Named Entity Recognition for a Large Scale Analysis of Individuals in Antiquity

Marijke Beersmans (KULeuven), ORCID: 0009-0002-0826-7319
Evelien de Graaf (KULeuven), ORCID: 0009-0006-8650-1595
Tim Van de Cruys (KULeuven), ORCID: 0000-0002-4650-0444
Alek Keersmaekers (KULeuven), ORCID: 0000-0003-4403-1143
Margherita Fantoli (KULeuven), ORCID: 0000-0003-3191-4860

Short paper abstract

In the past decades, we have seen a rise in large scale digitization efforts for historic and ancient texts, which facilitates the analysis of these texts at a much greater scale. As a manual analysis is often prohibitively laborious and expensive, automatic information extraction techniques, for example Named Entity Recognition (NER), have been increasingly studied in the field of Ancient Language processing (Ehrmann et al., 2020). NIKAW (Networks of Ideas and Knowledge in the Ancient World) is one such large scale interpretation project. Its focus lies on modeling the flow of ideas in Greek and Roman Antiquity by analysing and visualizing the mentions of people using Social Network Analysis (SNA). The first step of creating such a network is detecting the mentions in a large body of Latin and Ancient Greek texts. Therefore, we have been looking into NER as a way of extracting those automatically.

The state-of-the-art approach for NER (as for many other NLP-tasks) consists of finetuned classification architectures based on Large Language Models (LLMs). Recently, there has been a surge in experiments training these LLMs for Ancient and Historic languages, with Latin and Ancient Greek as the most prominent ones. The former is often included in large multilingual language models (Conneau et al., 2020). Additionally, Bamman & Burns (2020), Riemenschneider & Frank (2023), and Mercelis & Keersmaekers (2022) have created a monolingual language model for Latin. For Ancient Greek, researchers have finetuned modern Greek models (Singh et al., 2021; Yamshchikov et al., 2022) or multilingual models (Yousef et al., 2022) on a smaller sample of Ancient Greek texts, or trained a model from scratch (Mercelis, 2021; Riemenschneider & Frank, 2023). As such, there is no shortage of underlying LLMs for training NER systems; however, training these systems requires annotated training data, which are less readily available for the targeted languages.

For Latin, a sample of canonical Latin texts from around the beginning of the era (Caesar's *De Bello Gallico*, Ovid's *Ars Amatoria* and excerpts from Caesar's *Bello Civile*, Pliny the Elder's *Naturalis Historiae* and Pliny the younger's *Epistulae*) was annotated with person (PERS), location (LOC) and group (GRP) entity tags in the context of the Herodotos project. This dataset was used at the time to train a bi-LSTM-CRF NER model (Erdmann et al., 2016). As a first step of adding to this previous research, we finetuned a LatinBERT (Bamman & Burns, 2020) model on this dataset created by the Herodotos project. We annotated additional testing data following their guidelines, and found that on this newly created dataset, the LatinBERT model does outperform the original bi-LSTM-CRF by a small margin, proving the potential for Transformer-based methods for this task on Latin. After a detailed error analysis, we also drew conclusions about the necessity of annotation guidelines, the model's struggle

with the inherent ambiguity of Latin and the difficulty of transfer learning from prose to poetry (Beersmans et al., 2023).

For Ancient Greek, no dedicated NER-dataset was available. However, Ancient Greek texts have been annotated with entity information for other reasons, such as Geographical Information Systems (GIS) or general disambiguation and accessibility for a wider public. We collected four of these datasets. The first was *the Greek New Testament*, manually annotated by bible scholars in the context of the STEPBible project (*STEPBible Data Repository CC BY 4.0*, 2018/2023). The second consisted of the Deipnosophists, an encyclopedic dialogue by Athenaeus of Naucratis containing many detailed references to Ancient Greek authors. This work was semi-automatically annotated. The third is the Odyssey, annotated by Josh Kemp, to make the text more digestible for new classicists (Pelagios, 2021). The final one pertains the *Periegesis Hellados* of Pausanias of Magnesia, an Ancient Greek travelogue often used as an "information repository, particularly for the discovery and interpretation of peoples, sites, and, subsequently, for Hellenic heritage artefacts and monuments" (Foka et al., 2021, p. 57). This text was annotated in the context of the Periegesis project, aiming to visualize the travels in the *Periegesis Hellados* using GIS. After compiling and homogenizing them, we trained four different Ancient Greek transformer models, one modern Greek BERT model finetuned on a smaller subset of Ancient Greek texts (AG-BERT) and three other models trained from scratch on Ancient Greek data of various architectures and sizes; an ELECTRA model (Mercelis, 2021), a DeBERTa model (to appear) and an XLM-Roberta based model (Grɛberta; Riemenschneider & Frank, 2023). We found that while the models performed adequately metric-wise, the actual predictions still suffered from inconsistencies in the data. Further experiments (the use of a modified objective of only recognizing people on a word level, as well as the incorporation of gazetteers) are still ongoing; the outcome of these experiments will be discussed during the presentation.

References

Bamman, D., & Burns, P. J. (2020). *Latin BERT: A Contextual Language Model for Classical Philology* (arXiv:2009.10053). arXiv. https://doi.org/10.48550/arXiv.2009.10053

Beersmans, M., de Graaf, E., Van de Cruys, T., & Fantoli, M. (2023). Training and Evaluation of Named Entity Recognition Models for Classical Latin. In A. Anderson, S. Gordin, B. Li, Y. Liu, & M. C. Passarotti (Eds.), *Proceedings of the Ancient Language Processing Workshop* (pp. 1–12). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.alp-1.1

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Ehrmann, M., Romanello, M., Flückiger, A., & Clematide, S. (2020). Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum* (Vol. 2696). CEUR. https://ceur-ws.org/Vol-2696/#paper_255

Erdmann, A., Brown, C., Joseph, B., Janse, M., Ajaka, P., & Elsner, M. (2016). *Challenges and Solutions for Latin Named Entity Recognition*. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 85–93.

Mercelis, W. (2021). *Mercelisw/electra-grc · Hugging Face*. Retrieved March 17, 2023, from https://huggingface.co/mercelisw/electra-grc

Mercelis, W., & Keersmaekers, A. (2022). An ELECTRA Model for Latin Token Tagging Tasks. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 189–192. https://aclanthology.org/2022.lt4hala-1.30

Kemp, J., (2021, September 3). Beyond Translation: Building Better Greek Scholars. *Pelagios*. https://medium.com/pelagios/beyond-translation-building-better-greek-scholars-561ab331a1bc

Riemenschneider, F., & Frank, A. (2023). Exploring Large Language Models for Classical Philology. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15181–15199. https://doi.org/10.18653/v1/2023.acl-long.846

Singh, P., Rutten, G., & Lefever, E. (2021). A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, 128–137. http://hdl.handle.net/1854/LU-8726146

STEPBible Data Repository CC BY 4.0. (2023). [Computer software]. STEPBible. https://github.com/STEPBible/STEPBible-Data (Original work published 2018)

Yamshchikov, I. P., Tikhonov, A., Pantis, Y., Schubert, C., & Jost, J. (2022). *BERT in Plutarch's Shadows* (arXiv:2211.05673). arXiv. https://doi.org/10.48550/arXiv.2211.05673

Yousef, T., Palladino, C., Shamsian, F., d'Orange Ferreira, A., & Ferreira dos Reis, M. (2022). An automatic model and Gold Standard for translation alignment of Ancient Greek. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5894–5905. https://aclanthology.org/2022.lrec-1.634