Usage of Language Model for the Filling of Lacunae in Ancient Latin Inscriptions: A Case Study

Andrea Brunello[2], Emanuela Colombi[2], Alessandro Locaputo[1,*,†], Stefano Magnani[2], Nicola Saccomanno[1] and Giuseppe Serra[1]

[1]Department of Mathematics, Computer Science, and Physics, University of Udine, Italy
[2]Department of Humanities and Cultural Heritage, University of Udine, Italy

Abstract

This paper investigates the efficacy of LatinBERT in the task of infilling ancient Latin inscriptions. We contrast the baseline LatinBERT model with a version fine-tuned specifically for this task. A comprehensive experimental design evaluates the influence of various lacunae features, such as their length and relative position within the text, on the infilling process. In contrast to the results presented in LatinBERT's original publication, our findings indicate suboptimal performance. Interestingly, a parallel study of Greek inscriptions using models like PYTHIA and Ithaca demonstrated vastly superior performance in similar tasks. This disparity underscores the need for the development of more proficient models tailored for Latin inscriptions. Moreover, our study emphasizes the importance of robust and systematic evaluation methodologies to accurately assess model performance.

Keywords

Epigraphy, Lacunae, Latin, Deep Learning

1. Introduction

The examination and restoration of ancient inscriptions are key subjects of study in epigraphy, which is the field dedicated to analyzing and interpreting such inscriptions. Often, the condition in which these ancient inscriptions are preserved makes their legibility a challenging task, since portions of the text can, over time, become lost forever, rendering the inscription partially or totally unreadable. Hence, to be able to extract information from them, inscriptions have to undergo a restoration process that requires the involvement of an expert epigraphist well-versed in the language in which the inscriptions are written; finding such expertise can be hard, especially for less-spoken languages. Moreover, the reconstruction is a time-consuming process because epigraphists need to draw parallels

2nd Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2023, https://ai4ch.di.unito.it/), co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023). 6-9 November 2023, Rome, Italy

*Corresponding author.
†First author.
andrea.brunello@uniud.it (A. Brunello); emanuela.colombi@uniud.it (E. Colombi); locaputo.alessandro@spes.uniud.it (A. Locaputo); stefano.magnani@uniud.it (S. Magnani); nicola.saccomanno@uniud.it (N. Saccomanno); giuseppe.serra@uniud.it (G. Serra)
0000-0003-2063-218X (A. Brunello); 0000-0002-0384-6664 (E. Colombi); 0000-0003-1962-115X (A. Locaputo); 0000-0001-7869-344X (S. Magnani); 0000-0001-5916-3195 (N. Saccomanno); 0000-0002-4269-4501 (G. Serra)

with similar inscriptions relying on their knowledge and experience to fill the lacunae in the text.

Over the years, many prominent collections of ancient inscriptions, such as the *Corpus Inscriptionum Latinarum*, the *Corpus Inscriptionum Graecarum*, and *L'Année épigraphique* have been digitized and collected in digital corpora, among others. Notable examples of such digital corpora are the EAGLE project (Europeana network of Ancient Greek and Latin Epigraphy)[1], an online corpus that gathers inscriptions from various European epigraphic databases, and the Cuneiform Digital Library Initiative[2] for the preservation of text and images of cuneiform inscriptions.

This recent increase in the availability of such digitized corpora has enabled the application of machine learning methods to the field of epigraphy. For instance, PYTHIA [1] and Ithaca [2] are two neural networks designed for filling lacunae in ancient Greek inscriptions, to expedite the restoration process by assisting epigraphists .

Models such as PYTHIA and Ithaca emphasize how these types of tools serve as useful companions to epigraphists, by demonstrating the ability of the models to improve humans' capabilities.

In light of the success with Greek inscriptions, in this work we focus on Latin ones. Specifically, we study the capacity of LatinBERT, a BERT-based [3] model trained on Latin, to autonomously restore lacunae in ancient Latin texts. Through a thorough experimental design, based on a public dataset of ancient Latin inscriptions, we evaluate how LatinBERT's performance is impacted by the inherent characteristics of the lacunae. As we will see, our observed results are markedly inferior to those presented in the original LatinBERT article, highlighting two fundamental criticalities: the need for a higher-performing model that can be used to infill Latin inscriptions; and, the necessity of devising robust and systematic evaluation workflows for the latter task.

