SEARCHING IN ARCHAEOLOGICAL TEXTS

PROBLEMS AND SOLUTIONS USING AN ARTIFICIAL INTELLIGENCE APPROACH

Hans Paijmans & Alex Brandsen#*

*Tilburg University (Warandelaan 2, 5000 LE Tilburg, The Netherlands)

#Corresponding author: Leiden University, Faculty of Archaeology (Reuvensplaats 3, 2311 BE Leiden, The Netherlands) - Current address: University of York, Faculty of Archaeology (The King's Manor, York, Y01 7EP, UK) alex.brandsen@gmail.com

Paijmans, H. & A. Brandsen. 2010. Searching in Archaeological Texts. Problems and Solutions Using an Artificial Intelligence Approach. – Palarch's Journal of Archaeology of Egypt/Egyptology 7(2) (2010), 1-6. ISSN 1567-214X. 6 pages + 2 figures.

Keywords: archaeological texts, artificial intelligence, grey literature, metadata, name-dentity recognition

ABSTRACT

Searching in documents using full text indices is a powerful tool for retrieving relevant portions of text. However, performance is impeded by ambiguity in texts: similar words may have totally different meanings according to context. This also is true if the words are numbers, periods and place names, especially in archaeological and historical contexts. A new way of indexing texts allows for better and easier searching. This system has been developed in a collaboration between the RCE (The Dutch National Service for Cultural Heritage)[1] and the University of Tilburg. With Open Boek,[2] it is possible to search on chronological and geographical expressions, as well as regular keywords. In the newest version of Open Boek a number of additions to the system have been made to further improve the functionality.

Problems with Availability and Searchability

The importance of making information within texts available in an effective and durable way should be compared to the conservation of archaeological objects in museums; if the objects are not treated in a durable way, future study will be impossible. In the case of texts, we mean 'study' to be automated retrieval - the automated searching and finding of (archaeological) texts. A lot of information can be gained from past research, and texts containing such research should be treated with the same amount of attention as regular finds in museums, which at the moment does not happen very often. Because of this, large amounts of written information do exist, but are never used in research

© PalArch Foundation

because they cannot be found. "The information in all the published and unpublished, 'grey' reports are, in a way, just as hidden as the material that still resides underground. These facts are of course more or less read and known, but even the most erudite archaeologist or historian possesses but a small part of all that knowledge" (Paijmans & Lange, 2008: 14, free translation by authors).

The first problem with the storage of texts is availability. In the Netherlands this problem is addressed by DANS (Data Archiving and Networked Services)[3]: "Since its establishment in 2005, DANS has been storing and making research data in the arts and humanities and social sciences permanently accessible. To this end DANS develops permanent archiving services, stimulates others to follow suit, works closely with data managers to ensure as much data as possible is made freely available for use in scientific research."[4]

The other problem lies in finding the texts that are relevant to the user, based on the contents of those texts. The traditional way of solving this consists of manually adding metadata to the documents, such as author, subject, geographical and sometimes even temporal information. Unfortunately, this is a very time consuming and coarse-grained method and is dependant upon the knowledge of the subject at hand. An other possibility is to index keywords using a computer, which allows for searching in a Google-like way.

This technique is more fine-grained and less time consuming than adding metadata manually. The difference between indexes and metadata is the search level; metadata describe the documents while indexes describe the text in the documents. In other words, searching in an index goes one level 'deeper' than searching in metadata.

The problem with Google-like indexing however, is that the meaning of the words is not known to the system. On a search engine based on indexes, such as Google, a query on "middle ages and ceramics" will only show documents containing the words "middle", "ages" and "ceramics". It is possible to further define your query, for example by using quotation marks to search for exact word combinations or phrases. In this case this would result in finding any document containing the exact wording "middle ages and ceramics". However, the meaning of the words "middle ages" as a period between 500 and 1500 AD is lost. When making documents searchable, it is very important for the system to know that meaning; when searching documents about, for example, the Middle Ages the user also wants to find documents containing "late medieval period", "11th century" or "1100-1200 AD". Unfortunately, Google will not be able to do so.

