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

Brandsen, A.

5

User Requirement Solicitation

"*Improve a mechanical device and you may double productivity. But improve man, you gain a thousandfold.*"
Khan Noonian Singh, Star Trek TOS, s01e22 'Space Seed'

Previously published as: Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M., 2019. User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology*, 2(1), pp.21-30. DOI: 10.5334/jcaa.33

In this paper, we present the first results of applying Named Entity Recognition and Information Retrieval techniques to tackle the problem of unused grey literature in archaeology, specifically Dutch excavation reports. We used Conditional Random Fields to identify entities, with an average accuracy of 56%. This is a baseline result, and we identified many possibilities for improvement. These entities were indexed in ElasticSearch and a user interface was developed on top of the index. This proof of concept was used in user requirement solicitation and evaluation with a group of end users. Feedback from this group indicated that there is a dire need for such a system, and that the first results are promising.

5.1 Introduction

The archaeological world creates huge amounts of text in different formats, from books and scholarly articles to unpublished fieldwork reports. These reports are also known as grey literature. Easy access to the information hidden in these texts is a substantial problem for the archaeological field. Making these documents searchable and analysing them is a time consuming task when done by hand, and will often lack consistency. Text Mining and Information Retrieval (IR) provide methods for disclosing information in large text collections, allowing researchers to locate (parts of) texts relevant to their research questions, as well as being able to identify patterns of past behaviour in these reports (Richards *et al.*, 2015).

The Malta convention (or Valletta Treaty) is a European treaty, signed on 16 January 1992. It came into effect on 25 May 1995, and its aim is to protect archaeological remains by making "the conservation and enhancement of the archaeological heritage one of the goals of urban and regional planning policies" (Council of Europe, 1992, Art. 1). The convention was implemented in the Netherlands via the Archaeological Heritage Management Act in 2007 (Ministerie van Onderwijs Cultuur en Wetenschap, 2007). Preferably, preserving these remains is done by keeping them in situ, but when this is not possible, the developer disturbing the ground record is required by law to pay for the archaeological research. This research is generally performed by commercial archaeology units.

This archaeological research has created a collection of texts that is too large to be completely read by humans. The amount of reports created in the last 20 years is currently estimated at just under 60,000, and is growing by approximately 4000 per year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Most of these reports are categorised as 'grey literature' (Evans, 2015), and are likely to end up in a proverbial 'graveyard', unread and unknown, unless they are properly

archived, indexed and disclosed.

In the Netherlands, the SIKB (*Stichting Infrastructuur Kwaliteitsborging Bodembeheer*) creates and maintains the standards of activities relating to soil management. As stipulated in their BRL 4000 guidelines, a report has to be deposited into an e-depot within 2 months of completing the project (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016, Art. 2.6.2). While some companies and municipalities are still reluctant to deposit their reports into national e-depots (instead opting to deposit in small local depots) most reports and the associated metadata do end up in one of three of the main e-depots of the Netherlands; the Data Archiving and Networked Services (DANS) repository, the Document Management System of the *Rijksdienst voor Cultureel Erfgoed* (RCE) or the *Koninklijke Bibliotheek* (KB) e-Depot. There is considerable overlap between the DANS, RCE and KB data sets, and altogether it is estimated they hold around 70 percent of all so-called Malta reports. This means that a large portion of the reports is currently available, and access to the files is not a major problem at the moment.

This paper describes the work carried out in the first year of a PhD project. This project is in association with both the Faculty of Archaeology and the Data Science Research Programme (DSRP) at the University of Leiden, combining archaeological knowledge with the technical skills available in the Data Science department.

The work carried out in this project is motivated by the need from researchers in the archaeological field to be able to efficiently and effectively find information related to their research questions in the available grey literature. This requirement has been well documented in previous work (e.g. Richards *et al.*, 2015; Van den Dries, 2016) and some studies have investigated different applications of Text Mining from archaeological reports in English (Vlachidis & Tudhope, 2016; Amrani *et al.*, 2008; Byrne & Klein, 2010) and Dutch (Paijmans & Brandsen, 2010; Vlachidis *et al.*, 2017).

However no system is currently available that allows full-text access to at least a major part of the Dutch archaeological corpus, or document collection. As a result, relevant and valuable information is not being utilised by some researchers, mainly those who are not experts in their field yet. Information like a single Bronze Age find in a otherwise Medieval site is unlikely to be mentioned in the metadata, and is thus nearly impossible to find. This is a problem from a theoretical point of view, as key information could be overlooked at the moment, information that could change archaeological interpretations. It also devalues the monumental effort that has gone into collecting, digitising, archiving and publishing these documents, as well as the legislation that has been drawn up surrounding the archiving of these documents.

More and more Text Mining, data mining and IR tools and techniques have become available over the last years, which could potentially provide a way to access and extract information from this wealth of data currently hidden in these reports. This, combined with the relatively easy access to higher computer processing power, makes a systematic implementation of Text Mining techniques for Dutch archaeological reports not only desirable, but also feasible.

In this project we are developing AGNES (**A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch), a search system that aims to make archaeological grey literature more accessible and searchable by applying IR techniques to this big data set.

The goals of this paper are (1) to give an overview of previous work on Text Mining in archaeology, (2) to show the need for a search system by interviewing the user group, (3) soliciting user requirements for such a system, (4) presenting the results of the initial experiments with Named Entity Recognition (NER) and (5) presenting the indexing and front end software of the developed system.

5.2 Prior work

Some experiments have been carried out in Text Mining and NER in archaeology, across multiple countries and languages. In English, one of the earliest contributions is the work by Amrani *et al.* (2008), which helped experts to extract information from archaeological literature. Byrne & Klein (2010) also investigated the extraction of information, but focused solely on event information. The OPTIMA system, described by Vlachidis (2012), used a rules-based approach to semantic indexing, including NER. Another notable project is Archaeotools in the UK, which combined databases with information extracted from reports in an interesting faceted browser interface (Jeffrey *et al.*, 2009). A more recent paper is that by Kintigh (2015), which provides a detailed overview of the problems and possible solutions, but does not include the development of a search system.

For Dutch language reports, most of the previous research has been carried out by Paijmans with several collaborators, including extracting monument names from free text fields (Paijmans & Brandsen, 2009) and the OpenBoek system, which used memory-based learning to perform NER (Paijmans & Wubben, 2008; Paijmans & Brandsen, 2010). Like the work by Byrne & Klein (2010), this project focused mainly on time periods, but also applied some rules-based NER to detect place names. The OpenBoek system included an online search interface during the Continuous Access To Cultural Heritage (CATCH) project, but unfortunately this isn't available anymore.

A notable contribution is that by Mélanie-becquet *et al.* (2015), who ran a pilot study on texts from a part of France, dealing with the Iron Age till the Medieval period. They performed NER and other techniques similar to some of the previously discussed projects, but they did this multi-lingually, including French, German and English. Unfortunately, the technical details of their work don't seem to be published yet.

More specifically, this project builds upon the Text Mining experiments performed by researchers of the University of South-Wales in the European ARIADNE project between 2013 and 2017. They applied a rules-based technique to the problem, utilising the GATE framework[1]. Leiden University participated in this project and a limited number of eight Dutch reports were analysed and compared to manually tagged 'gold-standard' documents as a proof of concept, next to English, Swedish and German reports. In the same project, the ADS (Archaeological Data Service) in the UK applied machine learning techniques to English grey literature, and developed an API that can automatically create metadata based on entered text (Vlachidis *et al.*, 2017).

The contributions of this paper compared to previous work are twofold: (1) this system includes a user study which hasn't previously been undertaken; and (2) it combines the results of the NER with a full-text index in an effective search interface, instead of just focusing on the NER.

More broadly, this project is in cooperation with the DSRP, which gives us access to a high computing power cluster, allowing for the use of more computationally expensive techniques on bigger document sets. The length of this project is also of importance; most previous experiments were often performed over a short amount of time, making it difficult to create a finished system, while this project takes place over four years with the specific aim of creating a user-friendly web application.

5.3 Introducing AGNES

AGNES is an acronym that stands for **A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch, and is the name of the search system currently under development in this project, including both the front end of the web application, as well as the indexing software responsible for finding and indexing archaeological concepts. The current version of the system (v0.2) is available at `http://agnessearch.nl/index.php/search/agnesv02`.[2]

[1] See also https://gate.ac.uk

[2] Please note, free registration is needed to access the system.

	Synonomy		**Polysemy**
Main Term	*Neolithic*	**Main Term**	*Swifterbant*
Synonyms:	Late Stone Age	**Meanings:**	Time Period
	3000 BC		Excavation
	5000BP		Pottery Type
	4th Millenium BC		Location

Table 5.1: Synonymy and Polysemy examples

5.3.1 Named Entity Recognition

A standard full-text index, allowing researchers to search through all of the text instead of just the metadata, would already be an improvement on the current situation. However, such a full-text search would not account for synonymy and polysemy; multiple words that have the same meaning and one word having multiple meanings, respectively. See table 5.1 for two non-exhaustive examples, where a full-text search would either not return all results, or return possibly wrong results. This is why NER is needed to accurately index these documents.

Named Entity Recognition is a method that aims to identify and classify specific entities in natural language, also known as unstructured written text (Marrero *et al.*, 2013). In the case of this project, the entities are archaeological concepts, and the natural language are excavation reports. To give an example, in the following sentence the entities are underlined: "We found pottery dating from the Neolithic inside a rubbish pit", an artefact, a time period and a feature, respectively.

In the current version of the system, we used Conditional Random Fields (CRF). This is a form of machine learning specifically designed to label sequence data (Lafferty *et al.*, 2001), a common choice for NER tasks as words in a sen-

Figure 5.1: AGNES Logo

tence are sequential. We implemented the scikit-learn Python package (Pedregosa *et al.*, 2011), using the default algorithm (gradient descent using the L-BFGS method). The input for this algorithm were manually tagged Dutch reports (also known as a 'gold standard') created in the ARIADNE project (Vlachidis *et al.*, 2017), specifically selected to be a good sample of the corpus. In total, this training set consists of roughly 500,000 words, containing 11,000 tagged entities. Some issues with these documents are discussed later in this section.

These .docx files were tokenised and Part Of Speech (POS) tagged[3] using Frog (Van den Bosch *et al.*, 2007) and then converted to the FoLiA XML format (Van Gompel & Reynaert, 2013). Subsequently, the documents were converted to the format scikit-learn requires; a list of tokens including the token's POS and category (or concept) tag. At the moment, only three archaeological categories are used: artefact, time period and material, although more categories will be added in later versions. For each token, the following features were extracted for the word itself, as well as the word before and after the current one:

- Word in lowercase
- Word starts with uppercase character
- Word is all uppercase
- Word is all numbers
- Part of speech tag
- Word exists in materials wordlist
- Word exists in periods wordlist
- Word exists in artefacts wordlist
- Word is beginning or end of sentence

This is a fairly simple list, and is purely meant to provide a baseline result. As such, it was expected that the accuracy of the NER would not be very high.

To evaluate the results of the NER, a leave-one-out eight fold cross validation was done, meaning that the algorithm is run eight times, each time using seven of the documents as a training set, and using one document to test the model. It rotates through all eight possible combinations, and then calculates an average of the accuracy of the model. The total averaged accuracy (F1 score) is 56%, with the results for the different categories presented in table 5.2. As can be seen from this table, the average precision is fairly high at 71%, but the recall is much lower at only 48%.

[3]Tokenisation is the process of converting a character sequence (text) to individual tokens (words and punctuation). POS tagging is assigning a grammatical part of speech to each token, such as noun, verb, and so on.

	Precision	Recall	F1-Score
Artefact	0.76	0.40	0.53
Time Period	0.65	0.58	0.61
Material	0.72	0.46	0.56
Average	0.71	0.48	0.56

Table 5.2: Precision, recall and F1-scores for the 3 targeted entities, on a scale of 0 to 1.

When assessing the results of the NER, it was discovered that there are some issues with the gold standard documents which could affect the accuracy. It seems that some tagging decisions were made that mean that entities are expanded to the left or right. For example, wherever the word "before" or "after" occurs before a time period, these words are included in the tag, while ideally these shouldn't be included as they aren't part of the time period itself. If the NER then fails to classify these prefixes as the entity, the recall will be lower than the precision, which can also be seen in our results.

The artefact, time period and material wordlists that were taken from the *Archeologisch Basis Register* (ABR), a thesaurus for Dutch archaeology maintained by the RCE. It contains phrases that are written in such a way that they do not match the way we would find these phrases in natural language. For example, the entry for "*doorboorde bijl*" (perforated axe) is "*bijl, doorboord*" in the thesaurus, making it difficult to match the two. These two issues will be further discussed in section 5.5.

The code described in this section is available at https://doi.org/10.5281/zenodo.1238861.

5.3.2 Indexing & front end

For this version, 100 randomly selected reports from the DANS repository were selected to be indexed. For each page in these documents, the trained CRF model is used to extract the named entities. These are combined with the full text of the page and converted into a JSON structure, which can then be indexed directly by ElasticSearch (Gormley & Tong, 2015), an open source search engine running on a web server. ElasticSearch uses JSON over Hypertext Transfer Protocol (HTTP) to index and retrieve information, making it very easy to integrate with other systems. The other advantage of using ElasticSearch is that it includes a number of features by default that are very useful for these kinds of search systems, including a result ranking system.

To query the index, a front end has been developed. As a framework for the web application, the free and open source content management system Concrete5 was used (Concrete5, 2018).

To create a query, the user can use a query builder (Sorel, 2018) that allows for boolean AND / OR logic. You can specify exactly which entity you are looking for in each part of the query, or select a general full-text search. This allows for complex queries such as

```
artefact:scraper AND (period:neolithic
OR period:mesolithic) AND fulltext:burnt
```

which returns results on burnt scrapers from the neo- or mesolithic.

This query is then converted to a JSON format, so the ElasticSearch index can be queried using the ElasticSearch-PHP client (Tong, 2018), resulting in a list of matching results. It is useful to rank and sort these results by relevance, so the documents that are most likely to be relevant to a query are at the top of the list. To do this, ElasticSearch calculates a score for each result, which is based on the 'weight' of each query term that appears in that document. This weight is determined by three factors: term frequency, inverse document frequency and field length norm (ElasticSearch, 2018).

Once the results are displayed, the user can view a snippet of the text surrounding the keywords, preview the page of the report or go directly to the DANS repository to download the document. No PDFs are made available on the AGNES server to deal with the copyright of these files. A graphical representation of the full workflow of AGNES can be found in figure 5.2, which also displays the split between pre-processing of the documents on a high-performance cluster, and the indexing and querying that takes place on a standard web server.

5.4 User study

Part of this research includes a user study, to ensure the needs of the potential users are met. The focus group, as well as the methods and results of the first workshop, are detailed below.

5.4.1 Definition of target audience

To be able to make an effective search system, it is required to define the expected users of the system. As the main goal of this system is to make information available for research, the main expected user is a researcher working in Dutch

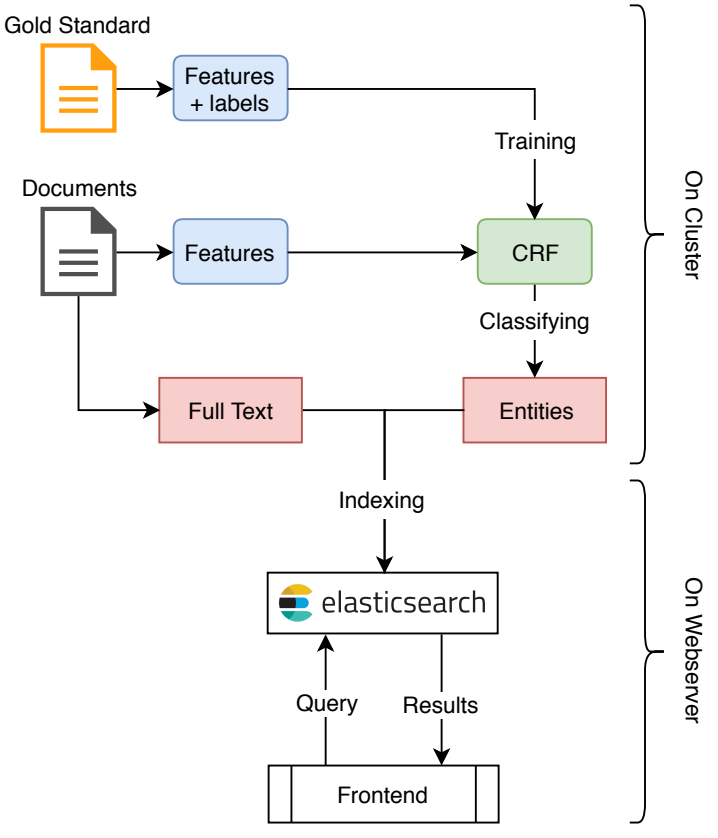

Figure 5.2: AGNES Workflow

archaeology. In the Netherlands, these researchers can be in a variety of organisational levels, including academia, commercial archaeology, regional/national government.

One of the main user groups expected to use this system are academics and people in higher education. However, this group is not homogeneous, as e.g. a professor will have much more in-depth knowledge and will already be aware of most of the literature and field reports related to their field, in stark comparison to e.g. a bachelor student or PhD who will still be exploring the literature and information available. Because of this difference in knowledge, these users will ask different questions of the data set and in different ways. However, regardless of their knowledge level it is expected that academic researchers will generally be asking thematic questions of the data set; questions about a particular time period, artifact type, context and/or location.

Another main user group is researchers in Dutch commercial archaeology. While this group will also be interested in the documents, it is likely that they will mainly want to use the system to find all information about a particular geographic area. This is because the main use of these reports for commercial archaeologists is to create desk assessments (*bureau onderzoeken*) and archaeological assessment/expectation maps (*archeologische verwachtingskaart*) about a specific area, generally because the area surrounds a potential building site. As some maps are also created by period, combined queries of place and time are also expected. There are three types of commercial archaeology, each are expected to have slightly different needs and requirements. These three types are inventarisation (investigating existing research), exploration or prospection (e.g. surveys and coring) and excavation (generally after the previous two types have been completed).

A third expected user group is municipal and regional (or provincial) archaeologists. Regarding their requirements, these will most probably fall in between academic and commercial archaeologists. While generally they will research a certain timespan in a particular area, it is likely that they will also want to research broader themes. However, generally they will be aware of all the available literature already, so perhaps a search system is less useful for this group.

Researchers at the RCE are a fourth user group, and will probably have similar needs to municipal archaeologists, except they are working on a country wide geographical scale. These researchers will commonly work on nation-wide synthesising research, combining the information from a large number of reports into a larger picture.

Outside of the archaeological sphere, it is possible that the system will also be used by historians researching more recent periods such as the Middle Ages, where there's an overlap between archaeology and history. It is expected these scholars will have similar requirements to archaeological academics.

Lastly, it is possible that this system might be used by amateur archaeologists, amateur historians, metal detectorists and other enthusiasts, for a variety of reasons.

5.4.2 Focus Group

In order to collect the requirements of archaeologists in the Netherlands, a voluntary focus group was set up. This group's function at the start of the project is to provide their needs and wishes for a system like this, while in further stages of the project they can provide feedback on the developed features. The size and make up of this group is fluid, and can be changed during the project to fit with

Group	Situation	Count
Academia	PhD Student	3
Academia	Assistant Professor	1
Academia	Lecturer	1
Commercial Archaeology	Excavation	1
Commercial Archaeology	Prospection	1
Government	Municipal	1
Government	National	1

Table 5.3: Overview of participants in focus group per category

the current goals and/or address issues of representativeness.

This group has been selected to be as representative as possible for the Dutch archaeological landscape, taking into account the target audience definition from section 5.4.1. The group consists of 5 academics, 2 commercial professionals and 2 archaeologists working on different levels in government. See table 5.3 for a more detailed break down of the participants.

No amateur researchers were selected for the focus group, mainly because they are not an intended user of the system, but also because their approaches to research are so wide ranging, it would be virtually impossible to assemble a representative group of people.

5.4.3 Prototype for discussion

From personal experience in commercial software development, as well as experiences from IR researchers in other fields (e.g. Verberne *et al.*, 2016), it seems that users in general, but the humanities specifically, find it difficult to express their requirements, oftentimes resulting in broad requirements that are too vague to interpret and implement. This can be further compounded by a lack of understanding of what is technically possible, leading to overly optimistic or very cautious expectations. We therefore first created a prototype with limited functionality (as discussed in section 5.3) as a starting point for discussions, in order to elicit feedback that is more detailed and can be implemented properly.

5.4.4 Workshops

The focus group will gather once a year during the project, for a total of 4 workshops. The initial workshop has been conducted, with the main aim of soliciting the requirements of the users. Later workshops will focus more on

assessing the system and its results. Minutes will be taken at each session to record the comments and feedback of the group, and these will be made public after anonymisation.

The first workshop started with an introduction to the problem, as well as some background information on IR and NER (see also section 5.3.1). The group was then asked what their current search behaviour is, and what problems they encounter, before being shown a prototype of the system (v0.2) and asked to provide feedback on both the functionality and the relevance of the results.

Finally, specific user requirements were discussed. A suggested list of features was provided to the participants, who then discussed amongst themselves in groups of 2 which features they would find most useful, on a scale of 0 to 3 with 0 being not useful or relevant at all, and 3 being very useful and high priority. The participants were also asked to think of features not currently on the list.

5.4.5 Results

From comments of the group, it was clear that the grey literature problem is very familiar to everyone present. Feedback on their current search behaviour showed that most people use the DANS search functionality[4] and find it not sufficient for their search needs, with most people having to manually search through individual documents to find information. Some participants, instead of using DANS, usually ask experts in the field to provide them with references, and the Archis[5] system is used to a lesser degree, again mainly because the search functionality is not sufficient. Some people explained that they create their own literature lists with keywords to be able to find materials previously accessed.

Initial feedback on the prototype indicates that the users find the returned results relevant to their queries, however much improvement is needed on the front end, as further discussed in the next paragraph.

The results from the feature elicitation were interesting; unanimously, everyone agreed that indexing by chapter and section would be more useful than indexing by page or document, and that this should be high priority. Another high priority feature across the board was to implement searching by drawing a polygon on a map as well as plotting results on a map, an indication that archaeologists have a strong need for geographical search. Another interesting result is that in general, everyone preferred to get many results with some irrelevant documents, than to get a smaller set of documents that are all relevant, with

[4] Found at http://easy.dans.knaw.nl

[5] Archis is a national database of archaeological sites in the Netherlands, maintained by the RCE, in Dutch. Located at https://archis.cultureelerfgoed.nl

Feature	Average
Search on map - plot results on map	2.78
Search on map - draw polygon	2.56
High recall over high precision	2.56
Search on map - morphology / expectation overlay	2.44
Index by chapter / section	2.33
Facets - time / artefact / place	2.22
Facets - research type	2.11
Personalise - alert if new docs in saved search	2.11
Related documents - by area	1.89
Facets - timeline	1.78
Personalise - save search	1.78
Related documents - by time	1.78
Ordering - by relevance	1.78
Personalise - mark documents as 'seen'	1.78
Ordering - by distance	1.67
Related documents - by artefact	1.67
Related documents - general	1.56
Plot terms in document	1.56
Ordering - by date added	1.11

Table 5.4: Features and average scores (0-3) across focus group (n = 9), sorted by average score, descending.

the risk of missing some documents. This means that the recall of the system is more important than the precision, which needs to be taken into account in assessing the results of the NER as well as the overall system assessment. For a full overview of the averaged result for each feature, please see table 5.4. In this table, facets mean the option for users to refine results by selecting categories, as often found on online shopping websites.

5.5 Future Work

The work discussed in this paper is the result of the first year of a 4 year project. Each year, a new version will be developed, tested, and assessed by the focus group.

The first issue that needs to be resolved is the gold standard. It seems that

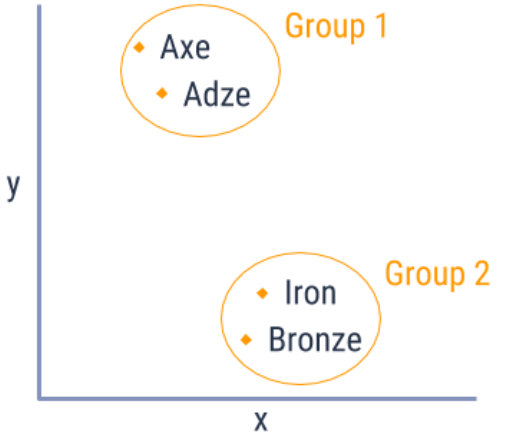

Figure 5.3: 2D representation of clustered word embeddings.

entities have been tagged sub-optimally for the NER task, and it is expected that improving the gold standard will increase the accuracy of the model. We intent to enlist the help of a group of archaeology students to re-tag these documents, and possibly tag new documents as well. We will have multiple people tag the same documents as well, so we can calculate the inter-rater agreement; a measure of how well this task can be accurately completed by humans, and ultimately, an upper limit for the accuracy of any machine learning model.

The other problem that will be addressed are the ABR wordlists. We are currently in discussion with the RCE, who manage these lists, to see if it's possible to add a new field for either the lemma of the word or to include multiple spellings of a word. After these two tasks have been completed, we will train the model again to see what difference these adjustments make.

Once that baseline has been established, we will integrate word embeddings as features, specifically word2vec (Mikolov *et al.*, 2013) and fasttext (Bojanowski *et al.*, 2016). These are both unsupervised machine learning techniques, that place words into a high-dimensional vector space based on their context in the text. The words can then be clustered using e.g. k-means clustering, with the idea that each cluster is a distinct 'type' of word. See figure 5.3 for a 2 dimensional (instead of high-dimensional) representation of this concept, where group 1 contains artefact types and group 2 contains materials. The advantage over using a word list is that related concepts not in the list, as well as misspellings of the concept, will also generally get assigned to the same cluster. Hopefully, this will increase the accuracy of the NER.

Regarding new features, according to the focus group the map functionality is the most required, including searching on a map and displaying results on a map. We are in the early stages of implementing this functionality and will hopefully present this in a future paper. Integration with common GIS systems is another avenue of research. Another feature with high priority is to index the documents by chapter or section, instead of by page as is currently the case.

To further evaluate the system, we will apply future versions to archaeological case studies. The plan is to find a specific archaeological information need, e.g. find all Iron Age cremations in the Netherlands and their geographical positions. We will then compare the results from AGNES with what experts currently know about this topic, and see if a significant increase in knowledge can be detected, probably by calculating the difference and overlap in numbers.

Currently, the system is focused on reports in Dutch, but as this problem is prevalent across the world, we will attempt to make the system multi-lingual, or at least provide ways of easily adapting the system to other languages.

Finally, one of the goals of the project is to expand the corpus from just the DANS documents to also including documents from the RCE and the KB, and creating a pipeline or API that allows for new documents added to these 3 repositories to be automatically added to the index.

5.6 Conclusions

From the user study, it is clear that a system such as AGNES is highly desirable for Dutch archaeology. The features assigned highest priority by the focus group are fairly uniform, which makes planning a road map of features straight forward. The first tentative feedback from the focus group is that results in AGNES are relevant to the queries, but more needs to be done to improve the functionality of the system.

From a technical viewpoint, the NER using CRF and a basic feature list resulted in an overall accuracy of 56%; fairly low, but partly explained by the problems with the gold standard and word lists. Fixing these problems, as well as introducing word embeddings as features, should increase the accuracy.

Overall, it seems that AGNES can address the problem of grey literature in Dutch archaeology, although this needs to be evaluated more thoroughly by comparing the results to expert knowledge. The systems developed should easily be adapted to other languages and areas as well. We are hopeful that AGNES will help archaeologists to answer their research questions more effectively and efficiently, leading to a more coherent narrative of the past.