed in the judgment. As a result, since 2000, the process of producing massime become an extractive summarization task guided by a topic presented in the last part of the judgment.

3.2. Producing *massime* as a classification task

Summarization is an inherently abstractive task. However, it can be treated as an extractive classification task once the target summary (i.e., a *massima*) is used to select the *relevant* or *irrelevant* sentences from the starting document (i.e., a judgment).

3.2.1. Supreme Court Extractive Data-set

As mentioned before, the first step of the extractive model used to deal with the Supreme Court dataset (see Sec. 2.1) consists in rephrasing a generic *abstractive* summarization dataset into something suitable for a (classical) classification model. This is achieved via the introduction of a *Oracle* meta-model[11].

For each pair (document, summary), all the sentences forming the set with the highest F1 Rouge [12] combination $R_1 + R_2$ concerning the summary are selected and annotated as *relevant*, while all the others are automatically identified as *irrelevant*. This automatically frames

[3] for these reasons

massima per judgment	Supreme Court	Constitutional Court
1	65%	36%
2	24%	26%
3	6%	24%
4	3%	11%
5+	2%	3%

Table 1
The fraction of distinct *massima* per judgment is displayed for both Courts under investigation.

the dataset into a binary classification perspective

$$(document, summary) \overset{\text{Oracle}}{\overbrace{\rightarrow}} (sentence, category)$$

with $category = 0, 1$. As shown in Tab. 1, when a judgment is related with more than one massima, the *Oracle* model acts independently on each judgment-maxima couple, and then the annotated sentences are merged together without repetitions. This is done because otherwise, it can most likely happen that in a multiple *massima* scenario, the same sentence in a judgment is related only with one *massima*, ending up with the same sentence annotated with opposite categories.

Starting from a dataset corresponding to 12000 couples of $(massime, judgments)$ we decided to keep only the data corresponding to at an Oracle rouge of $R_1 + R_2 \geq 0.55$, reducing our training data almost by half (6.849). We observed that, given the nature of the judgment, the number of relevant sentences in any judgment is a very small portion, inevitably producing a highly unbalanced dataset toward the *irrelevant* sentences.

3.2.2. Constitutional Court Extractive Data-set

Given the analysis of the *massime* and the related judgments of the Constitutional Court, we decided to define the task of producing *massime* as the classification task of selecting the appropriate sentences of the judgment given a target topic. The classification dataset is then built as follows, starting from the judgments and the related *massime*. For each judgment, we extracted the points of its operative part (*punti del dispositivo*). For each point, we selected the correlated *massima*. Then, we divided the judgments into sentences and produced a set of triples:

$$(sentence, point, in\_massima)$$

where $sentence$ is a sentence of the judgment, $point$ is a point of its operative part, and $in\_massima$ is True if the sentence overlaps for more than 90% with a sentence in the massima related to the $point$.

For our experiments, we extracted a subset of 40,000 data points from this expanded dataset. The selection process ensured a balanced distribution between positive and negative examples, maintaining a 50/50 ratio. It is important to note that the specific details and steps of the method used to derive the larger dataset from the original 14,316 rows are not provided in this paper. However, this method facilitated a focused analysis of the textual components, shedding light on the connections between phrases, device points, and the formation of massime.

4. Models

Our main challenge was identifying the most relevant parts of pronouncements to assist the massima producer in crafting legal maxims.

As mentioned before, extractive summarization models treat the task of automatic summarization generation as a straightforward sentence classification task. In this vision, the summary of a given document emerges by the concatenation of all the most relevant document *fragments* (i.e., sentences or sub-sentences) classified by the model, this could effectively provide the Massimario with the essential subparts of pronouncements for massima construction.

Both models proposed in the current work are essentially based on a deep encoder which maps the fragments to a vector representation in a high dimensional space subsequently classified into two classes: *relevant* sentences (i.e., candidates for the summary) or *irrelevant* sentences (i.e., not containing relevant information for the summary).

