Automatic Summarization of Legal Texts, Extractive Summarization using LLMs

David Preti[1], Cristina Giannone[1,*], Andrea Favalli[1] and Raniero Romagnoli[1]

[1]Almawave S.P.A., via Casal Boccone 10, Roma, 00133, Italy

Abstract

In this work, we describe the first results of experimentation with summarization systems based on large language models to produce an extractive summarization of the judgments (*massime*). We propose a novel approach for this task, exploiting the generative capabilities of LLM and removing all possibilities of hallucination. Our study aims to assess the effectiveness and efficiency of generative models in summarizing the court's decisions. Through a comprehensive analysis of several summarization system setups, we evaluate the quality of the summaries generated by each approach and their ability to capture the key legal principles and linguistic features in the courts' decisions.

Keywords

Legal Text, Summarization, LLM, Generative AI, Human in the Loop

1. Introduction

Artificial intelligence systems, now employed across a wide array of fields, can also serve as valuable aids for legal practitioners.

Increasingly sophisticated tools enhance information search capabilities, automate the drafting or verification of legal documents, and facilitate technical evaluations, such as predictive justice. Utilizing such tools can yield significant benefits by enhancing the efficiency and quality of legal processes. In civil and common law systems, accessing legal judgments to retrieve legal decisions is essential for various legal tasks, including defending clients, constructing cases for prosecution, and issuing judicial decisions. In Italy, to ensure widespread information on the courts' decisions, for this purpose, a dedicated body, the *Ufficio del Massimario*, was established, whose purpose is to produce *massime*.

In a concise yet detailed manner, these summaries (*massime*) encapsulate the legal principles articulated in judgments. Hence, legal professionals can consult these *massime* instead of delving into the entirety of legal decisions. The task of summarising legal texts and producing *massime* has been widely addressed in the last years [1], especially with the advent of the Generative AI [2, 3].

Given the complexity of the task, the approach outlined in [1] focuses on handling the automatic production of a *massima* as an extractive summarization task. This involves extracting the most pertinent part of the judgement to assist in the drafting of the *massima* by the designated office, utilizing a human-in-the-loop approach as discussed in [4].

The process of analyzing judgments and extracting relevant sentences can be significantly simplified through the use of pre-trained models [5, 6]. These models function as versatile universal sentence/text encoders, capable of addressing a range of downstream tasks, including summarization [7]. These models consistently outperform other approaches, particularly after fine-tuning or domain-adaptation [8].

Despite the success of pre-trained transformers and LLMs in other summarization tasks[9], certain phenomena, such as hallucination in the generation of the text [10], the task of producing *massime* is still challenging for current extractive and abstractive summarization systems. 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 [11].

In this paper, we present an approach to producing an extractive summary by exploiting the ability of an LLM to generate abstract summaries from a document. Our approach selects, from the abstract, the sentences that best match the sentences in the source document. This approach, described in Sec. 2, reduces the hallucination phenomena, achieving results in a zero-shot setting, described in Sec. 3 comparable with a model trained with a domain dataset.

2. Extractive Projection

It is well known that generative models, particularly when used in summarization systems, are prone to hallucination phenomena (see [12] and references therein). In this case, new terms or, in worst scenarios, even informa-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

d.preti@almawave.it (D. Preti); c.giannone@almawave.it (C. Giannone); a.favalli@almawave.it (A. Favalli); r.romagnoli@almawave.it (R. Romagnoli)

Model	$ROUGE_1$	$ROUGE_2$	$ROUGE_3$
`Oracle`	0.81	0.71	0.65
`Ext`	0.40	0.30	0.28
`Abs`(p_1)	0.32	0.10	0.05
`Gen-Ext`(p_1)	0.31	0.12	0.08
`Abs`(p_2)	0.35	0.13	0.07
`Gen-Ext`(p_2)	0.38	0.20	0.16

Table 1
Mean $ROUGE_n$-f_1 scores computed on test data for different models. `Ext` is an extractive model trained on the `Oracle`. `Gen-Ext` and `Abs` are the models based on pure abstractive summarization with and w/o extractive projection respectively. Results with different prompts p_1 (generic summarization prompt) and p_2 (domain tuned summarization prompt) are also displayed explicitly in Tab. 3

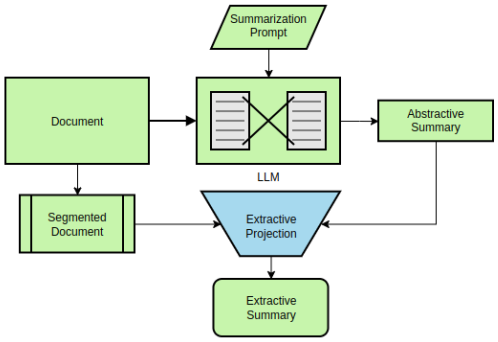

Figure 1: Sketch of the extractive summarization system proposed in this work.

tion and facts not present in the original document are generated in the output summary. Several attempts have been made to try to tame such unwanted behaviour (for instance, see [13]), which may lead to serious problems in sensitive domains.

