ercim-news.ercim.eu Number 141 April 2025

ERCIM NEWS

Special theme:

AI for Cultural Heritage

Automatic Extraction of Regesta for Medieval Latin Text Summarization

by Giovanni Puccetti (CNR-ISTI), Laura Righi (University of Modena and Reggio Emilia), Ilaria Sabbatini (University of Palermo), and Andrea Esuli (CNR-ISTI)

We produced a novel dataset of 4,533 medieval Latin regesta (summaries) paired with full texts, extracted through a meticulous pipeline involving manual annotation, custom model training, text extraction, and post-processing to ensure high-quality, structured data for AI-driven summarization tasks.

The dataset enables AI-driven summarization of historical documents. We evaluated the performance of Large Language Models (LLMs) in generating Latin regesta, offering insights into the challenges and potential of AI in digital humanities.

The REVERINO dataset (REgesta generation VERsus latIN summarizatiOn) [1], developed under the ITSERR [L1] project, focuses on summarizing medieval Latin documents, particularly pontifical texts from the 13th century. Regesta, concise summaries of historical documents, are essential for scholars studying large collections of latin documents. REVERINO pairs regesta with their corresponding full texts and apparatus (i.e., bibliographic information on the source of the regestum), extracted from two collections: Epistolae saeculi XIII e regestis pontificum Romanorum selectae (MGH, 1216-1268) and Les Registres de Gregoire IX (Auvray, 1227/41).

Generating the REVERINO Dataset

The input data for the process consist of high-resolution images from the pages of the volumes in the two collections. Creating the dataset involved a four-step pipeline:

1. Annotation: Manual annotation of pages by domain experts using the eScriptorium [2] platform to identify and segment text areas and lines. In this phase we simplified the work of experts requesting that they identify the main content of the page. The experts were not required to separate the different parts of the main content (i.e., regesta, apparatus, full text), as this was done automatically in post-processing.
2. Training: The annotated data allowed us to train customized segmentation models for the collections. The models segment the main content, separating it from any additional content (e.g., titles, headers, footers, footnotes, line numbers). Each collection required a specific model, as the MGH collection uses a single-column format, while the Auvray collection uses a more complex two-column format, requiring additional training data to improve segmentation accuracy.
3. Extraction: We used the trained models to segment and perform OCR to extract text from all pages. This step produced a continuous stream of text that includes the regesta, full texts, and apparatus, but without labels to distinguish them.
4. Post-processing: We could define simple heuristics based on content and positional information, leveraging the accurate information regarding text position and size produced by the trained models, to accurately split the text into regesta, apparatus, and full texts, producing the structured information we aimed to obtain.

The final dataset comprises 2,283 pairs from the MGH collection and 2,250 from Auvray. A t-SNE analysis (Figure 2) on the text of regesta revealed distinct clusters for each collection, highlighting their unique characteristics due to differences in topics, sources, and editorial style (t-SNE stands for t-distributed Stochastic Neighbor Embedding, a dimensionality reduction technique commonly used to visualise high-dimensional data in a lower-dimensional space).

Training LLM on Regesta Generation

We used REVERINO as a benchmark for evaluating the ability of LLMs to generate regesta. Three models—GPT-4, Llama 3.1 70b, and Llama 3.1 405b—were tested using two approaches:

1. Direct Summarization (format): The model generates a regestum directly from the Latin full text.

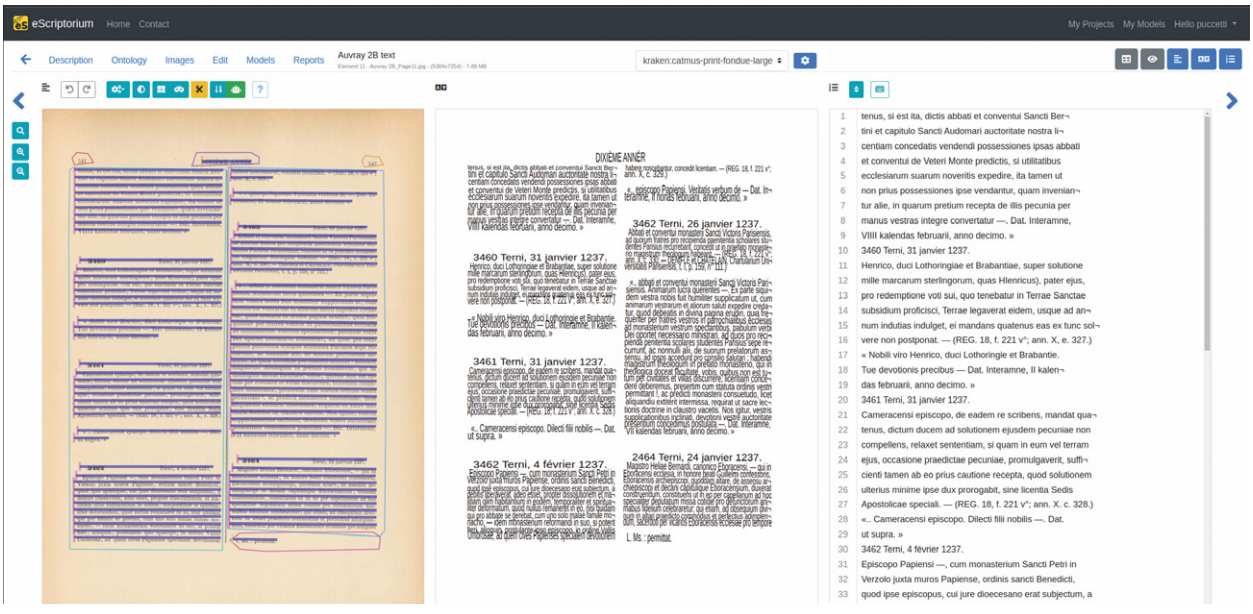

Figure 1: The annotation interface with the segmented regions and the output of the OCR process.

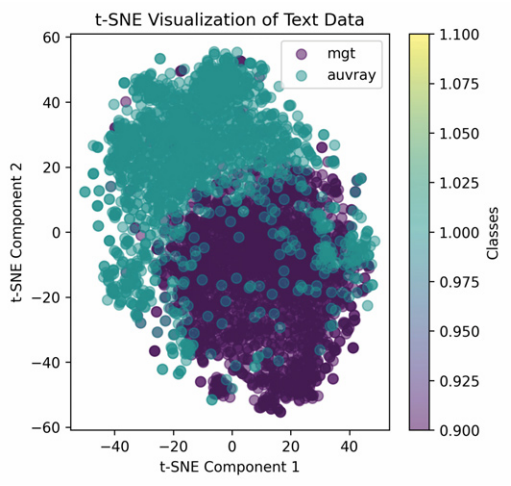

Figure 2: T-SNE plot showing the distribution of samples from the two collections MGH (dark) and Auvray (light).

2. Translation-Summarization (backtranslate): The model first translates the Latin full text into English, generates a summary in English, and then translates it back into Latin.

Quantitative evaluation using Rouge and Bleu metrics showed that GPT-4 outperformed the Llama models, achieving a Rouge-1 score of 0.39 on the MGH part of the dataset. Llama models performed better in the backtranslate setting, suggesting that the forced translation to English exploits the better summarisation capabilities of the model in their main language.

Qualitative analysis by a domain expert revealed challenges in accurately identifying key elements that are expected in regesta, such as the Pope's name, dates, and recipients. For example, GPT-4 correctly identified the Pope in 11 out of 20 cases and the recipient in 15, but only three dates were accurately extracted. These findings highlight the complexity of summarizing historical texts and the need for further refinement.

Conclusions

The REVERINO dataset provides a foundation for training models to automate the summarization of Latin texts, while the evaluation of LLMs offers insights into their strengths and limitations. Future work will expand the dataset to include more collections, further improve custom model training, and refine prompt engineering to improve summarization accuracy.

The ITSERR project is funded by the European Union - NextGenerationEU (CUP B53C22001770006).

Links:

[L1] https://www.itserr.it/
[L2] https://zenodo.org/records/14971613

References:

[1] G. Puccetti et al., "REVERINO: REgesta generation VERsus latIN summarizatiOn", IRCDL 2025.
[2] B. Kiessling, et al., "eScriptorium: an open source platform for historical document analysis", ICDARW, 2019.

Please contact:

Giovanni Puccetti, CNR-ISTI, Italy
giovanni.puccetti@isti.cnr.it

Who Are You, AI? – Artificial Intelligence as a Potential Metaphysical Humiliation of Humanity

by Michael Fleischhacker (University of Applied Sciences Burgenland), Igor Ivkić (University of Applied Sciences Burgenland and Lancaster University), and Friedrich Roithmayr (University of Applied Sciences Burgenland)

Artificial Intelligence as a potential fourth metaphysical humiliation—questioning human uniqueness in reason, freedom, and moral standing, and unsettling our place in the world after Copernicus, Darwin, and Freud.

Sigmund Freud identified three historical humiliations of humanity that have shaken the human sense of self [L1]. The first was the "cosmological humiliation" brought about by Nicolaus Copernicus' discovery that the Earth was not the centre of the universe, thereby depriving humans of their central position in the cosmos. The second humiliation, according to Freud, was the "biological humiliation" introduced by Charles Darwin's insight that humans were a product of evolution and not the creation of a higher being. Freud identified the third humiliation in the realm of the psyche and called it the "psychological humiliation". In this view, the human subconscious undermines the supposed autonomy of the conscious mind: the "ego" is not, as Freud put it, "master in its own house", but is driven by unconscious impulses. These three blows have effectively dethroned humanity from its self-appointed role as the centre and ruler of the world. In this article we explore whether AI is a potential fourth "metaphysical humiliation" of humanity, challenging the special status of humans as unique beings endowed with consciousness and reason, and the exclusivity of humanity as the only free and intelligent beings.

The emergence of Artificial Intelligence (AI) and the potential development of Artificial General Intelligence (AGI) poses a profound challenge to humanity's self-perception, touching on three critical dimensions. First, the domain of intelligence and problem-solving, long considered a hallmark of human uniqueness, is being redefined. AI systems such as AlphaGo and AlphaZero have demonstrated superhuman capabilities in complex tasks. AlphaZero, for example, mastered the strategy game Go by learning solely through playing itself, mimicking processes similar to human intuition and creativity. Such advances undermine the assumption that strategic thinking and intelligence are exclusively human traits. Second, it challenges the concept of consciousness and freedom. Jean-Paul Sartre famously described human beings as "condemned to be free", with freedom arising from consciousness, which allows individuals to shape their own existence. If AI were to develop consciousness and subjective experience, it could fulfil Sartre's notion of freedom, thereby challenging humanity's monopoly on self-determination. This would force us to rethink our understanding of autonomy and concede that AI might also possess the capacity to "make itself", as Sartre described it, dealing a profound blow to humanity's self-conception. Finally, the ethical and ontological supremacy tradition-