Archaeological Entity Recognition for Information Retrieval in Dutch Archaeological Reports

Alex Brandsen
Faculty of Archaeology, Leiden University
Leiden, The Netherlands
a.brandsen@arch.leidenuniv.nl

Suzan Verberne
Leiden Institute of Advanced Computer Science
Leiden, The Netherlands
s.verberne@liacs.leidenuniv.nl

ACM Reference Format:
Alex Brandsen and Suzan Verberne. 2017. Archaeological Entity Recognition for Information Retrieval in Dutch Archaeological Reports. In *Transactions of the DIR2017 conference*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROPOSAL

In this demonstration we will present AGNES; Archaeological Grey literature Named Entity Search, an online search tool aimed at unlocking the information hidden in archaeological reports.

Over 60.000 Dutch archaeological reports are available online, and this number is growing by around 4.000 a year. The main reason for the existence of these reports is the Malta Convention [4]; a European agreement, aimed at protecting archaeological remains. Every research project that is performed has to be documented and deposited according to Dutch law [1], which has created a collection of grey literature too vast to comprehend. Many of these reports threaten to end up in a proverbial graveyard, unread and unknown. However, the data contained in these reports are of immense value, and this hidden knowledge can be very useful in research, if researchers can find and access the specific information they need from this big data collection.

Currently it is only possible to search through the metadata of these documents, mainly via the DANS (Data Archiving and Networking Services) repository.[1] However, these metadata are often of poor and inconsistent quality, and generally do not describe the contents of a report well. Also, an archaeologist will generally want to search more fine-grained, and might be interested in what is known as the 'by-catch opportunity'; i.e. a single Bronze Age find in a Medieval excavation, not mentioned in the metadata. There is a strong need for a better way to search through these documents [5, 6, 8]. Also, archaeologists are eager to use multiple aspects in their searches; an example query might be to find all documents relating to the Iron Age, from a particular geographical area, that mention cremations. This is currently possible via DANS, but it is difficult and inaccurate.

To effectively index these texts, Named Entity Recognition (NER) is needed to correctly identify and distinguish between entities. Standard approaches to NER, and NER in related fields such as history, are insufficient to deal with the peculiarities of archaeological concepts and the wealth of potential classes. Some of the challenges include non-standard naming, extensive polysemy & synonymy, and complex word formation, including different spellings, entities being one or more words, and concepts including capitals, numbers and symbols. This is particularly true for archaeological time periods, which can be expressed in numerous ways. For example, the following entities all equate to roughly the same time period: Neolithic, Swifterbant culture, Early Neolithic, New Stone Age, 3500 v.Chr, 5000 to 4000 BP and 4915 ±40 Cal BC.

Some research has already been done on NER in archaeological texts in e.g. English [2, 3] and Dutch [7, 9], but these are not combined with full-text search, or tend to focus on limited entity types, and not the full breadth of archaeological concepts, which includes artefact, time period, place, material, ground context and monument. This means that currently there is no working system in place for Dutch archaeology.

The first version of AGNES is currently online at http://agnessearch.nl.[2] Our demonstration will present the next phase of AGNES, in which Conditional Random Fields trained on manually annotated data are used to perform NER on archaeological reports. These entities are combined with a full-text index to create an effective online search tool, with an intuitive user-led front-end in the pipeline for future versions.

REFERENCES

[1] 2015. Erfgoedwet. (2015). http://wetten.overheid.nl/BWBR0037521/2016-07-01
[2] A Amrani, V Abajian, and Y Kodratoff. 2008. A chain of text-mining to extract information in archaeology. *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008.* (2008), 1–5. http://ieeexplore.ieee.org/abstract/document/4529905/
[3] K Byrne and E Klein. 2010. Automatic extraction of archaeological events from text. *Proceedings of Computer Applications and Quantitative Methods in Archaeology* (2010).
[4] Council of Europe. 1992. European Convention on the Protection of the Archaeological Heritage (Revised). (1992). http://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/143
[5] Monique van den Dries. 2016. Is everybody happy? User satisfaction after ten years of quality management in European archaeological heritage management. In *When Valletta meets Faro. The reality of European archaeology in the 21st century*, P Florjanowicz (Ed.). Brussels, 126–135.
[6] Harry Fokkens, Bastiaan Steffens, and Stijn van As. 2016. *Farmers, fishers, fowlers, hunters.Knowledge generated by development-led archaeology about the Late Neolithic, the Early Bronze Age and the start of the Middle Bronze Age (2850 - 1500 cal BC) in the Netherlands.* Rijksdienst voor het Cultureel Erfgoed, Amersfoort.
[7] H. Paijmans and A. Brandsen. 2010. Searching in archaeological texts: Problems and solutions using an artificial intelligence approach. *PalArch's Journal of Vertebrate Palaeontology* 7, 2 (2010).
[8] Julian Richards, Douglas Tudhope, and Andreas Vlachidis. 2015. Text Mining in Archaeology: Extracting Information from Archaeological Reports. In *Mathematics and Archaeology.* CRC Press, 240–254. https://doi.org/10.1201/b18530-15
[9] A Vlachidis, D Tudhope, M Wansleeben, J Azzopardi, K Green, L Xia, and H Wright. 2017. *D16.4: Final Report on Natural Language Processing / Resources / Ariadne - Ariadne.* Technical Report. ARIADNE. http://www.ariadne-infrastructure.eu/Resources/D16.4-Final-Report-on-Natural-Language-Processing

[1] https://easy.dans.knaw.nl/

[2] Version 0.1, full text search only on 100 randomly selected reports

DIR2017, November 2017, Hilversum, The Netherlands
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn