Filling the Lacunae in ancient Latin inscriptions

Alessandro Locaputo[1], Beatrice Portelli[1], Emanuela Colombi[2] 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

Inscriptions are a testimony to the past but their poor condition, caused by the deterioration of the material on which they are engraved upon, often makes them partially or completely illegible. The process of restoring these inscriptions is time-consuming and requires the involvement of an expert epigraphist. It is possible to speed-up this process by adopting a semi-automatic assisting tool based on deep neural networks. This work describes a methodology, from the acquisition of the inscriptions to the description of four possible approaches, to predict the missing text in a Latin inscription, that our research team plans to implement in the near future as part of an interdisciplinary research project.

Keywords

Epigraphy, Lacunae, Digital Humanities, Deep Learning, Latin

1. Introduction

Epigraphy is the study of inscriptions, which can be described as text engraved on any durable material, such as stone and metals, but also painted text on almost all kind of surfaces [1].

In the ancient Roman society, inscriptions were used for numerous different purposes, such as military records, juridical records and public notices. Their wide use makes inscriptions an invaluable evidence of the past [2]. For instance, honorary inscriptions give us information about the *cursus honorum*, the sequence of public offices held by aspiring politicians in the Roman Republic [1].

The material on which the inscriptions were carved upon is subject to deterioration over time. It is estimated that only between 2% and 3% of all Latin inscriptions have survived to this day [1]. For instance, a slab of stone could be missing some parts due to crumbling and thus creating a gap in the text. This newly created gap is commonly referred to as *lacuna*. The presence of *lacunae* in an inscription makes it partially or completely illegible (Figure 1). The process of filling these *lacunae* is time-consuming and requires the involvement of an expert epigraphist, who has to conjecture the size of the gap as well as the content of the missing text. The standard epigraphic convention for reporting these conjectures is to put the presumed text within square brackets [3].

19th IRCDL (The Conference on Information and Research science Connecting to Digital and Library science), February 23—24, 2023, Bari, Italy

locaputo.alessandro@spes.uniud.it (A. Locaputo); portelli.beatrice@spes.uniud.it (B. Portelli); emanuela.colombi@uniud.it (E. Colombi); giuseppe.serra@uniud.it (G. Serra)
0000-0003-1962-115X (A. Locaputo); 0000-0001-8887-616X (B. Portelli); 0000-0002-0384-6664 (E. Colombi); 0000-0002-4269-4501 (G. Serra)

During the years, there have been several attempts to create corpora that collect all available inscriptions for specific languages. Two notable examples for Latin inscriptions are the *Corpus Inscriptionum Latinarum* (CIL) [4] and *L'Année épigraphique* (AE) [5]. These corpora also contain, for each inscription, the interpretations made by the epigraphists regarding the resolution of abbreviated texts, and the addition of missing text.

The objective of this paper is to present our research project for ancient Latin inscriptions restoration. We will describe our action plan to develop a neural network aided companion tool for scholars to speed-up the restoration process, by presenting our proposals of four different deep learning architectures based on previous literature and innovative ideas. Our work focuses on Latin inscriptions, as this ancient language still lacks automatic tools for filling *lacunae* and it has been investigated less than other ancient languages (e.g., Ancient Greek). This specific type of inscription is written for the most part in the Latin language but may also contain Greek words or transliteration of them, as well as lexical borrowings and hybridisation with foreign languages [1].

The joint research group will be composed of epigraphy experts from the Department of Humanities and Cultural Heritage, as well as deep learning experts form the Department of Mathematics, Computer Science, and Physics of the University of Udine. Thanks to this interdisciplinary expertise we will be able to solve a problem that will benefit both research fields.

In fact, aside from the humanistic point of view, the problem of filling *lacunae* in Latin is nevertheless a challenging problem from the deep learning perspective, since these kinds of methods require to be trained on a large amount of data, which is not available when working with an ancient language such as Latin.

Figure 1: Damaged inscription from the Fortress of Deva, EDCS-09400486. (CC BY 3.0, Wikimedia)

2. Related work

In the recent years, there has been a growing focus on the adoption of Artificial Intelligence (AI) for the analysis of historical documents [6]. For example, AI has been applied to the study of Ancient Christian inscriptions in order to automatically extract new information by analysing their features (e.g., language used, writing style, and material) [7], or to develop tools such as HisDoc [8] to analyse medieval Latin manuscripts from the 9th century.

