Augustus: Enhancing Epigraphic Language Models with POS Tagging, Material Classification, and Generative Capabilities

Federico Gerardi
Sapienza University of Rome
gerardi.1982783@studenti.uniroma1.it

Abstract

Ceriotti, Gerardi, et al. [4] propose a BERT-based model trained on EDR epigraphs using masked language modeling and century classification. In this Homework as Project, I aim to extend their work by incorporating Part-of-Speech (POS) tagging and material type classification to enhance the model's overall performance. Additionally, I will experiment with the T5 model [11] to generate realistic synthetic epigraphs, enabling a generative approach to epigraphic text modeling.

1 Task description/Problem statement

Epigraphs from the Ancient Roman Empire are invaluable historical artifacts, offering insights into significant events, public figures, and official decrees. These inscriptions not only serve as historical records but also as instruments of political propaganda, providing a window into the daily life and cultural practices of the time. However, the passage of centuries has led to significant erosion, resulting in the loss of crucial words and information.

Building upon the work of Ceriotti, Gerardi, et al. [4], which combines masked language modeling with century classification, this project aims to enhance the approach by integrating part-of-speech (POS) tagging and material type classification. Additionally, we will explore the generation of synthetic epigraphs to further improve the model's performance and applicability. These enhancements are designed to provide a more robust and accurate tool for historians and archaeologists, facilitating deeper and more precise analyses of these ancient texts.

Sentence	*valeria marci et liberta philumina. in **fronte** pedes ix, in **agro** pedes xiix.*
Century	Before Fifth Century B.C.
POS	NOUN NOUN CONJ NOUN NOUN PREP NOUN NOUN NUM PREP NOUN NOUN NUM
Material	Marble

Table 1: Output

1.1 Examples

The input is the text of the epigraph:
valeria marci et liberta philumina. in <unk> pedes ix, in <unk> pedes xiix.

The output is showed in Table 1

Considering that the task also includes a generative component, the objective is to produce a completely new epigraph based on a small set of input words. Specifically, the model will take 2–3 keywords as input and generate a full, coherent epigraph as output.

Input: *valeria marci et*

Output: *valeria marci et liberta philumina. in fronte pedes ix, in agro pedes xiix*

1.2 Real-world applications

In real-world scenarios, this tool is highly valuable for archaeologists, as it significantly enhances both the speed and quality of epigraphic analysis. In Ceriotti, Gerardi, et al. [4], we introduced a system that achieved high accuracy in top-10 predictions for reconstructing missing text. This top-10 approach is particularly practical for archaeologists, as it narrows the range of possibilities to a manageable number while preserving analytical depth. In the present work, the objective is to further improve the model's reliability by increasing the accuracy of top-5 and top-1 predictions.

Moreover, epigraph generation offers additional real-world applications. It can be used to synthetically expand epigraphic databases, thereby supporting better model training and improved accuracy. It also serves as a valuable educational tool, enabling students to practice gap-filling tasks on realistic but artificially generated inscriptions, even in cases where no original epigraph exists.

2 Related work

In our previous work, **Ceriotti, Gerardi, et al.**[4], we proposed a novel approach to predicting epigraphic *lacunae* by leveraging masked language modeling combined with century classification. To this end, we fine-tuned **LatinBERT**[1], a transformer-based language model [12] pretrained on a diverse corpus of Latin texts. While LatinBERT has demonstrated strong performance on literary data, its direct application to the epigraphic domain—characterized by fragmentary and formulaic inscriptions—warrants dedicated adaptation and evaluation.

Building upon this foundation, Locaputo et al. [8] outlined a comprehensive research agenda for lacuna restoration in Latin inscriptions, advocating for the use of deep learning techniques such as fine-tuning pretrained models (e.g., LatinBERT) and exploring specialized architectures like Ithaca.

A key resource underpinning our work is the **Epigraphic Database Roma (EDR)** [6], which provides over 82,000 curated Latin inscriptions. This corpus, with its rich metadata and standardized formatting, is essential for training and benchmarking epigraphic NLP models.

The **Classical Language Toolkit (CLTK)** [7] represents a major advancement in natural language processing for ancient languages such as Latin and Greek. It offers a suite of tools including tokenization (used by LatinBERT) and part-of-speech (POS) tagging, both of which are integral to this project. However, a notable limitation of CLTK is its lack of GPU support, which makes large-scale POS tagging, such as processing the 83,000 epigraphs, computationally intensive.

To address this bottleneck, we also consider the use of **Stanza** [10], a neural NLP library developed by Stanford that includes support for Latin. Stanza offers GPU acceleration and neural pipeline components, making it a suitable and scalable alternative for high-volume epigraphic annotation tasks.