Open Boek

In 2006, CATCH projects started all around the Netherlands. CATCH (Continuous Access To Cultural Heritage)[5] is an initiative by NWO,[6] The Netherlands Organisation for Scientific Research. Each project is a collaboration between a heritage organisation and a university. In the RICH project (Reading Images in Cultural Heritage)[7] the RCE (The Dutch National Service for Cultural Heritage)[8] collaborates with the University of Tilburg. In the context of this collaboration, a new way of indexing was developed and integrated into a usable system; Open Boek.

Technology

Chronological expressions (timespans) can be divided into two categories. The first are period names such as "middle ages" and "early neolithic". These expressions are easily recognised and can be correlated to a certain timespan in years. For Open Boek a table was created containing a list of the most common period names and their corresponding timespans. During the indexing process the periods in the text are labeled with the timespans indicated in this table. Problems arise when dealing with regional variations of timespans, such as the Neolithic, which differs from place to place, even in a relatively small area as the Netherlands. This issue will be discussed later in more detail.

The second kind of expression concerns the dates themselves, and the many variations that exist in writing these dates: '1100', 'eleven hundred', 'XI century', '11th century', '1100 ± 60 BP', etcetera. It is not hard to write a so called 'parser' (a program recognising words) that recognises and converts written cardinal and ordinal numbers to real numbers (integers). However, it becomes more difficult when trying to recognise all the variations such as '11th', '11-th', 'eleventh' or even in superscript: '11^{e}'.

It is even more difficult when trying to interpret numbers mentioned in a text. Is '1100' a year, a serial number, the number of artifacts or the price of a shovel? It turns out to be extremely difficult to construct a set of rules that can distinguish between timespans and other numbers. This is where artificial intelligence comes in.

MBL (Memory Based Learning), and more specific TiMBL (Daelemans *et al.,* 2004), is used in Open Boek to recognise timespans. MBL is a form of artificial intelligence that uses examples to learn that certain words in certain contexts have certain meanings. In the case of Open Boek the process is summarised as follows (also see figure 1).

- First, a number of example documents are processed by a language dependent parser which extracts all numbers and their context, and lists these in a table (see also figure 2).
- When this is completed, all the extracted numbers are evaluated and labeled by hand in the following categories: timespan, bibliographical or other. This results in an example database.
- New documents are processed by the same parser as in step one. The extracted numbers are compared to the example database by TiMBL and every extracted number is given the tag of the example it resembles the most.

Open Boek will tag the numbers in the text with the corresponding timespan after TiMBL has interpreted these numbers. An expression such as '100-200 AD' will be tagged '01/01/100 AD – 31/12/200 AD'. Period names and named centuries are also tagged in this phase. This completes the chronological index and it now becomes possible to search periods and years. Artificial intelligence, such as MBL, never reaches total accuracy and makes mistakes, just as human intelligence. Still, with this system an accuracy between 90% and 96% can, and has been reached.

The indexing of place names takes place in approximately the same manner, in which the context is used to define wether a word is a place name or not. The specific problem is the identification of which place the name denotes, because most place names are far from unique (Pouliquen *et al.* 2006). At the time of indexing, all place names are tagged with their corresponding longitude and latitude. After this step it is possible to resolve queries such as "middle ages, in a radius of 10 miles around Amsterdam".

Indexing versus Metadata

Open Boek allows for resolving queries relating to time, space and regular keywords (as explained above). This greatly improves the functionality, as opposed to the regular keyword indexes such as the one used by Google. However, this system also has its drawbacks.