PYTHIA [9] was the first deep learning model capable to perform fully-automated ancient text restoration. It is able to fill the *lacunae* in damaged ancient Greek inscriptions using a sequence-to-sequence architecture with a bidirectional-LSTM encoder. In order to restore incomplete and missing words, it works at both word and character level, which also allows it to build a better internal word representation [10].

When it comes to restoring Greek inscriptions, Ithaca [11] is the state of the art. In order to enable large-scale processing, it uses a Transformers-based architecture. Additionally to filling *lacunae*, Ithaca can also determine the original location of an inscription and place it in time.

PYTHIA and Ithaca have demonstrated that the adoption of this kind of assistive tools can improve both the accuracy and the speed of the epigraphist's restoration activity.

The problem of filling *lacunae* in ancient texts can be seen as a specific case of what in NLP is generally referred to as *text infilling* [12], which is the task of filling the gaps in a text. This, in turn, is a generalisation of the cloze task. Recently, there have been some advancements in this field thanks to researchers adapting Language Models to perform text infilling [13]. As an example, the Blank Language Model (BlankLM) [14] generates sequences of text by dynamically creating and filling gaps, and it has also been used to perform ancient text restoration, showing an accuracy similar to PYTHIA's.

The use of a Masked Language Modeling (MLM) approach has been proven to be effective not only for the restoration of ancient documents, but also for their translation [15]. A recent work [16] proposed a different approach for restoring texts written in the Akkadian language,[1] showing that it is possible to use a pretrained multi-language Language Model such as Multilingual-BERT [17] and fine-tune it on a small dataset, such as the Akkadian language one, to achieve state-of-the-art performances.

When it comes to the Latin language, Latin BERT [18] has been proposed, a contextual language model, trained on a corpus of documents spanning from the Classical era to contemporary sources. Due to the expensiveness of training a BERT [17] model, both in terms of computational requirements and data availability, researchers have also proposed an ELECTRA-based model [19] trained on the Latin language.

Interestingly, the problem of filling *lacunae* can also be framed in the context of image processing, as it is akin to the task of completing patterns on ancient pottery [20]. This is a particular case of the most general inpainting task in computer vision for which, recently, methods based on diffusion models such as RePaint [21] have been proposed. Additionally, although diffusion models are mainly used for computer vision tasks, recently they have been

[1] Between the Late Bronze and Early Iron Ages, the Akkadian language was the *lingua franca* used in the Middle East

also applied to the field of NLP, making them an interesting method to bridge these two fields. For instance, DiffusER [22] and Diffusion-LM [23] are two generative text models based on denoising diffusion models. The first one is a discrete diffusion model that corrupts the text by applying the four Levenshtein edit operations,[2] while the second one is a continuous diffusion model where the diffusion process is continuously applied to word embeddings.

3. Methodology

Many of the physical epigraphic corpora have been digitised and are available as part of various online corpora. These corpora contain not only the transcription of the text of an inscription, but also the annotations made by expert epigraphists when performing restoration. For example, they correct obvious misspelled words, they estimate the number of missing characters and words, as well as make conjectures on how to fill the *lacunae*. All this additional information is valuable for the restoration task, but given the large number of different sources, the potential presence of noise, and the heterogeneity of the annotation styles, it becomes necessary to perform some data cleaning and normalisation of the input text to make it machine-readable. In order to do so, a pipeline specific for Latin inscriptions will be created, analogously to what has already been done for Ancient Greek [9] [11].

The newly acquired data will be used to train a model on Ithaca's architecture, which currently represents the state-of-the-art for the restoration of Greek inscriptions, thus obtaining a baseline for the Latin language. Then, we will develop and analyse three new different approaches to improve on the baseline performance.

3.1. Dataset

One of the main issues when working with ancient languages such as Latin is the scarcity of available data. For historical reasons, the most important corpora of Latin inscription were available only on physical media, namely books. In the recent years, there have been efforts to digitise them all.

This research will make use of the *Epigraphik-Datenbank Clauss/Slaby* (EDCS)[3], an online database comprehensive of 45 different corpora, including the *Corpus Inscriptionum Latinarum*, which contains inscriptions until the Fall of the Western Roman Empire, and *L'Année épigraphique*, a collection of inscriptions, mainly in Latin or Ancient Greek, concerning Ancient Rome. The database is also updated with new findings not available in any printed work and inscriptions originating from numerous online corpora such those part of the EAGLE (Europeana Network for Greek and Latin Epigraphy)[4] project, which gathers the information collected by EDB[5], EDR[6], EDH[7] and other European epigraphic database. The database contains

[2]The Levenshtein edit operations are: Insert, Delete, Keep and Replace
[3]https://db.edcs.eu/
[4]https://www.eagle-network.eu/
[5]https://www.edb.uniba.it/
[6]http://www.edr-edr.it/
[7]https://edh.ub.uni-heidelberg.de/

the transcription of approximately 532 thousands Greek-Latin inscriptions, and it is the most extensive digital resource of Latin inscriptions [24].

Each inscription is annotated, when available, with the hypothesis made by expert epigraphist about the number of missing characters that form the *lacunae* and the eventual conjecture of the missing words, including the reconstruction of abbreviations, and the correction of obvious misspellings by inserting or erasing characters.

In addition to the full transcript of the inscription, EDCS makes use of some special characters to report the conjectures made by the epigraphists. For example, Figure 2 reports the transcript of the damaged inscription in Figure 1, taken from EDCS, where the symbol `[3]` is used to represent a *lacuna* within the line.

```
[3] missa div[3]
[3]A castris qua[3]
[3]corum clause[3]
[3] contra regim[3]
[3]orum fem[3]
[3]S per M[3]
```

Figure 2: EDCS entry for the fragment of inscription in Figure 1. (EDCS-09400486)

To acquire the inscriptions we will make use of the Latin Epigraphy Scraper (LatEpig) [25], a tool able to extract information from EDCS in an easy to read format (e.g. *json*).

3.2. Models

This research project will study four different approaches (Figure 3) to solve the problem of filling *lacunae* in Latin inscriptions.

The first approach (Figure 3a) is to adopt the same Transformer architecture used by Ithaca, inspired by the BigBird [26] model, which is currently the state of the art for restoring ancient Greek inscriptions. The output sequence, generated by Ithaca's torso starting from the text of the inscription where the characters to be restored are marked with the "**?**" symbol, is then given as input to a two-layer feedforward network followed by a softmax function which handles the restoration task, returning the predicted characters. Since Ithaca is the state-of-the-art for ancient Greek inscriptions, it could also serve as a baseline for Latin inscriptions. Since Transformer models can be expensive to train, in the eventuality that this represent a problem, the Transformer architecture will be replaced with an LSTM-based sequence to sequence architecture, similar to the one proposed by PYTHIA.

The second approach (Figure 3b) is based on the one proposed by [16] for the Akkadian language, which identified that the problem of restoring text corresponds to the objective of the Masked Language Modeling task in NLP. Thus, it is possible to restore inscriptions using a fine-tuned pretrained Language Model, such as Multilingual BERT. This approach was proven to be effective in context where the amount of training data is scarce, as in the case of ancient languages. As regards the pretrained model, Multilingual BERT was trained on 104

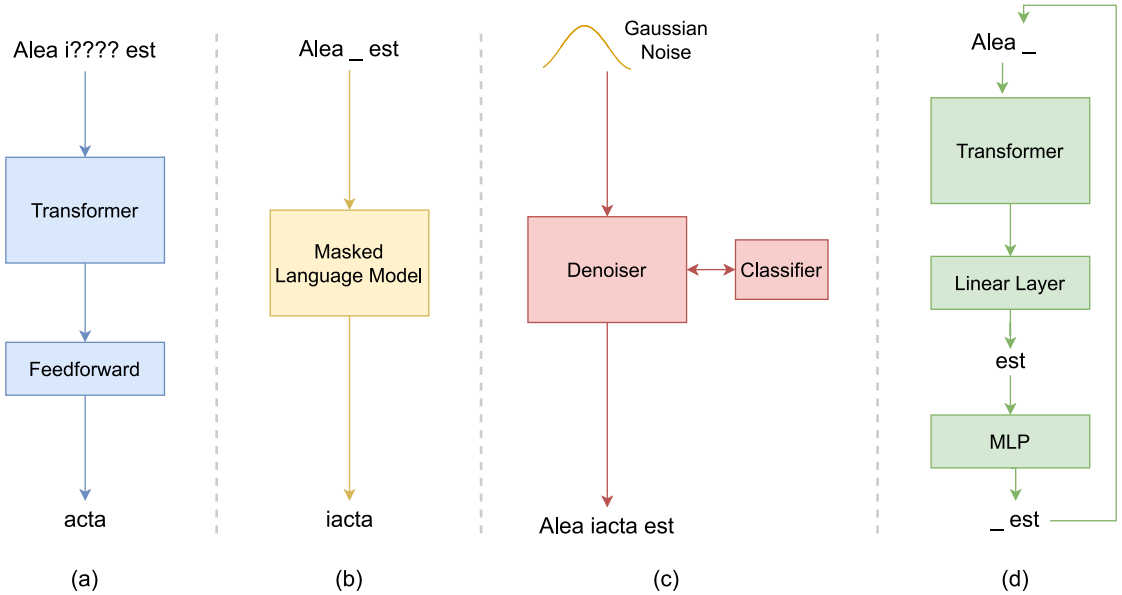

Figure 3: Comparison of the architectures of the four discussed approaches: (a) first approach, based on Ithaca's architecture; (b) second approach, a Masked Language Model; (c) third approach, a di□usion-based model; (d) fourth approach, based on Blank Language Model's architecture.

different languages (including Latin), so it is a suitable candidate. Furthermore, it has also been proposed a BERT-based model pre-trained on a Latin corpus consisting of 642.7 million tokens and containing documents spanning for 22 centuries. This implies that the model has also been trained on documents which are not coeval with the inscriptions contained in EDCS. Despite this, Latin BERT could still be more suitable that Multilingual BERT for fine-tuning thanks to its focus on the Latin language, so it will be taken into considerations in our experiments.

The third approach (Figure 3c) will investigate a possible application of diffusion-based models, given their remarkable results in the inpainting task in computer vision, as well as their recent applications to the NLP field for text generation. This approach will consider the advancements in conditional text generation [23], as it is an essential feature in order to perform text infilling and for being able to control the length of the output, which are both desirable features in our scenario. A Diffusion model works by gradually denoising some random Gaussian Noise using a neural network, in order to synthesize new data. To be able to control this generation process, DiffusionLM proposed to use a classifier, which measures how well the generated text satisfies some constraints.

The main limitation of the first and second approaches is that they rely heavily on the conjectures made by expert epigraphists regarding the size of the *lacunae* and the number of missing characters. These conjectures might be erroneous and lead to low-quality model predictions. The fourth approach (Figure 3d) aims to bypass this limitation by using the same strategy adopted by BlankLM, a model capable of generating sequences of text by fillings the gaps, and possibly introducing new ones (given the predicted word, a MLP determines whether

to introduce a new blank on the left of the word, on the right, on both, or none), and repeating the process until all blanks are filled. Doing so, it is possible to fill a *lacuna* without knowing its exact size. In particular, BlankLM has shown great performance for the restoration of Greek inscriptions, therefore the same approach could also be applied to Latin inscriptions.

4. Conclusions

Despite the presence of encouraging studies that show the effectiveness of automatic methods to fill *lacunae* in ancient texts, this field is still rather unexplored and leaves great opportunities to develop new technologies and create better tools to aid the experts, especially for the Latin language.

This paper describes the foundations and objectives of our research project, which aims to build a deep learning model to restore ancient Latin inscriptions. To this end, first we identified a suitable database which comprises several ancient Greek-Latin inscriptions coming from different renowned corpora. We plan to develop a comprehensive pipeline to pre-process the data, denoise them, normalise them and unify their annotation schema. Finally, we identified four promising deep-learning approaches to fill the *lacunae* in the Latin texts, based on different techniques. The first one is based on the current state-of-the-art model used to restore ancient Greek inscriptions, the second one relies on pretrained language models, the third one leverages the recent advancement in the field of computer vision and diffusion models, and a final one aims to bypass the limitations of the first and the third proposal, that is the need for epigraphists' conjectures about the dimension of the *lacunae*.

The project will be carried out by a multi-disciplinary team, and it aims to create useful and powerful deep-learning tools to assist expert epigraphists in the task of restoring ancient inscriptions.

If successful, the proposed approaches could be easily applied to any other ancient language or, indeed, to any language with limited data availability.

Acknowledgments

This work is partially supported by the Artificial Intelligence project and the interdepartmental DIUM-DMIF project - Department Strategic Plan (DSP) of the University of Udine.

References

[1] A. Buonopane, Manuale di epigrafia latina, Beni culturali, Carocci, Roma, 2009. Tex.lccn: 2009478450.
[2] A. Cooley, The cambridge handbook of latin epigraphy, Cambridge Univ. Press, 2012.
[3] J. Bodel, Epigraphic Evidence, Routledge, 2012.
[4] D. A. der Wissenschaften zu Berlin, B.-B. A. der Wissenschaften, Corpus inscriptionum latinarum, Apud G. Reimerum, 1862. Tex.lccn: 43020276.