2. Related Work

The first specialized neural network designed to aid epigraphists in restoring ancient Greek inscriptions is PYTHIA [1], which utilizes a bi-directional LSTM to produce 20 hypotheses for filling the specified lacuna. The same bi-directional LSTM architecture was later employed in the restoration of Akkadian inscriptions [4]. Both of the aforementioned models require the epigraphist to specify not only the location of the gaps to be filled but also their dimensions in characters. To overcome this limitation, the Blank Language Model (BLM)[5], a Transformer-based model [6] capable of filling gaps with an arbitrary number of characters, was introduced. When evaluated on the same dataset used for assessing PYTHIA, BLM demonstrated similar accuracy.

Instead of just focusing on the infilling task, Ithaca [2] addressed the problem together with two other fundamental tasks in the epigraphist workflow: the temporal and spatial attribution of ancient Greek inscriptions. Ithaca's architecture is inspired by BigBird [7] (i.e., another

[1] https://www.eagle-network.eu/
[2] https://cdli.mpiwg-berlin.mpg.de/

Transformer-based language model), with its output passed on to three different Multi-Layer Perceptrons, one for each epigraphical task. The model's Top-1 accuracy (62%) in filling the lacunae surpassed PYTHIA's Top-1 accuracy (32%). Moreover, the authors showed that the best performance could be achieved when employing Ithaca to assist trained epigraphists, improving their accuracy from 25% to 72%.

When it comes to Latin, the only model whose performance has been assessed for the problem of infilling is LatinBERT [8], a BERT-based [3] masked language model pre-trained on an extensive corpus of 642.7 million Latin words, encompassing texts from the Classical age to contemporary documents originating from the Latin Wikipedia. In the paper in which LatinBERT was introduced, one of the case studies considered was the filling of literature documents extracted from the Latin Library[3]. Unlike the other previously mentioned models, the performances of LatinBERT were not assessed by artificially creating the gaps and comparing them with the model's predictions but rather by comparing the concordance of the model's predictions to the emendation made by an epigraphist, in which it scored a Top-1 accuracy of 33.1%.

In this regard, our experimental workflow is radically different. First, we show how LatinBERT performance, when tested on the same setting as the other previously described approaches, becomes unsatisfactory for the filling of ancient inscriptions. Then, we fine-tune a LatinBERT model by focusing precisely on ancient Latin inscriptions infilling, and finally, we study its performance and how the latter changes when dealing with different types of *lacunae*.

3. Materials and Methods

3.1. Dataset

The dataset used for our experiments has been obtained from the Epigraphik-Datenbank Clauss/Slaby (EDCS) [4], the most comprehensive collection of ancient inscriptions from the Roman Empire. It also includes information from 45 external corpora, including the Corpus Inscriptionum Latinarum and the inscriptions that are part of the EAGLE project, for a total of over 537,000 inscriptions.

Most of the inscriptions retrieved from EDCS are marked up according to a custom notation that slightly differs from the standard Leiden Conventions [9], a set of rules and symbols used by corpora's editors to annotate inscriptions [10]. This markup includes the expansion of abbreviated words, restoring erroneously omitted characters, and proposing missing letters. As a result, the inscriptions underwent filtering, which involved discarding the empty inscriptions and the repeated ones, and they were also cleaned of such notation. This resulted in a total of 211,601 cleaned inscriptions. During this process, due to the scarcity of data and the lack of ground truth, we decided to retain all the emendations proposed by the editors, including the integration of some of the lacunae.

[3] http://thelatinlibrary.com/
[4] https://db.edcs.eu/

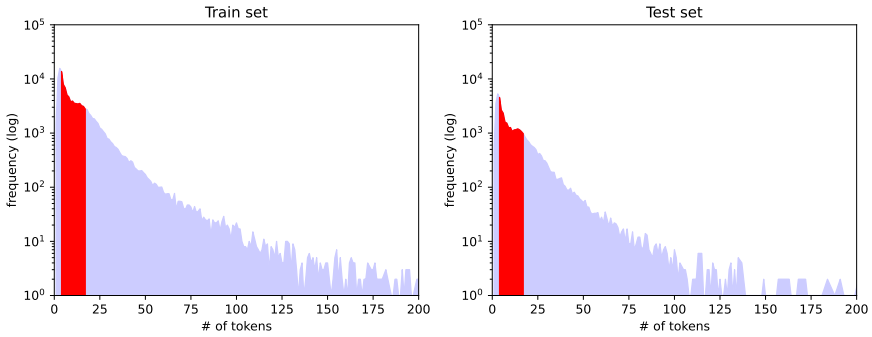


Figure 1: Frequency of token length occurrences in the train and test sets (logarithmic scale). The shaded area in red represents the interquartile range (between the first and third quartiles).

The preprocessed dataset was subsequently divided into three subsets: a training set, a validation set, and a test set, with a split of 60% for the training set and 20% each for the validation and test sets. For the experiments, in the test set, only the inscriptions with a number of tokens that fall between the first and third quartiles are considered (Figure 1), resulting in a total of 22,926 inscriptions. This is done to ensure a balanced and representative sample that avoids extreme outliers.

3.2. Model

The training data served for our realization of LatinBERT-epi, a specialized version of LatinBERT fine-tuned specifically for the infilling of lacunae in ancient Latin inscriptions. The model has undergone fine-tuning for 15 epochs, determined by its performance on the validation set, with an early stopping patience of 5 epochs and a learning rate of 1e-5. This fine-tuning was conducted on a single NVIDIA RTX A5000 GPU with 24GB of VRAM. Meanwhile, the test set mentioned above, limited to inscriptions within the first and third quartiles, was again used for establishing its final performance.

It is important to note that while the results of our experiments are significant, they may not be directly comparable to those of Ithaca due to differences in fundamental units of operation: LatinBERT operates on tokens corresponding to sub-words, whereas Ithaca predicts at the character level. On top of that, Ithaca requires the epigraphist to specify both the exact number and position of the missing characters and utilizes this information to generate predictions of the requested size exclusively. In contrast, LatinBERT only necessitates knowledge of the positions, without regard for the length of the lacunae. This can be a disadvantage as its predictions may not always match the size of the gap.

4. Experiments

In the following experiments, to better reflect real-world scenarios, we refrain from masking entire words and instead focus on sub-word tokens. This decision is based on the understanding

Hic requiesc<i=E>t in pace
Paulus v(i)r l(audabilis) ser<v=B>us d(e)i
mil<es=IX>
de num(ero) Zal(iorum) qui <v=B>ixit annis
plus minus XL depositus est
in pace su<b=D> die tertiu(m) Kal(endas)
Februar(ias)
per ind(ictionem) XI

Hic requiescit in pace
Paulus vir laudabilis servus dei miles
de numero Zaliorum qui vixit annis
plus minus XL depositus est
in pace sub die tertium Kalendas Februarias
per indictionem XI

hic requiesc it in pace
paulus vir laudabili s servus dei miles
de numero za liorum qui vixit annis
plus minus xl depositu s est
in pace sub die tertium kalendas februari as
per indict ionem xi

1

hic requiesc it in pace
paulus vir laudabili s servus dei miles
de numero za liorum qui vixit annis
plus minus xl depositu s est
in pace sub die tertium kalendas februari as
per indict ionem xi

3

hic requiesc it in pace
paulus vir laudabili s servus dei miles
de numero za liorum qui vixit annis
plus minus xl depositu s est
in pace sub die tertium kalendas februari as
per indict ionem xi

2

hic requiesc it in pace
paulus vir laudabili s servus dei miles
de numero za liorum qui vixit annis
plus minus xl depositu s est
in pace sub die tertium kalendas februari as
per indict ionem xi

4

Figure 2: The text of each inscription is cleared of any markup notation (in red) and then tokenized. The resulting texts are then used to evaluate the models in four experiments: 1) Randomly masking 15% of the tokens; 2) masking consecutive spans of 10%, 20%, and 30% starting from the beginning, middle, and end; 3) randomly masking a single token at a time based on its length, starting from 1 up to 9; 4) masking a single token at a time, one for each Part-of-Speech role.

that erosion typically affects portions of inscriptions rather than removing entire words. In Experiment 4.1, we assess the performance of the fine-tuned model compared to the base model by applying the same Masked Language Model objective used during pre-training. In Experiment 4.2, we analyze how model accuracy varies based on the location of the lacuna within the inscription. In Experiment 4.3, the models are evaluated based on the number of characters that make up the lacuna. Finally, in Experiment 4.4, we study the performance of the models by masking tokens according to their POS (part-of-speech) tags in the sentence.

Model	Top-1	Top-10	Top-50
LatinBERT-base	0.0242 ± 0.0008	0.0628 ± 0.0012	0.1189 ± 0.0016
LatinBERT-epi	**0.0402** ± 0.0007	**0.0832** ± 0.0003	**0.1547** ± 0.0010

Table 1
Average accuracy of two models with 15% of tokens masked, measured at 1, 10, and 50.

4.1. First experiment: Mask 15% of the tokens

To compare the performance of the LatinBERT base model (from now on referred to as LatinBERT-base) and the one fine-tuned on the inscriptions, we evaluated both by applying the same MLM (Masked Language Model) objective used in the pre-training phase. Specifically, we masked 15% of the tokens in each inscription and measured the accuracy at 1, 10, and 50. As shown in Table 1, the performances of LatinBERT-base on inscriptions are far from those presented in the original paper, where the reported Top-1 accuracy was equal to 33.1%. This may be due also to the fact that LatinBERT-base is predominantly pre-trained on literary documents. Although these documents are written in a language similar to that of the inscriptions, they exhibit a different syntactic structure, which is less strict than the one found in inscriptions. This is the reason why we also considered LatinBERT-epi which nevertheless, although outperforming the base model, still exhibits lower accuracy than expected.

4.2. Second experiment: Lacunae occurring in different positions

Lacunae can occur in any part of the text and can spread for any given length. Considering this, the two models are here evaluated in three different scenarios: when the gap occurs at the beginning, in the middle, or at the end of the text. For each scenario, we masked consecutive spans of tokens equal to 10%, 20%, and 30% of the total number of tokens of each inscription.

When examining the results in Table 2, it is important to consider that the set of inscriptions used to evaluate the models contains short inscriptions (Figure 1). To ensure that at least one token per text is masked, the number of tokens to be masked has been calculated as the ceiling of the specified percentage. Thus, in many cases, masking 10% of the tokens results in the same number of tokens being masked as when masking 20% of them.

It should also be noted that inscriptions are often highly formulaic. For instance, in funerary inscriptions, it is very common to begin with '*Dis Manibus*'[5]. Since this kind of inscription is prevalent in the dataset and typically quite short, it helps explain why the performance of the fine-tuned model is best when masking only a few tokens at the beginning. Meanwhile, the lowest overall performance reported for LatinBERT-epi is when the lacuna lies in the middle part. This can be attributed to the fact that, unlike the beginning, the middle part contains higher variability even when in formulaic inscriptions, for instance in funerary ones it is where the name of the deceased is mentioned, which is very challenging for the model to predict.

It is surprising to see that both Top-10 and Top-50 accuracy for LatinBERT-epi are higher when the lacunae occur at the end of the text than when they occur in the middle. In the former

[5]It translates to: '*to the spirits of the dead*'

%	Top-1	Top-10	Top-50	Position
10	0.0102	0.0354	0.0646	B
	0.0471	**0.0896**	**0.1414**	
20	0.0106	0.0595	0.1025	
	0.0267	**0.0956**	**0.1772**	
30	0.0205	0.0918	0.1444	
	0.0308	**0.1200**	**0.2119**	
10	0.0223	0.0627	0.1192	M
	0.0372	**0.0860**	**0.1564**	
20	0.0186	0.0629	0.1218	
	0.0326	**0.0840**	**0.1550**	
30	0.0163	0.0677	0.1326	
	0.0297	**0.0856**	**0.1611**	
10	0.0183	0.0669	0.1379	E
	0.0366	**0.0926**	**0.1872**	
20	0.0128	0.0575	0.1229	
	0.0269	**0.0853**	**0.1775**	
30	0.0114	0.0639	0.1328	
	0.0232	**0.0868**	**0.1780**	

Table 2
Accuracy at 1, 10, and 50 of the two models (grey = LatinBERT-base, white = LatinBERT-epi) as the size and position (B = beginning, M = middle, E = end) of the masked token varies.

case, the context is limited to just one side, while in the latter case, the model can take advantage of context on both the left and right.

