The iDAI.publication: extracting and linking information in the publications of the German Archaeological Institute (DAI)

Francesco Mambrini
Deutsches Archäologisches Institut
Podbielskiallee 69-71, Berlin
francesco.mambrini@dainst.de

Abstract

English. We present the results of our attempt to use NLP tools in order to identify named entities in the publications of the Deutsches Archäologisches Institute (DAI) and link the identified locations to entries in the iDAI.gazetteer. Our case study focuses on articles written in German and published in the journal *Chiron* between 1971 and 2014. We describe the annotation pipeline that starts from the digitized texts published in the new portal of the DAI. We evaluate the performances of geoparsing and NER and test an approach to improve the accuracy of the latter.

Italiano. *Il paper descrive i risultati dell'esperimento di applicazione di strumenti di NLP per annotare le Named Entities nelle pubblicazioni del Deutsches Archäologisches Institute (DAI) e collegare i toponimi identificati alle rispettive voci dell'iDAI.gazetteer. Il nostro studio si concentra sugli articoli in tedesco pubblicati nella rivista* Chiron *tra il 1974 e il 2014. Descriviamo la pipeline di annotazione impiegata per processare gli articoli disponibili nel nuovo portale per le pubblicazioni del DAI. Discutiamo i risultati della valutazione degli script di geoparsing e NER e, infine, proponiamo un approccio per migliorare l'accuratezza in quest'ultimo task.*

1 The iDAI.publications and the iDAI.world

The Deutsches Archäologisches Institute (German Archaeological Institute, henceforth DAI) is a German agency operating within the sphere of responsibility of the federal Foreign Office; the goal of the institue is to promote research in archaeological sciences and on ancient civilizations worldwide. Founded in Rome in 1829, the DAI has developed into a complex institution, with branches and offices located around the world. The Institute has participated in several projects, including missions of paramount importance like those in Olympia, Pergamon or Elephantine.

One of the most visible output of this activity is the amount of scientific publications produced by the DAI. The Institute currently publishes 14 international journals and 70 book series on different topics.[1] Since 2018, part of this collection is now accessible to the public on a new online portal named idai.publications for books and journals.[2] This ongoing initiative will not only enable researchers to have easier access to the published works; even more importantly, it will allow the Institute to integrate the data contained in articles and books (such as persons, places and archaeological sites, artifacts and monuments) into a network of all the other digital resources of the DAI.

All the digital collections of the DAI are indeed designed to operate within a network known as the idai.welt (or idai.world).[3] This network includes web collections such as "Arachne",[4] the database of archaeological monuments and artifacts of the DAI, and "Zenon",[5] the central bibliographic catalogue that serves all the libraries of the DAI offices around the world, but also compiles

[1] A list of journal is provided at: https://www.dainst.org/publikationen/zeitschriften/alphabetisch; for the list of book series: https://new.dainst.org/publikationen/reihen.

[2] See https://publications.dainst.org/journals/ and https://publications.dainst.org/books/.

[3] https://www.dainst.org/de/forschung/forschung-digital/idai.welt

[4] https://arachne.dainst.org/

[5] https://zenon.dainst.org/

some of the most comprehensive bibliographies in the areas of activity of the different branches.

The other cornerstone of the idai.world is represented by the layer of web-based services such as thesauri and controlled vocabularies. The idai.gazetteer,[6] in particular, connects names of locations with unique identifiers and coordinates; the gazetteer is intended to serve both as a controlled list of topnyms for DAI's services and to link the geographic data with other gazetteers. Unique identifiers defined in the idai.gazetteer are already used to connect places and entries in Zenon and Arachne. In this way, users of these services can already query monuments and artifacts in Arachne or books in Zenon that are linked to a specific place.

