NER on Ancient Greek with minimal annotation

Chiara Palladino, Farimah Karimi, Brigitte Mathiak

Introduction

This paper presents the results in the adaptation of a new workflow of Named Entity Recognition and classification applied to Ancient Greek. We used a model of language-independent data extraction and pattern discovery based on machine learning algorithms, which allowed the creation of a dataset of automatically classified place-names and ethnonyms starting from a small manually annotated list. We worked on the assumption that premodern textual sources display a recognized systematicity in their linguistic encoding of space, which provides a test-case for automatic context-based methods[1]. The idea is that we should be able to train the machine to recognize an entity from recurring elements in the context, without providing a large training dataset in advance.

Background

The problem of automatic Named Entity Recognition in inflected historical languages with scarce annotated data is very debated. Current text mining and NLP methods tend to work with high accuracy on modern Western languages, while non-Western and historical languages have a considerably less solid infrastructure. For Ancient Greek and Latin, many reasons affect the reliability of automated methods, including scarce stable services of lemmatization and named entity classification, issues with Unicode characters[2], and little data on place-names in the original language in gazetteers. Currently, several new approaches are being experimented with, most of which require a large manually annotated training dataset[3].

---

[1] C. Palladino, *New Approaches to Ancient Spatial Models: Digital Humanities and Classical Geography*, BICS 59.2 (2016), 56-70.

[2] J. Tauber, Character Encoding of Classical Languages. In M. Berti (ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, 137-158. De Gruyter Saur, Berlin-Boston, 2019.

[3] See for example: M. Berti, Named Entity Annotation for Ancient Greek with INCEpTION. In K. simov and M. Eskevich, *Proceedings CLARIN Annual Conference* 2019, 1-4. CLARIN 2019, Leipzig; A. Erdmann et al., Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 85-93. The COLING 2016 Organizing Committee, Osaka; T. Yousef, *Word Alignment and Named Entity Recognition applied to Greek text-reuses*, Masterarbeit. University of Leipzig, 2015. Recently on the topic of NLP infrastructure for Classical Languages P. J. Burns, Building a Text Analysis Pipeline for Classical Languages. In M. Berti (ed.), *Digital Classical Philology,* 159-176.

Method

We tested our approach on Herodotus' *Histories* (5th century BCE). We focused on detecting place-names and ethnonyms, and tested the model to verify if we could reliably train an algorithm to recognize and classify names based on the patterns where they appear. Our choice of Herodotus depended on the large number of available open data resources for this text: we used the XML text of the Perseus Digital Library (http://www.perseus.tufts.edu/hopper/), POS-tag information from the PROIEL Treebank (https://proiel.github.io/) and, for comparison, the place-names database of the Hestia project (http://www2.open.ac.uk/openlearn/hestia/), which, however, is based on the English version and does not include ethnonyms.

We created a manually annotated training dataset of 11 place-names and 11 ethnonyms (ca. 1000 occurrences in the text), including inflected forms, to avoid relying on a lemmatization service and to train the algorithm to recognize different endings. We extracted all the sentences[4] where a name from this list occurred together with another word (1046 sentences), working on the assumption that a frequently occurring name would often be mentioned in the same context as other, less frequent, names.

We used a Conditional Random Fields (CRF) algorithm[5] and trained it with various features, considering the word itself and a 1-token window preceding and following it: these included lower-case words, last three and last two letters, capitalization of the first letter, POS-tag, first letter of the POS-tag. We deliberately did not include the full word itself into the feature space, to avoid having the algorithm simply learn the words, we had used for training data generation and forcing it to learn general patterns from the context. We choose CRF as a robust, easy-to-configure alternative to more state-of-the-art Deep Learning[6], because of the easy interpretability it provides.

The "misclassified" tokens were inspected by our domain expert and newly found correct ethnonyms and toponyms were added to the list for a second run of the procedure above.

Results

Overall, the algorithm was able to correctly classify ethnonyms and place-names with considerable accuracy and allowed us to automatically find at first 202, then 182 ethnonyms and x+45 place names, starting from only 11 each. Not all the words identified by the algorithm were genuine places. The 34 wrong terms were mostly personal names with endings very similar to places or, in 3 cases, ethnics. For the ethnonyms, the yield was much better: 202 out of 209 tokens were correct ethnonyms, with only 7 incorrectly identified words, including some patronymics and adjectival toponyms, e.g. "the Ikarian sea").

The algorithm assigns the most weight on word-endings for positive identification, while it used context features for negative identification. For example, we observed that ethnonyms rarely

[4] We indexed the text according to the number of sentences, selecting strings ending with a full stop as sentence units (total 6754 sentences).

[5] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." *Foundations and Trends® in Machine Learning* 4.4 (2012): 267-373. As implemented by scikit-learn.org

[6] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436.

occur at the beginning of the sentence. We will propose a discussion of these results in the presentation.

Table 1: Results from the CRF NER cross-validation on the first/second run

	precision	recall	f1-score	support
normal word	0.99/0.99	0.99/0.99	0.99/0.99	19416/22585
ethnonym	0.70/0.76	0.72/0.77	0.71/0.77	628/856
toponym	0.88/0.87	0.85/0.84	0.87/0.86	608/633
accuracy			0.97/0.98	20652/24074
macro avg	0.86/0.87	0.85/0.87	0.85/0.87	20652/24074
weighted avg	0.97/0.98	0.97/0.98	0.97/0.98	20652/24074

Future Work

We could show that using NER with this method on the original Ancient Greek is feasible, even on very small data sets and with reasonable effort on manual annotations. We want to emphasize that the availability of open source material was crucial to our work, as we needed to rely on many different resources to process the text. Inspired by these projects, we chose to publish our results as well[7].

Using an incremental feedback loop of adding the newly found words to our gold standard and re-running the experiment, enabled us to improve the quality without large amounts of manual annotation. We were quite surprised by the sheer number of different ethnonyms and would expect even more by repeating the process further. The same methodology can be used for other named entities, such as personal names, by starting with a manually compiled list of high-frequency words and refining from there.

We also plan to look at the differences in patterns indicating places and ethnicities, especially in instances in which these could be used interchangeably, for example, “man from Athens” vs. “Athenian”, to derive information as to when Herodotus would choose one over the other. We also intend to compare our results to annotated translations. By using sentence-wise comparison, we can match a translation to the original and thereby find instances in which ethnicities have been translated as places, or vice versa.

[7] https://github.com/farikarimi/NERonAncientGreek