Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports

Brandsen, A.

1 Introduction

"*A library serves no purpose unless someone is using it.*"
Mr. Atoz, Star Trek TOS, s03e23 'All Our Yesterdays'

In the last decade, archaeology has joined other disciplines and has started generating what is known as 'big data': "Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" (De Mauro *et al.*, 2015, p. 103). The challenge is how to best mine, mix and analyse these incredibly rich data sets. While a lot of research in this area is dedicated to processing structured information – such as spatial information, databases, and so on – less attention has been given to processing unstructured information: the texts describing archaeological research (Bevan, 2015).

Easy access to the information hidden in these texts is a substantial problem for the archaeological field. Making these documents accessible, searchable and analysing them is a time consuming task when done with the existing metadata search systems or by hand, and will generally lack consistency. In addition, in the last fifteen years or so, the amount of archaeological texts in general – and fieldwork reports specifically – have seen an explosive growth. It is practically impossible to keep up with the rate of documents being produced, and the available literature is so extensive that the current search systems are not effective enough for detailed search. Text Mining[1] provides methods for automatically extracting meaningful information from these large data sets, allowing researchers to locate texts relevant to their research questions, as well as being able to identify patterns in the literature (Richards *et al.*, 2015).

This dissertation will describe the use of Text Mining techniques on Dutch language grey literature (further defined in Section 2.2.2): field reports from excavations in the Netherlands, deposited in the Data Archiving and Networked Services (DANS) digital archive. More detailed information on the data set will be provided in chapter 2 and 3.

1.1 Research motivation

The work carried out in this project is motivated by the need of researchers in the archaeological field to be able to efficiently find information related to their research questions in the available literature. This requirement has been well documented around the globe (e.g. Richards *et al.*, 2015; Van den Dries, 2016; Habermehl, 2019) and some studies have investigated different applications of Text Mining in archaeology in English (Vlachidis & Tudhope, 2016; Amrani *et al.*, 2008; Byrne & Klein, 2010), French (Mélanie-becquet *et al.*, 2015) and

[1] Please see the Glossary at the end of this document for a definition of this term, as well as other technical terms and acronyms used throughout this dissertation.

Dutch (Paijmans & Brandsen, 2010; Vlachidis *et al.*, 2017, also see section 2.4).

However no system is currently in place that allows easy access to the full text in the Dutch archaeological corpora (document collections), meaning that relevant and valuable information is not being utilised by researchers. This is a problem from a theoretical point of view, as key information that is currently being overlooked could change archaeological interpretations, but it also devalues the monumental effort that has gone into collecting, digitising, archiving and publicising these documents, as well as the legislation that has been drawn up surrounding the archiving of these documents. The scientific value of these reports will be further discussed in section 2.2.4.

More and more text and data mining tools and techniques have become available over the last years, which provide a way to access and extract information from this wealth of information currently hidden in the text data of these reports. This, combined with the relatively easy access to higher computer processing power available to us now (see section 2.5.2), makes a systematic implementation in Dutch archaeology not only feasible, but also highly desirable.

The end goal of this project is to develop a search engine that combines Text Mining techniques with a full text search, allowing archaeologists to search through archaeological reports stored in the DANS repository (also see section 2.2.3). The search system is named AGNES, and will be further discussed in the following chapters.

1.2 Research questions

When creating a tool for archaeologists, it is important to ensure it can positively impact their work. Unfortunately, many digital tools in archaeology have been created, and subsequently ended up unused in a corner of someone's web server, or even worse, not available online at all, such as the OpenBoek project (Paijmans & Wubben, 2008) and the work by Tudhope *et al.* (2011) (also see section 2.4). To ensure an impact is made, it is required to investigate the user requirements, as well as the effectiveness and usability of the developed system, but also to evaluate AGNES's output using a real-world case study.

Main Research Question: To what extent can a search engine using Text Mining improve archaeological research and aid information discovery in grey literature data sets?

To answer this question, the following subquestions have been formulated:

1. Can we use existing labelled data sets for Named Entity Recognition[2] in the archaeological domain, or do we need to create our own data set? If so, to what extent does the accuracy increase? (Chapter 3)
2. To what extent can we automatically generate time period and site type metadata for Dutch excavation reports? (Chapter 4)
3. Which questions do archaeologists want to ask of this data set, and which user requirements do they have for a search system? (Chapter 5)
4. What user interface features of AGNES are experienced as positive or negative, and how can we optimise the usability of the system for archaeologists? (Chapter 6)
5. To what extent does adding more domain-specific training data to BERT models improve Named Entity Recognition accuracy? (Chapter 7)
6. What is the impact of the developed system on archaeological research? (Chapter 8)

Sub questions 3 & 6 are closely linked, as the impact of AGNES will be evaluated using a case study. The case study research question is defined in chapter 8.

1.3 Research Methodology

The research described in this dissertation is relatively varied, and as such a number of different research methodologies are combined. For the development of the search engine, we apply the Agile principles, which are described below. To evaluate classification methods, we use human labelled data, and to evaluate the system functionality we use a focus group and case study.

1.3.1 Agile Development Principles

To define the agile development principles, and to explain why they are relevant to this project, it is useful to draw an analogy to a non-programming project. Highsmith *et al.* (2002) uses a battle as an example: a commander will plan extensively, but also realise that plans are just the beginning. Probing enemy defences (creating change) and responding to enemy actions (responding to change) are more important. A commander is successful by defeating the enemy (the mission), not by strictly conforming to a plan. Battlefields are uncertain, constantly changing, and turbulent, so planning everything up front simply isn't feasible. A

[2]Further defined in section 2.3.3

similar parallel can be drawn to archaeological excavations: although they are planned to a high degree, initial findings in the field can massively change the approach taken during excavation.

Both the battlefield and excavation scenarios are examples of projects with a relatively clear mission, but the specific requirements to complete that mission are (partly) unknown, volatile, and constantly evolving as change unfolds. Many programming and research projects are of a similar nature, where the full extent of the work is not known from the onset. These projects are dubbed "high-exploration factor projects" (Highsmith *et al.*, 2002, p.4), and this is where agile approaches are most suitable, as they can not simply be completed by plan-driven methods. The more experimental the technology and the more volatile the requirements are, the more agile development improves the chance of success.

This project certainly falls under this category as well: while the mission is clear (to disclose archaeological reports), the specific techniques that should be used to attain this goal and the exact user requirements were not clear at the start of the project. As this makes it a prime candidate for agile development, it has been decided to use this methodology for this project.

As a more formal definition, agile development is a combination of a philosophy and a set of development guidelines. The philosophy encourages user satisfaction, quick, incremental development of systems, minimal development work products, and overall development simplicity. The guidelines describe iterative feature-driven cycles, delivery over analysis and design, and continuous communication between the developer and the end users, often by using user focus groups (Pressman, 2005).

In the Manifesto for Agile Software Development, Beck *et al.* (2001) describe four main ideas to define agile development:

- **Individuals and interactions** over processes and tools
- **Working software** over comprehensive documentation
- **Customer collaboration** over contract negotiation
- **Responding to change** over following a plan

This describes the agile mindset in general terms. In more practical terms, in the software development in this project:

- The first version of AGNES shall be a proof of concept, with only basic capabilities, to act as a starting and discussion point, enabling better feedback
- The system shall be developed in small cycles, each leading to a working prototype that can be demonstrated and assessed

- After each cycle, a user panel will provide feedback on AGNES to integrate into the next development cycle. More information on the user panel can be found in section 1.3.3

While this specifically describes the software development process, the rest of the research is undertaken with the same principles in mind. This makes the research more suited to be published in papers than a monograph, which is further detailed in section 1.5 below.

1.3.2 Machine Learning Assessment with Labelled Data

To develop and evaluate supervised machine learning technologies, we use human-labelled data. In Chapter 3 we describe the problems with the existing labelled data set for Named Entity Recognition, and how we created a new one which led to better results for this task. This labelled data is used to evaluate the work in Chapters 5 and 7. For the document classification task in Chapter 4, we created a labelled data set by converting the manually added metadata to a controlled list of labels. For any of these tasks, we report the recall, precision and F1 score, as further described in section 2.3.4.

1.3.3 System Evaluation by Focus Group

In this project, a focus group of potential users has been assembled, including academic researchers, PhD students, commercial archaeologists, and archaeologists working at the government level. An initial meeting has been organised with the group to synthesise a list of user requirements: the objectives for the system (chapter 5). Based on these requirements, the first iteration of AGNES has been created. The user interface of this version was evaluated, and the feedback was used to improve the system (chapter 6). Finally, we evaluated the quality of results retrieved by AGNES in a case study (chapter 8).

1.3.4 System Evaluation by Case Study

There is an abundance of research questions that could be answered if archaeologists had easy access to the full text of the reports. One example is presented by the 'by-catch opportunity': many excavations focus their research questions on a specific time period (e.g. on a Roman cemetery), but often also reveal objects and features from other periods (e.g. a small cluster of flint objects or a single stone axe from the Stone Age) or other types of contexts (e.g. a single residential

find). These other finds will always be presented in the publication, but will escape the attention of Stone Age archaeologists since they probably would ignore a report titled 'The Roman cemetery at Vlodrop (Limburg)'. However, these individual finds and small clusters do express a very valuable component of human behaviour, often called off-site (or off-settlement) activities (Foley, 1981).

In this project, Femke Lippok – a PhD researcher at Leiden University – has formulated an archaeological research question, which will be used as a case study, or test case, for the system. The first stage is to create a set of baseline results using current approaches, i.e., using existing search systems and the archaeologists' knowledge of their field.

Once this baseline is established, it is possible to analyse and compare this to the results obtained by AGNES qualitatively (comparing the interpretations) as well as quantitatively, via statistical and geographical analyses of the resulting document sets. This way we can demonstrate the increase in knowledge discovery by using the system, as well as the change caused to the archaeological view on a particular research question by being able to integrate more knowledge into research. More information on the case study can be found in chapter 8.

1.4 Contributions

Of course, the main contribution of this research is the AGNES search system, which is currently online and being used by archaeologists for their literature research. Besides AGNES, we also contribute a number of data sets, software and language models that can be used in other research, and we end this section with an overview of scientific contributions.

1.4.1 Software and Data

An application-oriented research project such as this inevitably produces resources that can be used by other researchers. Below is a list of the most prominent data sets and software shared publicly:

1. A manually annotated training data set for Named Entity Recognition (NER) in the archaeology domain (doi.org/10.5281/zenodo.3544543)
2. A training data set for the classification of archaeological reports on time period and subject (doi.org/10.5281/zenodo.3676702)
3. A JavaScript Object Notation (JSON) export of all the entities extracted from our corpus (doi.org/10.17026/dans-zcs-7b72)

4. A trained Conditional Random Fields (CRF) model for Dutch archaeological NER, with code to generate such models (doi.org/10.5281/zenodo.1238860)
5. A BERT model further pre-trained on our corpus, called ArcheoBERTje, and a specific model for NER inference, hosted on HuggingFace for ease of use (huggingface.co/alexbrandsen)

1.4.2 Scientific Contributions

Besides the resources described in the previous section, we make the following unique contributions to the scientific field.

In Chapter 3 we show that annotation of a NER data set with rigorous annotation guidelines, tailored to machine learning, leads to higher performance than a previously available data set. We also argue that for NER data, the pairwise F1 score between annotators is a better indicator of Inter Annotator Agreement than the commonly used Cohen's Kappa.

In the following chapter (4), we investigate the difficult task of multi-label text classification with many classes. We are the first to do this in the archaeology domain, and show that this method can contribute to either faceted search (by filtering documents by topic) or even metadata assignment at the time of deposition in an archive.

In Chapter 5 we show the need for a system such as AGNES, and make a case for the adoption of user requirement solicitation and short development cycles for digital tools in the archaeology domain. We also present our CRF based NER method which outperforms previous rule-based approaches.

The usability evaluation of our search system described in Chapter 6 is (as far as we know) the first of its kind in the archaeology domain, and we contribute to the general discussion of information needs in archaeology. We show the importance of a diverse group of test users, and argue that usability evaluation should be a core part of tool development.

Chapter 7 describes our work on the use of BERT language models for NER in Dutch archaeology. We present the first Dutch domain specific BERT model, which is also the first archaeology specific BERT model. We show that adding language-specific and domain-specific training data to an existing language model (by further pre-training) increases the performance of the model.

Perhaps the most important contribution for archaeologists is described in Chapter 8: in this case study we show that for Early Medieval cremations, using AGNES increased the amount of sites known to experts by 30%. This indicates that this type of search through grey literature can lead to more efficient and

more detailed research.

And finally, in Chapter 9, we contribute to the discussion on development-led archaeology more broadly, and how computational tools might solve existing problems and shape future research.

1.5 Dissertation outline

This dissertation consists of a collection of papers, sandwiched in between the introduction / background chapters and a discussion chapter. A majority of the papers have already been published in – or submitted to – peer-reviewed journals and conference proceedings during the course of the PhD. The papers are not in chronological order of publication, but in the order that makes the most sense for the narrative. Each paper can be read independently from the other chapters.

In the following Chapter (2), we will give an overview of the current state of affairs in Digital Archaeology, grey literature, and the value of excavation reports. We will also introduce Text Mining techniques, so the following chapters can be understood by anyone. Finally, we present previous research on Text Mining in archaeology, and the resources we use for this research.

In Chapter 3 we discuss the difficulties with an existing training data set for Named Entity Recognition, and how we have created a new data set with rigorous guidelines that improves the accuracy (Brandsen *et al.*, 2020).

Chapter 4 describes the work in collaboration with Martin Koole, where we trained a number of models to automatically classify excavation reports in subject and time period categories. This information can then be used for faceted search: allowing users to filter documents based on these categories (Brandsen & Koole, 2021).

The following chapter (5) describes the user requirements solicitation process, where we held a workshop with users to determine what features they would like when searching through excavation reports. The results of this process are the basis for how we developed the search system. We also describe the first version of AGNES (v0.1), and how it was used to elicit more feedback from the users (Brandsen *et al.*, 2019).

In Chapter 6, we evaluate the user interface created based on the input of the previous chapter. We specifically look at how quickly the users learn the interface, and which interface components are experienced as positive and negative. The outcomes have been used to improve the search system (Brandsen *et al.*, 2021b).

Chapter 7 describes how we experimented with different BERT language models to perform Named Entity Recognition, and how adding more domain-specific

training data increases the accuracy (Brandsen *et al.*, 2021a).

The case study is described in Chapter 8, where we worked together with Femke Lippok to investigate Early Medieval cremations. We used AGNES to retrieve relevant excavation reports, and assessed to what extent these documents are new information to the researcher and to what extent her view of this topic changed.

In Chapter 9 we discuss the results, and what these mean in a wider context. We then provide answers to the research questions in the conclusion, and end with proposed future research.