Given its specific lexicon, the vast amount of fixed forms and judicial references, the legal domain is very delicate and unsuitable for a straightforward application of generative systems. To overcome such a problem, we introduce what we refer to as *extractive projection*, meaning, a transformation mapping a generated text into sentences of the original document.

Defining the source documents $d \in D$, the abstractive summary as $a \in A$ with D, A respectively the space of documents and abstractive summaries, the summarization prompt $p \in P$. The *generative summarization* transformation is defined as:

$$G : D \times P \to A$$
$$a = G(d|p)\,. \tag{1}$$

We introduce the *extractive summary* as $a' \in A' \subset A$, and the extractive projection Γ

$$\Gamma : \tilde{D} \times A \to A'$$
$$a' = \Gamma(\tilde{d}; a)\,, \tag{2}$$

where $\tilde{d} \in \tilde{D}$, and $\tilde{D}$ is the space of segmented documents (i.e., containing the same documents as D, but each one is split into a set of segments).

The projection Γ used in this work is a slightly modified version of the algorithm proposed in [7] to pre-process the data. As a main difference from [7] we allow the algorithm to select up to all the segments present in the document, without any parameter fixed a priori. Moreover, while in [7] this greedy selection algorithm is used to obtain an oracle summary for each document used as a reference to train the extractive model, here this algorithm is used to *project* the (abstractive) generated summary into *the segments* of the original document. Note that such procedure completely removes *by construction* any possibility of hallucination since the projection cuts off all possible novelties and generations.

The greedy selection procedure employed is then simply a combinatorial optimization algorithm based on *coverage* metrics. In this respect, we tested several metrics, ranging from the average of *ROUGE-1* and *ROUGE-2* [14] as originally proposed in [7] to different linear combinations of *Rouge-n* and more sophisticated similarity metrics (e.g., *BERTscore* [15].

We observe that with the exception of very rare cases where the generated summary is produced in a different language with respect to the original document, all the coverage metrics produce accurate results (see Tab. 1). In the multilingual setup, only a similarity metric based on multilingual embeddings, which is insensitive to language shifts, produces reasonable results, while *ROUGE* does not work correctly.

3. Results

As discussed in Sec. 1, we trained and tested the extractive summarization systems introduced in [1] on a dataset composed by judgments and *massime* from different courts[1]. Starting from a whole dataset of 1340 couple of (judgement, *massima*), we randomly selected a subset of 199 of them as a *validation* set, 940 as a *train* set, and the remnant 201 as *test* set. The latter has been further refined down to 61 "high quality" examples. For such

[1]The data are publicly available on the website https://www.inps.it/it/it/inps-comunica/atti/sentenze.html

Prompt	Text
p_1	Write a summary in Italian of 150 words of the following text delimited by triple backquote: “‘content“‘
p_2	Scrivi una massima in Italiano di 150 parole della seguente porzione di testo delimitata dalle virgolette. La massima deve rispondere ai seguenti generali requisiti: a) fedeltà alla decisione; b) sintesi nell’enunciazione del principio; c) chiarezza e precisione del principio enunciato La massima costituisce l’enucleazione del principio di diritto e non il riassunto della decisione e non può tradursi nella mera riproduzione di passaggi argomentativi della motivazione. “‘content“‘

Table 2
Prompts used for generic summarization (p_1) and domain tuned task summarization (p_2).

selection, we first used the greedy algorithm proposed in Sec. 2 based on the average of $ROUGE_1$ and $ROUGE_2$ and then selected only data with that value larger or equal to 0.6 (see Fig. 2). The scores for the extractive model `Ext`, compared with the `Oracle` and those produced using generative models, are collected in Tab. 1. More specifically, we used two different prompts p_1 and p_2 (for details see Tab. 3) to estimate the effect of a "generic" summarization prompt, with a "task tuned" prompt specifically referring to the features of a *massima* [16]. As expected, we observe a small improvement in scores with all generative models using p_2 over p_1. Moreover, we compare the scores of a straightforward abstractive summarization `Abs`, with the setup proposed in this work, i.e., including the *extractive projection* called `Gen-Ext` in Tab. 1. For all the evaluations, we used a generative model of the *gpt-turbo* [17] family[2]. Interestingly, the scores obtained using zero-shots (no fine-tuning or contextual examples are involved) generative models, in both their types: abstractive (`Abs`) and extractive (`Gen-Ext`), seem to perform reasonably well when compared to the `Ext` model. An example of the summaries produced in all the setups are displayed in Tab. 3.
It is worth noting that the scores obtained in this work should be interpreted only as a reference. They are affected by large statistical fluctuations, which make a direct comparison among the scores very tricky. Moreover, coverage scores are known to have a limited correlation with the effective quality of the summary produced, which requires some human evaluation by domain experts.