4.1. Supreme Court Model

Data-sets with very long documents (as the one introduced in Sec. 2.1) are usually difficult to handle using a BERT-based [2] transformer encoder. The well known self-attention (introduced in [13]) which characterizes most of the transformer networks is plagued by a fast scaling of computational and memory requirements with the input sequence. Instead of proceeding with a more memory-efficient attention implementation (for instance, see [14]), we decided to act on data and restrict the context length. In this perspective, we introduced a fixed

Court	Prec	Rec	F_1	R_1	R_2	R_3	$\bar{R}_1$	$\bar{R}_2$	$\bar{R}_3$	$\tilde{R}_1$	$\tilde{R}_2$	$\tilde{R}_3$
Supreme	0.40	0.31	0.35	0.47	0.32	0.28	0.64	0.52	0.50	1.97(8)	3.69(35)	4.45(54)
Constitutional	0.53	0.80	0.52	0.32	0.29	0.24	-	-	-	-	-	-

Table 2
Classification and Coverage results of the two models considered. Normalized values with respect to the *Oracle* coverage $\bar{R}_n = R_n/R_n^{\text{Oracle}}$ and random baseline $\tilde{R}_n = R_n/R_n^{\text{Random}}$ where sentences are extracted with the same frequency as in the train set.

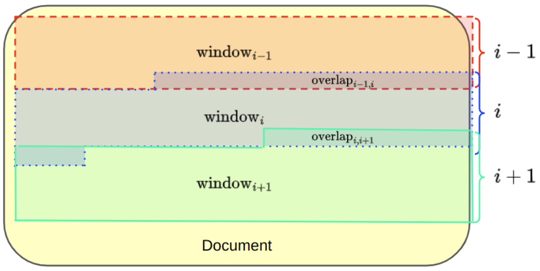

Figure 2: Sketch of the sliding window procedure with overlap. The shaded area in different colors corresponds to different windows (i.e., contexts).

length *sliding window* similar to the implementation described in [15]).

Our goal was to optimize the context length and mitigate context-truncation effects. To achieve this, we defined the context window based on word-pieces and introduced the possibility of overlapping windows up to a maximum number of sentences. The latter, while still under investigation, offers an intriguing tool to probe the context effects on the model. Even in the simplest implementation, with only one overlapping sentence (see Fig. 2), it is interesting to see the effect of a preceding or subsequent context on the same sentence.

The model used, with the aforementioned modification in the document pre-processing, is based on the one proposed in [16, 3] referred to as *BERTSUM*[4]. It is worth mentioning that while in their works, the predicted probability is used only as a ranking score and a fixed number of sentences are extracted neglecting the actual probabilities, we select the relevant sentences accordingly to a cutoff parameter. The latter is a necessary introduction since we observed that, in our data-set, it is common to have documents with a clear separation between sentences that are very likely to be extracted compared to others whose predicted probability is minimal. Therefore, fixing the number of extracted sentences introduces a strong bias toward the extraction of irrelevant sentences.

We observe that the model does not seem to reach high performances both in terms of absolute and relative scores, normalized with the *Oracle* Rouge scores, which are the maximum scores such a model can achieve (see Tab. 2). This is partially due to the violent class unbalance present in the dataset, even if marginally mitigated by the introduction of a weighted loss, with weights inversely proportional to the category frequency in the train set. As a baseline comparison, we decided to include the scores normalized with the one of a random classifier to assess that no random classifications are being performed.

4.2. Constitutional Court Model

The challenge lay in selecting the most useful subparts of legal judgments for the purpose of the massima producer. We sought to leverage BERT[2] to provide the best subparts of pronouncements to aid in massima construction. However, we soon realized that the task was exceptionally complex, requiring the ability to summarize and generalize the text in a unique manner.