[5] J. (Organization), A. d. i. . b.-l. (France), L'Année épigraphique: revue des publications épigraphiques relatives a l'antiquité romaine, Presses Universitaires de France., 1894. Tex.lccn: 2009213588.

[6] J. P. Philips, N. Tabrizi, Historical Document Processing: A Survey of Techniques, Tools, and Trends (2020) 30.

[7] G. Pio, F. Fumarola, A. E. Felle, D. Malerba, M. Ceci, Discovering Novelty Patterns from the Ancient Christian Inscriptions of Rome, Journal on Computing and Cultural Heritage 7 (2015) 1–21. URL: https://dl.acm.org/doi/10.1145/2629513. doi:10.1145/2629513.

[8] A. Fischer, H. Bunke, N. Naji, J. Savoy, M. Baechler, R. Ingold, The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries, in: M. Stolz, Y.-C. Chen (Eds.), Internationalität und Interdisziplinarität der Editionswissenschaft, DE GRUYTER, 2014, pp. 91–106. URL: https://www.degruyter.com/document/doi/10.1515/9783110367317.91/html. doi:10.1515/9783110367317.91.

[9] 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. URL: https://www.aclweb.org/anthology/D19-1668. doi:10.18653/v1/D19-1668.

[10] Z. Zhang, Y. Huang, P. Zhu, H. Zhao, Effective Character-augmented Word Embedding for Machine Reading Comprehension, 2021. URL: http://arxiv.org/abs/1808.02772, arXiv:1808.02772 [cs].

[11] 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. URL: https://www.nature.com/articles/s41586-022-04448-z. doi:10.1038/s41586-022-04448-z.

[12] W. Zhu, Z. Hu, E. Xing, Text Infilling, 2019. URL: http://arxiv.org/abs/1901.00158, arXiv:1901.00158 [cs, stat].

[13] C. Donahue, M. Lee, P. Liang, Enabling Language Models to Fill in the Blanks, 2020. URL: http://arxiv.org/abs/2005.05339, arXiv:2005.05339 [cs].

[14] T. Shen, V. Quach, R. Barzilay, T. Jaakkola, Blank Language Models, 2020. URL: http://arxiv.org/abs/2002.03079, arXiv:2002.03079 [cs].

[15] K. Kang, K. Jin, S. Yang, S. Jang, J. Choo, Y. Kim, Restoring and Mining the Records of the Joseon Dynasty via Neural Language Modeling and Machine Translation, 2021. URL: http://arxiv.org/abs/2104.05964, arXiv:2104.05964 [cs].

[16] K. Lazar, B. Saret, A. Yehudai, W. Horowitz, N. Wasserman, G. Stanovsky, Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach, 2021. URL: http://arxiv.org/abs/2109.04513, arXiv:2109.04513 [cs].

[17] 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.

[18] D. Bamman, P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, 2020. URL: http://arxiv.org/abs/2009.10053, arXiv:2009.10053 [cs].
[19] W. Mercelis, A. Keersmaekers, An ELECTRA Model for Latin Token Tagging Tasks (2022) 4.
[20] S. Lengauer, R. Preiner, I. Sipiran, S. Karl, E. Trinkl, B. Bustos, T. Schreck, Context-based Surface Pattern Completion of Ancient Pottery (2022) 9 pages. URL: https://diglib.eg.org/handle/10.2312/gch20221234. doi:10.2312/GCH.20221234, artwork Size: 9 pages Publisher: The Eurographics Association.
[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, L. Van Gool, RePaint: Inpainting using Denoising Diffusion Probabilistic Models, 2022. URL: http://arxiv.org/abs/2201.09865, arXiv:2201.09865 [cs].
[22] M. Reid, V. J. Hellendoorn, G. Neubig, DiffusER: Discrete Diffusion via Edit-based Reconstruction, 2022. URL: http://arxiv.org/abs/2210.16886, arXiv:2210.16886 [cs].
[23] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, T. B. Hashimoto, Diffusion-LM Improves Controllable Text Generation, 2022. URL: http://arxiv.org/abs/2205.14217, arXiv:2205.14217 [cs].
[24] C. Bruun, J. Edmondson, The Oxford Handbook of Roman Epigraphy, Oxford handbooks, Oxford University Press, 2015.
[25] B. Ballsun-Stanton, P. Heřmánková, R. Laurence, LatEpig (version 2.0). GitHub, 2022. URL: https://github.com/mqAncientHistory/Lat-Epig/. doi:10.5281/zenodo.5211341.
[26] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big Bird: Transformers for Longer Sequences, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 17283–17297. URL: https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.