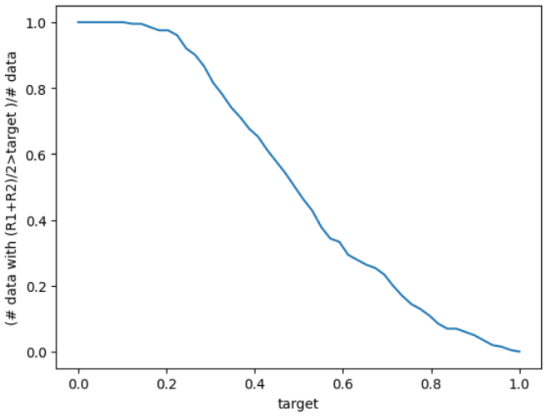

Figure 2: Fraction of test data as a function of the score $(ROUGE_1 + ROUGE_2)/2$ computed on the segments extracted by the `oracle` combinatorial algorithm.

4. Conclusions

In this work we discussed the first results of a novel approach that can be used to obtain "hallucination"-free results out of a generative model. We applied such procedure in a legal domain, where preserving factuality is mandatory.
While obtain only partial results, we find them to be reasonably promising but requiring some further investigation.
A comparison with different "open source" LLMs as generative models, estimating the parameter scaling effects on the performances and a complete or partial (see for instance [18]) fine-tuning or domain-adaptation is dedicated to future studies. In conclusion, while discounting the difficulty of the task, given both by the inherent complexity of the structure of a *massima* that cannot be treated as a simple summary, and by the difficult evaluation of the results found, as well as the fact that automatic "token"-coverage metrics require some evaluation by human domain experts, we believe that LLMs, if appropriately applied, can offer a valuable tool even in domains where factuality is paramount.

[2] gpt-3.5-turbo-1106.