To address the above multifaceted challenge, we focused on using BERT to assist us in identifying the most relevant subparts of legal judgment. Through this approach, we aimed to equip the massima producer with essential tools for constructing the maxim more effectively. While our efforts resulted in the development of a tool to assist the massima, producer, we must acknowledge that the results achieved with BERT were not as remarkable as initially hoped. The complexity of the problem, combining the tasks of summarization and generalization in a unique manner, presented formidable hurdles.

Nevertheless, we view this endeavor as a stepping stone toward understanding and tackling the intricacies of legal text processing. Our tool, despite its limitations, serves as a valuable resource for the Massimario, aiding them in the maxim construction process. We recognize that further research and advancements in natural language processing will be crucial in making substantial strides in this domain.

Even in this case, results are interesting but not yet satisfactory (see Tab. 2). Indeed, R1, R2, and R3 are 0.32,

[4] sometimes named *BertSumExt* in literature.

0.29, and 0.24 respectively. This suggests that the task of producing massime is indeed a challenging task.

5. Conclusions

In conclusion, while we did not achieve outstanding results, our efforts shed light on the intricacies of this challenging problem. During our analysis, we also noticed notable differences between the two courts, which further emphasizes the complexity of generating accurate *massime*. We find it particularly intriguing to explore the factors contributing to these variations and understand how they impact the summarization process. Despite the challenges, we remain committed to refining our approach and exploring innovative techniques. Recent advances in the field further motivate us to seek a proper solution that addresses data privacy concerns and significantly improves the task of summarization in the legal field for the Italian language.

Acknowledgments

This work was conducted within the *DATALAKE Giustizia project*; we acknowledge the partners and the scientific committee for their support.

References

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018. arXiv:1802.05365.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 2019. URL: https://arxiv.org/abs/1908.08345. doi:10.48550/ARXIV.1908.08345.

[4] X. Jin, D. Zhang, H. Zhu, W. Xiao, S.-W. Li, X. Wei, A. Arnold, X. Ren, Lifelong pretraining: Continually adapting language models to emerging corpora, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 1–16. URL: https://aclanthology.org/2022.bigscience-1.1. doi:10.18653/v1/2022.bigscience-1.1.

[5] E. Bauer, D. Stammbach, N. Gu, E. Ash, Legal extractive summarization of u.s. court opinions, 2023.

[6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[7] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, N. Aletras, LexGLUE: A benchmark dataset for legal language understanding in English, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4310–4330. URL: https://aclanthology.org/2022.acl-long.297. doi:10.18653/v1/2022.acl-long.297.

[8] M. Cherubini, F. Romano, A. Bolioli, N. De Francesco, I. Benedetto, Summarization di testi giuridici: una sperimentazione con gpt-3, Rivista Italiana di Informatica e Diritto (2023). doi:10.32091/RIID0103.

[9] F. M. Zanzotto, Viewpoint: Human-in-the-loop artificial intelligence, Journal of Artificial Intelligence Research 64 (2019) 243–252. URL: https://doi.org/10.1613%2Fjair.1.11345. doi:10.1613/jair.1.11345.

[10] F. Costantini, P. D'Ovidio, Sintesi dei criteri della massimazione civile e penale, https://www.cortedicassazione.it/cassazioneresources/resources/cms/documents/SINTESI_CRITERI_DELLA_MASSIMAZIONE_CIVILE_E_PENALE.pdf, 2023.

[11] C.-Y. Lin, E. Hovy, The potential and limitations of automatic sentence extraction for summarization, in: Proceedings of the HLT-NAACL 03 Text Summarization Workshop, 2003, pp. 73–80. URL: https://aclanthology.org/W03-0510.

[12] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[14] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. arXiv:2004.05150.

[15] Q. Grail, J. Perez, E. Gaussier, Globalizing BERT-

based transformer architectures for long document summarization, 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. 1792–1810. URL: https://aclanthology.org/2021.eacl-main.154. doi:10.18653/v1/2021.eacl-main.154.

[16] Y. Liu, Fine-tune bert for extractive summarization, 2019. URL: https://arxiv.org/abs/1903.10318. doi:10.48550/ARXIV.1903.10318.