3 Datasets and benchmarks

Our dataset is the **Epigraphic Database Roma (EDR)**[6], which contains over 82,000 curated Latin inscriptions, making it one of the largest resources of its kind in the world. The inscriptions are richly annotated by archaeologists, including elements such as parentheses, expansions, and editorial suggestions. In Ceriotti, Gerardi, et al.[4] we applied a comprehensive cleaning process to this corpus, involving case folding and the removal of archaeological annotations. This normalization step was crucial to preserve only the raw textual content and to reduce potential biases in the language modeling process.

There are also other epigraphic datasets available, such as the Epigraphische Datenbank Heidelberg (EDH) [5], which contains a large collection of Latin inscriptions from the Roman Empire, particularly from the provinces. However, in our work, we focus exclusively on Ancient Roman Latin inscriptions—that is, inscriptions originating from Rome itself or its immediate cultural sphere. This choice is made to ensure linguistic and stylistic consistency and to avoid introducing regional variations or external influences that could affect the quality and reliability of our analysis.

4 Existing tools, libraries, papers with code

The Classical Language Toolkit (CLTK) [7] is the primary framework for processing ancient languages, offering a wide range of tools specifically designed for this domain. In this report, CLTK plays a central role, as it represents the standard toolkit for addressing challenges in Ancient Language Processing. Notably, its tokenizer is employed in the training pipeline of LatinBERT [1].

LatinBERT [1] is a transformer-based [12] language model trained on a diverse Latin corpus, including classical literature and Latin Wikipedia, making it well-suited for tasks involving historical Latin texts.

PyTorch [9] is the deep learning framework used for model implementation and training, providing the flexibility and efficiency required for fine-tuning and experimenting with neural architectures such as LatinBERT.

Stanza [10] is a neural NLP library developed by Stanford, offering support for Latin and GPU acceleration. It plays a crucial role in this project, as the POS tagging of the dataset is generated using this library.

5 State-of-the-art evaluation

Ceriotti, Gerardi, et al. [4] evaluated their **Masked Language Modeling** task using Perplexity, Top-1 Accuracy, Top-5 Accuracy, and Top-10 Accuracy.

In line with the recommendations by Celikyilmaz et al. [3], the task of Epigraph Generation requires *expert-centric human evaluation*, as only trained Epigraphers and Latinists can properly assess the output's historical authenticity, lexical appropriateness, and grammatical correctness.

6 Comparative evaluation

6.1 Dataset

For this work, I will use the same dataset presented in Ceriotti, Gerardi, et al. [4], as it represents the largest and most comprehensive epigraphic dataset currently available. It has been carefully curated and normalized, making it particularly well-suited for training language models on ancient Latin inscriptions.

The dataset will be improved adding the POS Tagging to each epigraph evaluating it with the Stanza Toolkit by Stanford University.

6.2 System

The proposed system builds upon the approach introduced in Ceriotti, Gerardi, et al. [4], which fine-tunes the LatinBERT base model by modifying the training objective to include both Century Classification and Masked Language Modeling (MLM). In this work, I extend that system by incorporating two additional objectives into the loss function: Material Classification and Part-of-Speech (POS) Tagging.

By jointly optimizing for Century, Material, and POS classification tasks alongside MLM, the model is encouraged to learn richer representations of the input inscriptions. This multi-task learning setup aims to enhance the model's understanding of linguistic and contextual patterns in the data, thereby improving its performance on the primary MLM task.

PPL	Acc@1	Acc@5	Acc@10
5.7667	0.697	0.8	0.84

Table 2: Masked Language Modeling Results

	Acc	P	R	F1
Date	0.6971	0.6864	0.6971	0.6896
Material	0.6212	0.5284	0.6212	0.5554
POS	0.8874	0.8872	0.8874	0.8866

Table 3: Other Tasks Results

The overall loss function for the multi-task learning setup [2] is a weighted sum of the individual losses from the 4 tasks:

$$\begin{aligned}\mathcal{L}_{\text{Total}} = {} & \mathcal{L}_{\text{Masked Language Modeling}} \\ & + \lambda_1 \cdot \mathcal{L}_{\text{Date Classification}} \\ & + \lambda_2 \cdot \mathcal{L}_{\text{Material Classification}} \\ & + \lambda_3 \cdot \mathcal{L}_{\text{POS Tagging}} \end{aligned} \tag{1}$$

The individual task losses are combined using a weighted sum, where the λ coefficients were set based on the hypothesized contribution of each auxiliary task to the primary MLM objective. The MLM loss retains an implicit weight of **1.0**. For the auxiliary tasks, the weights were strategically chosen: POS Tagging (λ_3) was assigned the highest weight of **0.7**, as its token-level linguistic signal is most directly relevant to MLM. Date Classification (λ_1) was given a moderate weight of **0.5** for its sequence-level contextual information. Finally, Material Classification (λ_2) received the lowest weight of **0.2**, de-prioritizing this potentially less correlated signal to prevent it from unduly influencing the model's training on more critical linguistic patterns.