2 A pipeline for textual annotation

This network of references holds a great potential for the DAI publications. Places, persons, artifacts, monuments, and other entities of interest mentioned within the publications can be identified and linked to the concepts in the appropriate knowledge bases of the DAI. The linking of the different relevant entities would allow researchers not just to retrieve the texts that, independently from the language of the publication, make reference to certain concepts of interest, but also to study such epistemologically relevant questions as the variation in the patterns of locations cited in the studies across decades.

While the linking between entries in Zenon and Archne and the idai.gazetteer had been conducted manually, the volume and nature of the textual information to be processed in the publications encouraged us to turn to Natural Language Processing (NLP). We set up a pipeline for text annotation that aims to process the full texts of the publications, perform Named Entity Recognition (NER) to identify the mentions of the relevant entities, and finally link them to the appropriate entries in the idai.world.

We chose to build the first version of the pipeline around a series of open-source software that offer support for multiple languages and are widely used in the Digital Humanities (DH); at present, the annotation is limited to persons, places and organization, and only the linking of place-names to the idai.gazetteer is supported.

2.1 Preprocessing and NER

The pipeline is programmed in Python and takes advantages of modules of the NLTK platform for several task (Bird et al., 2009), like sentence- and word-tokenization.

The input of our annotation pipeline is, in the case of articles and books for which no other versions survive, the full text extracted from the PDF files of the articles.[7] The automatic recognition of the publication's main language is carried out by the Python library langid (Lui and Baldwin, 2011).

NER is performed using the Stanford Named Entity Recognizer (Finkel et al., 2005), which implements Conditional Random Field (CRF) sequence models. For a preliminary evaluation, we used pre-trained models for English, Spanish,[8] German (Faruqui and Padó, 2010), and Italian (Palmero Aprosio and Moretti, 2016). All these models are trained to recognize comparable classes of entities (persons, places, organizations and miscellaneous). We then chunked together the annotated tokens with a simple regular-expression chunker that takes consecutive, non-empty (O) tags together and labels them with the same label as the first token in the series.

Part-of-speech (POS) tagging, though not strictly necessary for NER and geoparsing, as the out-of-the-box models for Stanford NER do not require it, is also supported by our pipeline. TreeTagger (Schmid, 1999) was chosen since it offered a vast array of pre-trained models for many languages.

2.2 Geoparsing

The task of resolving place names by linking them to identifiers from a gazetteer is commonly referred to as "georparsing". The Edinburgh Geoparser[9] is a suite of tools that is often employed in DH (Grover et al., 2010; Alex, 2017) and allows users to preprocess texts, extract toponyms and resolve them by identifying the possible candidates in a gazetteer and scoring them. Users have the option to select between 4 gazetteers, and to set some parameters, like the coordinates of areas that will

[6] https://gazetteer.dainst.org/

[7] All the PDF files of the publications already include texts, so no Optical Character Recognition (OCR) is needed.

[8] Models for English and Spanish are available for download at https://stanfordnlp.github.io/CoreNLP/; for English we used the 4 Class model CoNLL 2003 English training set.

[9] http://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/html/

be given preference while ranking the candidates. The scoring process makes use of some properties recorded for places in gazetteers (e.g. the type of location, such as inhabited place or archaeological site) and especially by comparing locations pairwise with all other places identified; preference is thus given to places that cluster together.

Although Edinburgh works only with English and the `idai.gazetteer` is not supported, the CLI software is built as a suite of scripts, so that the input of a process is the output of the preceding one. By knowing the script that performs a task and the input it expects, it is therefore possible to inject a pre-processed text into any given step, while most processes (like scoring) are language-agnostic. We integrated the ranking script of Edinburgh within our pipeline to score, for any location that we extracted with our own NER scripts, any list of possible candidates matched in the `idai.gazetteer`.

3 Testing and Improving The Pipeline: a case study

In this section we discuss the preliminary results obtained by running the pipeline described above on the complete series of one journal now available in the `idai.publications`. The results will serve as a baseline for future improvement.