Summary	Text
Target	L'inserzione automatica di clausole, prevista dall'art. 1339 cod. civ., costituisce una restrizione significativa del diritto di libertà economica consacrato dall'art. 41 Cost. di cui è espressione l'autonomia privata, e deve quindi trovare il suo fondamento in una legge formale o in un altro atto avente valore di legge in senso sostanziale o da esso richiamato tramite rinvio integrativo; pertanto, con riferimento al contratto avente ad oggetto la prestazione di servizi di vigilanza da parte di un istituto privato, tale integrazione non può aver luogo in base al decreto con cui il prefetto approva la relativa tariffa ai sensi dell'art. 135 del r.d. n. 773 del 1931, trattandosi di un mero atto amministrativo espressivo di un sindacato di congruità, avente natura di merito, del prezzo di una prestazione contrattuale, né l'ingerenza con efficacia reale sulla tariffa proposta dall'Istituto e liberamente accettata dal committente può trovare fondamento nella generale potestà del prefetto di impartire prescrizioni nell'interesse pubblico, in sede di rilascio di autorizzazioni di polizia, ai sensi dell'art. 9 del r.d. n. 773 cit..
Abs(p_1)	L'Istituto cooperativo di V. ed I. C. s.r.l. ha citato in giudizio l'Istituto Nazionale Previdenza Sociale per ottenere il pagamento di una somma di denaro a titolo di maggior corrispettivo per il servizio di vigilanza. Dopo un lungo iter giudiziario, la Corte d'appello di Catanzaro ha respinto la domanda, sostenendo che il provvedimento prefettizio in materia di tariffe per la vigilanza non poteva derogare al contenuto contrattuale liberamente scelto dalle parti. Il fallimento dell'Istituto Cooperativo di V. e I. C. ha proposto ricorso per cassazione, ma la Corte ha respinto il ricorso, condannando al pagamento delle spese di giudizio. La Corte ha sostenuto che l'inserzione automatica di clausole ex art. 1339 c.c. riguardava solo i prezzi imposti con legge e che l'ingerenza con efficacia reale sulla tariffa proposta dall'istituto di vigilanza e liberamente accettata dal committente non poteva trovare fondamento nella generale potestà di impartire prescrizioni nell'interesse pubblico.
Abs(p_2)	L'autonomia privata non può essere derogata da atti amministrativi, ma solo da leggi formali. Il principio dell'autonomia privata è fondamentale e non può essere interpretato estensivamente. Le tariffe prefettizie non hanno efficacia normativa e non possono essere conferite da una presunta convergenza di vedute delle parti. Inoltre, il dovere di solidarietà sociale non può essere utilizzato come motivo di condanna al pagamento del maggior prezzo. La libertà economica può subire restrizioni solo in forza di una disposizione di legge. L'inserzione automatica di clausole ex art. 1339 c.c. riguarda solo i prezzi imposti con legge e non può essere derogata da atti amministrativi. La fedeltà alla decisione giuridica richiede il rispetto di tali principi.
Gen-Ext(p_1)	- con atto di citazione notificato il 5 luglio 1995 l'istituto cooperativo di v. ed i. c. s.r.l. conveniva dinanzi al tribunale di lamezia terme l'istituto nazionale previdenza sociale per ottenerne la condanna al pagamento della somma di l. 639.590.300 a titolo di maggior corrispettivo per il servizio di vigilanza espletato in favore dell'ente, presso la sede di lamezia, dovuto in forza dell'inserzione automatica della tariffa inderogabile emanata dal prefetto di catanzaro in materia di vigilanza, superiore al prezzo pattuito con il contratto stipulato il 25 agosto 1983. dopo l'espletamento di consulenza tecnica d'ufficio il tribunale di lamezia terme con sentenza 23-27 novembre 1999 condannava l'inps al pagamento della somma di l. 1.148.787.862, oltre le spese di giudizio. avverso la sentenza, non notificata, proponeva ricorso per cassazione il fallimento dell'istituto cooperativo di v. e i. 157 del relativo regolamento di esecuzione, nonché l'art. 1175 e 1375 cod. civ. e dell'art. diritto - con il primo motivo il ricorrente deduce la violazione degli artt. 9, 134 e 135 del citato testo unico delle leggi di pubblica sicurezza e dell'art. 1339 cod. civile. con il secondo motivo ricorrente censura l'omessa motivazione nel discostarsi dalla concorde interpretazione delle parti. con l'ultimo motivo il fallimento deduce la violazione degli artt. 2 della costituzione.
Gen-Ext(p_2)	1339, cod. civ. 1339 cod. civ. l'inserzione automatica di clausole, prevista dall'art. 1339, cod. civ., costituisce una deroga incisiva al principio dell'autonomia privata e deve quindi trovare il suo fondamento in una legge formale - come testualmente previsto dalla norma - o in altro atto avente valore di legge in senso sostanziale o da esso richiamato tramite rinvio integrativo. il diritto di libertà economica consacrato dall'art. 41 cost., di cui è espressione l'autonomia negoziale delle parti nel modellare il contenuto di un contratto, può soffrire restrizioni solo in forza di una disposizione di legge, insuscettibile di interpretazioni estensive (ibidem, terzo comma). ne consegue la vigenza, in subiecta materia, di un principio di stretta interpretazione dell'art. civ. ; vieppiù giustificato da esigenze di tutela della concorrenza e del mercato, che verrebbero lese da una pratica di prezzi amministrati. l'asserita convergenza di vedute sull'efficacia cogente delle tariffe prefettizie non può, neanche in astratto, valere a conferire loro l'efficacia normativa di cui sono intrinsecamente prive. l'invocazione di un inderogabile dovere di solidarietà sociale che avrebbe imposto la maggiorazione del prezzo non ha, infatti, alcuna attinenza con l'operatività dell'eterointegrazione ex art.

Table 3
Example summary comparison among the various summarization systems proposed.

References

[1] F. Achena, D. Preti, D. Venditti, L. Ranaldi, C. Giannone, F. M. Zanzotto, A. Favalli, R. Romagnoli, Legal summarization: to each court its own model, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023, volume 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3596/paper1.pdf.

[2] T. Dal Pont, F. Galli, A. Loreggia, G. Pisano, R. Rovatti, G. Sartor, Legal Summarisation through LLMs: The PRODIGIT Project, arXiv e-prints (2023) arXiv:2308.04416. doi:10.48550/arXiv.2308.04416. arXiv:2308.04416.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking Large Language Models for News Summarization, Transactions of the Association for Computational Linguistics 12 (2024) 39–57. URL: https://doi.org/10.1162/tacl_a_00632. doi:10.1162/tacl_a_00632.

[10] D. de Vargas Feijó, V. P. Moreira, Improving abstractive summarization of legal rulings through textual entailment, Artif. Intell. Law 31 (2023) 91–113. URL: https://doi.org/10.1007/s10506-021-09305-4. doi:10.1007/S10506-021-09305-4.

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

[12] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. arXiv:2311.05232.

[13] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating hallucination in large language models via self-reflection, 2023. arXiv:2310.06271.

[14] 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.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, CoRR abs/1904.09675 (2019). URL: http://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[16] C. di Cassazione, Sintesi criteri della massimazione civile e penale (2024). URL: https://www.cortedicassazione.it/resources/cms/documents/SINTESI_CRITERI_DELLA_MASSIMAZIONE_CIVILE_E_PENALE.pdf.

[17] OpenAI, Gpt-3.5-turbo-1106 large language model (2023).

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.