4.3. Third experiment: Mask tokens of different length

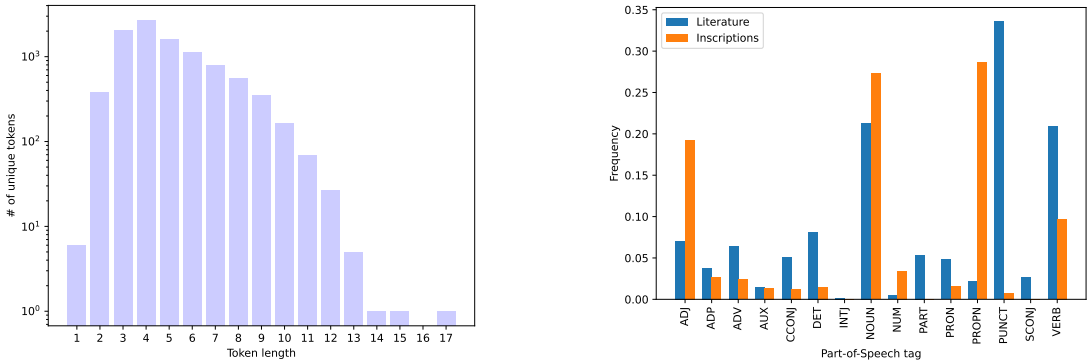
Intuitively, a model should find it easier to fill single small gaps rather than long ones. To evaluate this aspect, we mask tokens based on their length, ranging from single-character tokens to tokens with a length of 9 characters (Table 3). For each token length, the metrics are computed using a subset of the testing set, consisting of inscriptions that contain at least one token of that length.

This experiment highlights the difficulty of the model to correctly predict tokens of length equal to 5. This can be put into context by looking at the number of unique tokens per token length (Figure 3a). It can be noticed that tokens with a length equal to 5 are among those that present a very high variability in the dataset, and thus are the hardest to predict.

4.4. Fourth experiment: Mask according to the PoS tag

For this experiment, the accuracy is measured according to the Part-of-Speech role of the masked token. Thus, to distinguish between the different roles of each token, it is necessary to train an additional model for this sole purpose.

The Part-of-Speech (PoS) tagging of the test set was performed using a specialized version



(a) (b)

Figure 3: (a) Number of unique tokens for each token length.; (b) Frequencies of Part-of-Speech Tags in the Inscriptions Test Set and in the Literary Dataset used for fine-tuning the PoS tagger

Size	# of inscriptions	Top-1	Top-10	Top-50
1	9916	0.0908 ± 0.0020	0.1709 ± 0.0022	0.2300 ± 0.0020
		0.1303 ± 0.0017	**0.7166** ± 0.0020	**0.7983** ± 0.0030
2	14210	0.0297 ± 0.0007	0.1027 ± 0.0012	0.1986 ± 0.0031
		0.1178 ± 0.0009	**0.2219** ± 0.0020	**0.5560** ± 0.0013
3	16607	0.0140 ± 0.0004	0.0412 ± 0.0012	0.0883 ± 0.0014
		0.0251 ± 0.0005	**0.0611** ± 0.0013	**0.1299** ± 0.0011
4	18245	0.0148 ± 0.0014	0.0274 ± 0.0013	0.0471 ± 0.0007
		0.0377 ± 0.0001	**0.0508** ± 0.0001	**0.0738** ± 0.0011
5	14080	0.0063 ± 0.0006	0.0152 ± 0.0002	0.0278 ± 0.0006
		0.0213 ± 0.0007	**0.0327** ± 0.0008	**0.0491** ± 0.0013
6	14581	0.0088 ± 0.0008	0.0205 ± 0.0005	0.0336 ± 0.0003
		0.0335 ± 0.0006	**0.0499** ± 0.0011	**0.0715** ± 0.0013
7	10990	0.0061 ± 0.0008	0.0147 ± 0.0012	0.0254 ± 0.0014
		0.0288 ± 0.0003	**0.0465** ± 0.0008	**0.0632** ± 0.0008
8	9866	0.0071 ± 0.0010	0.0157 ± 0.0011	0.0393 ± 0.0010
		0.0720 ± 0.0007	**0.0830** ± 0.0005	**0.0924** ± 0.0006
9	3760	0.0046 ± 0.0002	0.0087 ± 0.0004	0.0145 ± 0.0002
		0.0163 ± 0.0004	**0.0266** ± 0.0003	**0.0327** ± 0.0005

Table 3
Accuracy at 1, 10, and 50 of LatinBERT-base (light grey) and LatinBERT-epi (white) as the size of the masked token varies.

of LatinBERT, fine-tuned specifically for the PoS tagging task[6]. This model was trained on 18,184 tokens[7] from the Perseus Latin Treebank [11], a corpus comprising Classical Latin texts sourced from the Perseus Digital Library [12]. It is worth noting that while the Perseus Digital

[6]Differently from the pretraining phase of LatinBERT that is done in an unsupervised manner, the fine-tuned model produced is a classifier trained in a supervised manner.
[7]Each of these tokens has been manually tagged with the corresponding PoS tag, which serves as the ground truth for the classifier.