3.1 *Chiron*: the data set

The first complete publication series that was added to the portal was *Chiron*, a journal published by the DAI's "Kommission für Alte Geschichte und Epigraphik" from 1970. Volumes from 1 to 44 (2014) are currently available,[10] for a total of 942 articles. The focus of the publication is in Graeco-Roman history and epigraphy; several articles contain lengthy quotations (or even full editions) of inscriptions in Greek or Latin.

Table 1 reports the total number of articles per language. As can be seen, quotations in Greek and Latin are sufficiently frequent and long to confuse the automatic recognition. In 39 cases, Latin or Greek were considered the main language of the publication. Luxembourgish (a West Germanic language) is also a clear mistake for German, also possibly prompted by lengthy quotations (Nollé and Wartner, 1987, for one likely case). The 44 volumes of the journal show an interesting distribution of languages, with German playing the most relevant role by far.[11]

Language	Nr. Articles	Auto rec.
German	645	580
English	211	222
French	59	55
Italian	17	15
Spanish	10	12
Luxembourgish	0	19
Greek and Lat.	0	39

Table 1: *Chiron*: number of article per language (actual count vs automatically recognized)

3.2 Evaluating the annotation

In this preliminary stage, we decided to focus on the 580 automatically identified German articles in order to evaluate the performances of our pipeline and to improve its accuracy.

We have manually corrected the NER annotation and geoparsing of 4 articles (Linke, 2009; Hammerstaedt, 2009; Sänger, 2010; Haensch and Mackensen, 2011), for a total of 36,159 words. The articles were selected so as to represent a broad scope of subjects (from papyrology, to social and religious history, to military archaeology) and geographic areas (North Africa, Asia Minor, Rome and Italy).

For the evaluation of our NER tools we adopted the same metrics (precision, recall and $F_{\beta=1}$ score) and methods of the CoNLL-2000 shared task (Tjong Kim Sang and Buchholz, 2000). Note, in particular, that the scores are calculated at the level of the phrase, not of the single tag. The evaluation of the geoparser is also based on the same principles, but instead of evaluating its performances on the automatically annotated texts, we re-ran the geoparser on the gold-standard and evaluated that output.

The scores reported in Table 2 are considerably below the state of the art in NER for German, as documented e.g. in the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003). These results would very likely be considered insufficient or too noisy for the needs of researchers in the (Digital) Humanities.

[10] Readers are however requested to register an account.

[11] A word count on the automatically recognized languages confirms this conclusion: German has 7,394,004 words (60.48% of total), English 2,955,640, and French 899,888. Greek and Latin total 481,596 words; the other languages count between 193k and 148k words.

Entity	Precision	Recall	$F_{\beta=1}$
Person	73.21%	47.13%	57.34
Location	67.18%	34.56%	45.64
Organization	9.23%	35.71%	14.66
TOTAL	56.27%	43.22%	48.89

Table 2: NER: results of the first evaluation round; 1423 phrases; found: 1093; correct: 615

Modules for NER trained on general corpora do not seem to be suited to annotate texts that belong to such a specific domain with acceptable accuracy. The poor performances with organizations, in particular, point to some peculiarities of the archaeological literature in comparison to texts included in most general-use corpora: companies, firms and other institutions, which are frequent in the news, are rarely found in scholarly texts of our domain; the organization tag is more often reserved either to ancient institutions (like "the Roman Senate") or peoples and tribes ("the Aquitani") which are hardly represented in ordinary corpora.

Article	Precision	Recall	$F_{\beta=1}$
L09	76.53%	73.53%	75.00
H09	97.87%	95.83%	96.84
S10	72.66%	80.17%	76.23
H&M11	86.67%	74.71%	80.25
TOTAL	83.49%	79.13%	81.25

Table 3: Geoparsing: results per article; 575 phrases; found: 545; correct: 455. Articles: L09 (Linke 2009), H09 (Hammerstaedt 2009), S10 (Sanger 2010), H&M11 (Haensch and Mackensen 2011)

The performances of the geoparser, on the other hand, seem encouraging (Table 3). With gold-standar named entity recognition, the Edinburgh Geoparsers combined with the `idai.gazetteer` attained scores that closely approximate, or even surpass 80%. The evaluation of our annotation was also a valuable occasion to assess the accuracy and granularity of the `idai.gazetteer`: 38 locations in North Africa mentioned in one article (Haensch and Mackensen, 2011) did not have any record in DAI's gazetteer.