For the generative task, I will adopt an encoder-decoder architecture by pairing the fine-tuned LatinBERT encoder with a T5 decoder [11]. This configuration enables the model to generate complete epigraphs from minimal prompts, such as two or three input words.

6.3 Results

After conducting ten epochs of training on an NVIDIA RTX 4070, we observe an improvement in the Masked Language Modeling (MLM) task, with accuracy@1 increasing by 0.06 compared to the results reported by Ceriotti, Gerardi, et al. [4] with a similar increase also in Top-5 and Top-10, and a reduction in Perplexity by 3 points. Regarding the other tasks, the performance on Date

Classification remains comparable to that of Ceriotti, Gerardi, et al. [4]. In contrast, Material Classification showed limited progress: although the loss dropped significantly in the initial epochs, it plateaued early and did not improve further. Part-of-Speech (POS) Tagging achieved notably strong results, demonstrating high accuracy and robustness.

After conducting three epochs of training on an NVIDIA RTX 4070, the results for epigraph generation were generally underwhelming. The model demonstrated proficiency in producing coherent outputs when provided with well-structured prompts. However, it struggled with more challenging or ambiguous inputs, often generating incoherent or random words.

For instance, consider the following well-structured prompt:

Input: *aurelius <extra_id_0>vixit annis <extra_id_1>*

Output: *saturninus xxxv.*

This example yields a satisfactory result. In contrast, the model's performance deteriorates with less structured inputs, as illustrated below:

Input: *iulia filia*

Output: *iulia filia cai filia. posuit. secundus. cnaei. suae. vn.*

In this case, the output appears to be a jumble of random words, highlighting the model's limitations with ambiguous prompts.

6.4 Discussion

The primary limitation of the baseline system is its exclusive reliance on textual content and date information. In contrast, our approach incorporates Part-of-Speech (POS) tagging, which has yielded impressive results and shows strong potential to further improve model accuracy. Notably, these gains were achieved with only 10 training epochs, constrained by limited time.

The task of generating new epigraphs remains largely unexplored in the current literature, preventing direct comparisons with prior work. However, this area holds great promise for future research, offering opportunities to expand existing datasets and substantially improve epigraphic modeling performance. The current lack of strong results is likely due to the dataset being heavily inflated with unknown tokens which constitute 40% of the dataset, and poses significant challenges.

7 Conclusions

In this project, I built upon the multi-task learning framework introduced by Ceriotti, Gerardi, et al. [4] by integrating Material Classification and Part-of-Speech (POS) Tagging. Among these additional tasks, POS Tagging emerged as particularly advantageous, substantially enhancing overall performance.

Furthermore, I investigated the fine-tuning of a T5 model for epigraph generation, with the goal of expanding the existing dataset and facilitating downstream tasks.

Looking ahead, I propose the implementation of K-Fold cross-validation combined with grid search to optimize hyperparameters, thereby improving Masked Language Modeling (MLM) task performance. Additionally, a more thorough exploration of epigraph generation—conducted in collaboration with Latin linguists and archaeologists—could uncover novel opportunities for data augmentation and historical analysis. Such partnerships would enable archaeologists to leverage the MLM model for dataset refinement, ultimately fostering the generation of more accurate and meaningful epigraphs.

References

[1] David Bamman and Patrick J. Burns. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv preprint arXiv:2009.10053*, September 2020. 2

[2] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, July 1997. 3

[3] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. 3

[4] Olmo Ceriotti and Federico Gerardi. Language modeling for epigraphs: a bert model for edr's latin epigraphs text completion. In *IEEE Cyber Humanities*, 2025. To appear. 1, 2, 3, 4

[5] EDH - Epigraphic Database Heidelberg. EDH Official Website. https://edh.ub.uni-heidelberg.de/. Accessed: 2025-05-16. 2

[6] EDR - Epigraphic Database Roma. EDR Official Website. http://www.edr-edr.it/default/index.php. Accessed: 2025-05-16. 2

[7] Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. The Classical Language Toolkit: An NLP framework for pre-modern languages. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online, August 2021. Association for Computational Linguistics. 2

[8] Alessandro Locaputo, Beatrice Portelli, Elisabetta Colombi, Giuseppe Serra, et al. Filling the lacunae in ancient latin inscriptions. In *Proceedings of the 19th Italian Research Conference on Digital Libraries (IRCDL 2023)*, pages 68–76, 2023. 2

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 2

[10] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020. 2, 3

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. 1, 3

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 2