PoS	Top-1	Top-10	Top-50
PRON	0.0000 ± 0.0000	0.0081 ± 0.0035	0.0122 ± 0.0000
	0.0203 ± 0.0035	**0.0447** ± 0.0035	**0.1443** ± 0.0035
ADV	0.0071 ± 0.0031	0.0089 ± 0.0031	0.0089 ± 0.0031
	0.0071 ± 0.0031	**0.0160** ± 0.0053	**0.1543** ± 0.0232
ADJ	0.0109 ± 0.0014	0.0244 ± 0.0003	0.0401 ± 0.0007
	0.0437 ± 0.0012	**0.1142** ± 0.0017	**0.1997** ± 0.0015
VERB	0.0085 ± 0.0000	0.0173 ± 0.0008	0.0285 ± 0.0013
	0.0823 ± 0.0018	**0.1301** ± 0.0012	**0.1608** ± 0.0031
NOUN	0.0090 ± 0.0000	0.0180 ± 0.0013	0.0283 ± 0.0003
	0.0300 ± 0.0010	**0.0979** ± 0.0020	**0.1451** ± 0.0018
DET	0.0039 ± 0.0000	0.0092 ± 0.0023	0.0157 ± 0.0000
	0.0223 ± 0.0023	**0.0289** ± 0.0023	**0.0407** ± 0.0023
PUNCT	0.0000 ± 0.0000	0.0024 ± 0.0000	0.0201 ± 0.0007
	0.0614 ± 0.0012	**0.1821** ± 0.0056	**0.6963** ± 0.0018
SCONJ	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.0000 ± 0.0000	0.0000 ± 0.0000	0.1515 ± 0.0525
CCONJ	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.0000 ± 0.0000	**0.0571** ± 0.0000	**0.3143** ± 0.0000
PROPN	0.0124 ± 0.0014	0.0277 ± 0.0005	0.0423 ± 0.0025
	0.0562 ± 0.0003	**0.1696** ± 0.0042	**0.2257** ± 0.0042
ADP	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.0074 ± 0.0128	**0.0815** ± 0.0128	**0.4000** ± 0.0222
PART	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.0000 ± 0.0000	**0.1667** ± 0.1443	**0.4167** ± 0.1443
NUM	0.0040 ± 0.0006	0.0162 ± 0.0013	0.0361 ± 0.0006
	0.0085 ± 0.0006	**0.0408** ± 0.0022	**0.1089** ± 0.0061
AUX	0.0058 ± 0.0000	**0.0633** ± 0.0000	**0.0801** ± 0.0000
	0.0437 ± 0.0000	0.0604 ± 0.0000	0.0750 ± 0.0000
INTJ	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.0000 ± 0.0000	**0.0175** ± 0.0304	**0.1053** ± 0.0526

Table 4
Accuracy at 1, 10, and 50 for LatinBERT-base (light grey) and LatinBERT-epi (white) as the part of speech role of the masked token varies.

Library contains contemporaneous documents related to the considered inscriptions, it primarily consists of literary sources with distinct syntactic structures, which becomes evident when comparing the frequency of occurrence of each PoS tag of the two (Figure 3b). Consequently, the PoS tagging accuracy of the model for inscriptions may be lower than the 94.3% reported in the LatinBERT paper for Classical documents.

In Table 4, the lowest performance is observed when masking coordinating conjunctions (CCONJ), subordinating conjunctions (SCONJ), particles (PART), and interjections (INTJ). One possible explanation for this is their infrequent use in inscriptions, which prevents the fine-tuned model from learning how to correctly fill them.

The only PoS tag for which the base model outperforms the fine-tuned one is the prediction of auxiliary verbs (AUX). This can be attributed to the fact that inscriptions typically prioritize

brevity and conciseness, resulting in limited usage of auxiliary verbs. Another challenging task for both models is predicting numerals (NUM) because, similar to proper nouns, they can be difficult to infer from the context, as often there are multiple solutions that, while not correct, still make perfect sense.

Overall, LatinBERT-epi's accuracy is higher than the base model for each PoS tag, highlighting and confirming the different syntactical structures between inscriptions and literary documents.