3.3 Applying in-domain NER models

We decided to use the manually corrected articles to see whether we could improve on the baseline with the help of in-domain models. We trained a CRF model adding a series of linguistic features, like POS, which may help capturing non-German expressions, or type-set features such as the use of small- and full-caps.[12] As the articles in *Chiron* focus on the Greco-Roman civilization, we expect a lookup in lists of known toponyms of the Ancient Word to sensibly improve the performances of NER for locations. We chose to add a gazetteer lookup to the list of features; we preferred to resort to a more specific resource like the "Digital Atlas of the Roman Empire" (DARE)[13] instead of the general-purpose `idai.gazetteer`.

Entity	Precision	Recall	$F_{\beta=1}$
Person	80.00%	71.41%	75.30
Location	76.26%	58.90%	65.87
Organization	22.02%	23.08%	16.94
TOTAL	79.32%	65.75%	71.75

Table 4: NER: results of the in-domain model; average scores of 10-fold cross-validation

Table 4 reports the results of this second round of testing, which was conducted using the same methodology as before and performing a 10-fold cross-validation. As can be seen, the in-domain model considerably improves over the baseline. The performance with organizations is still largely insufficient, mainly on account of the scarcity of examples (70 phrases, vs 970 persons, 387 locations). The improvement with locations is significant, but the overall performance still leaves room for substantial improvement.

4 Conclusions and future work

The use of in-domain CRF models trained specifically for the target journal and adopting a specialized gazetteer for place names improves on the baseline of the out-of-the-box NER tools in our initial pipeline. It is likely that the accuracy on the *Chiron* data can be further increased with additional training. Given that an accurate recognition is a prerequisite for geoparsing, we plan to con-

[12] The CRF implementation that we used is provided by the Python library `sklearn-crfsuite` (0.3.6).

[13] `http://dare.ht.lu.se/`

centrate our effort on the NER components. We intend to progress in the direction discussed above, in particular by: a. training and evaluating models for the other languages (French, English, Italian, Spanish) b. testing the models on other publications in the portal.

In a more distant future, we also intend to include support to the identification (and subsequent linking) of other named entities of interest for archaeologists, such as artifacts, monuments and chronological references.

References

Beatrice Alex. 2017. Geoparsing English-Language Text with the Edinburgh Geoparser. https://programminghistorian.org/en/lessons/geoparsing-text-with-edinburgh.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly, New York.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368:3875–3889.

Rudolf Haensch and Michael Mackensen. 2011. Das tripolitanische Kastell Gheriat el-Garbia im Licht einer neuen spätantiken Inschrift: Am Tag, als der Regen kam. *Chiron*, 41:263–286.

Jürgen Hammerstaedt. 2009. Warum Simonides den Artemidorpapyrus nicht hätte fälschen können: Eine seltene Schreibung für Tausender in Inschriften und Papyri. *Chiron*, 39:323–338.

Bernhard Linke. 2009. Jupiter und die Republik. Die Entstehung des europäischen Republikanismus in der Antike. *Chiron*, 39:339–358.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Johannes Nollé and Sylvia Wartner. 1987. Ein tückischer Iotazismus in einer milesischen Inschrift. *Chiron*, 17:361–364.

A. Palmero Aprosio and G. Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*.

Patrick Sänger. 2010. Kommunikation zwischen Prätorianerpräfekt und Statthalter: Eine Zweitschrift von IvE Ia 44. *Chrion*, 40:89–102.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Processing*, pages 13–26. Kluwer Academic Publishers, Dordrecht.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 127–132, Stroudsburg, PA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.