5. Discussion

As hypothesized, Latin used in literature documents, which has been relied upon for the pre-training of LatinBERT greatly differs from the Latin that appears in ancient inscriptions, both due to a different syntactic structure and the evolution that the language has witnessed over the centuries. Thus, it should not entirely come as a surprise that the performances of the base model are lower than the ones reported in the original LatinBERT's paper, nevertheless, the fine-tuned model, while improving the accuracy, still reported underwhelming performances, especially when compared to PYTHIA and Ithaca results. In light of this, it is important to point out the way in which LatinBERT evaluation was conducted: the authors did not randomly mask parts of the text, but they rather measured the concordance of the model's predictions with epigraphists emendations; for doing so they restricted to those inscriptions where there is a single emendation made of a single word of at least two characters, thus Experiment 4.3 is the one closer to their experimental setting. However, it is important to notice that, when using PYTHIA and Ithaca, the epigraphist has to specify which characters (their number and position) the model has to predict, thus providing the model additional information regarding the characteristics of the lacunae. This is not the case with LatinBERT, where the only information provided by the epigraphist is about the location of the lacunae.

The experiments did not uncover a specific aspect in which LatinBERT is lacking but rather showed consistent difficulties in correctly filling the gaps. Given the lower-than-expected performance of our model and the fact that many papers in this field often emphasize collaboration with epigraphists rather than in-depth analysis of model performance, we recognize the importance of establishing a well-defined pipeline of experiments to assess language models' accuracy in filling lacunae and to develop a model based on Ithaca's architecture also for Latin.

We believe that the pipeline should possess at least the following requirements:

- It must consider the various positions where inscriptions can occur, given that inscriptions are often highly formulaic. Consequently, certain parts may be easier to predict than others, as emerged in Experiment 4.4.
- It must favor models with a higher Top-10 accuracy over those with a higher Top-50 accuracy since these tools are expected to be used in conjunction with an epigraphist which has to evaluate every prediction of the model.

6. Conclusions

In this work, we presented a fine-tuned version of LatinBERT for filling lacunae in Ancient Latin inscriptions and then evaluated it by comparing its performance to the baseline LatinBERT model in the task of filling the lacunae without human intervention in different scenarios, analyzing how the features of the inscriptions affect the model's predictions. The experiments highlighted the suboptimal performances of LatinBERT in this task, which, when compared to the results showed by PYTHIA and Ithaca with ancient Greek inscriptions, underscores the necessity of establishing a comprehensive and standardized set of experiments to more accurately assess the performance of these models and the need of a more proficient Latin-specific model.

The remark made by PYTHIA and Ithaca about involving domain experts to better evaluate these models still remains valid, although it should be considered that this is something that is not always feasible, especially for the less-spoken languages.

Acknowledgments

This work was supported by the Department Strategic Plan (DSP) of the University of Udine—Interdepartmental Projects: Artificial Intelligence, Artificial Intelligence for Cultural Heritage (AI4CH); PRIN 2022 - Project code: 2022YTE579.

References

[1] Y. Assael, T. Sommerschield, J. Prag, Restoring ancient text using deep learning: A case study on Greek epigraphy, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6367–6374. doi:10.18653/v1/D19-1668.

[2] Y. Assael, T. Sommerschield, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, J. Prag, N. de Freitas, Restoring and attributing ancient texts using deep neural networks, Nature 603 (2022) 280–283. doi:10.1038/s41586-022-04448-z.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805.

[4] E. Fetaya, Y. Lifshitz, E. Aaron, S. Gordin, Restoration of fragmentary Babylonian texts using recurrent neural networks, Proceedings of the National Academy of Sciences 117 (2020) 22743–22751. doi:10.1073/pnas.2003794117.

[5] T. Shen, V. Quach, R. Barzilay, T. Jaakkola, Blank Language Models, 2020. arXiv:2002.03079.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[7] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.
[8] D. Bamman, P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, 2020. arXiv:2009.10053.
[9] C. Bruun, J. Edmondson, The Oxford handbook of Roman epigraphy, Oxford University Press, 2014.
[10] J. Flanders, C. Roueché, Introduction to epidoc guidelines, 2006. URL: https://epidoc.stoa.org/gl/latest/intro-eps.html, accessed on October 17, 2023.
[11] D. Bamman, G. Crane, The Design and Use of a Latin Dependency Treebank (2006).
[12] Perseus Digital Library, Perseus digital library, 2023. URL: http://www.perseus.tufts.edu, accessed on September 19, 2023.