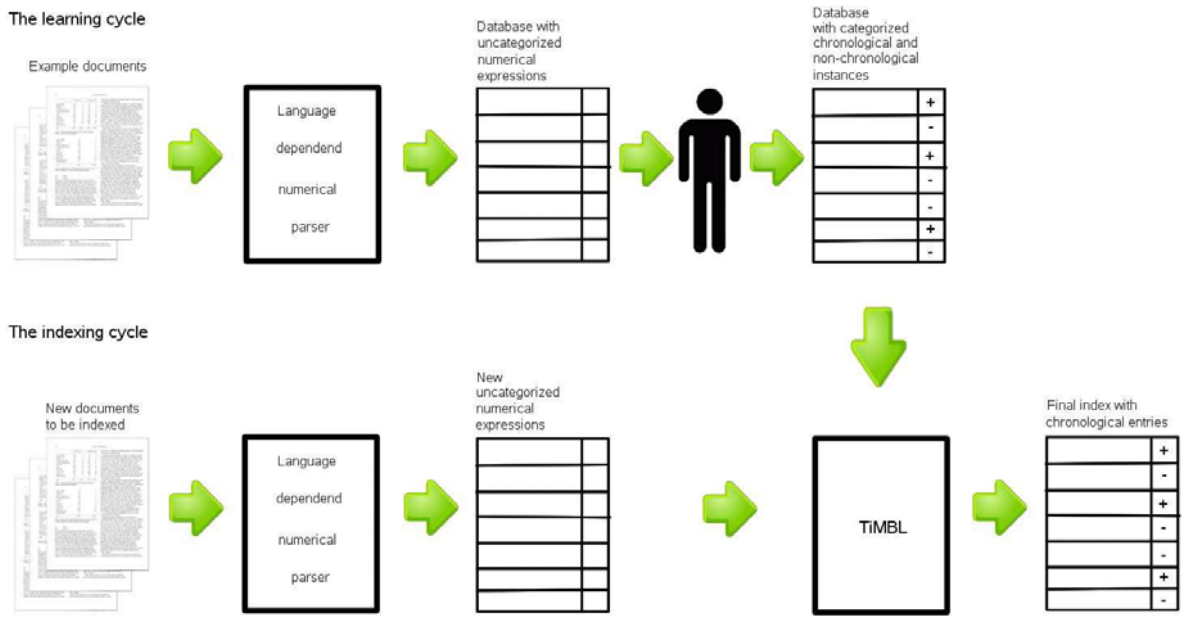

Figure 1. The MBL cycle, adapted from Paijmans & Lange (2008).

Leiden	2003	Postbus	9515	2300	RA	Leiden	info	[Other]
81,1	Bd	:	beschadigd	59,9	GL	:	beschadigd	[Other]
.	Bloo	)	Bijlage	3	Overzicht	van	het	[Reference]
vroegste	complex	(	vindplaats	21	)	valt	op	[E42_Object_identifier]
op	onge	-	veer	0,2	m	+NAP	,	[E54_Dimension:depth]
slechts	RAAP	-	rapport	969	/	eindversie	12	[E42_Object_identifier]
(	Module	3	)	3	Vondsten	Het	toekomstige	[Other]
De	Franse	kaart	van	1811	vormt	de	oudste	[E52_Timespan]
te	geven	.	Tabel	6	Aantals	-	en	[Reference]
cultuur	,	uit	de	vierde	eeuw	van	de	[E52_Timespan]
uit	de	Midden	-	Bronstijd	zijn	vijf	sites	[E52_Timespan]
Midden	-	Bronstijd	zijn	vijf	sites	aangetroffen	.	[E60_Number]
IJstijd	:	ca	.	8800	jaar	voor	Chr	[E52_Timespan]
aangetrof	-	fen	.	6	.	conservering	:	[Other]

Figure 2. Example of a table with parser-selected numbers and their context. The last column contains the tags given to these numbers. From Paijmans & Lange (2008).

There are a number of differences between adding metadata and automated indexing of texts. The first is the difference in search level; metadata are document-orientated while indexes are page- or even sentence-oriented. The greatest difference between these orientations is the grainedness, with metadata being very course-grained and indexes being fine-grained. Another difference is the time needed for processing the text. Adding metadata manually is a very time-consuming process; the texts need to be read and tagged by hand. On the other hand, automated indexing is done by the computer and takes a lot less time. The drawback is that the semantics of the keywords are less accurate than metadata.

When dealing with manually added metadata there is always a certain deal of subjectivity. Subjective metadating may be semantically more precise, but it depends on the skills and opinions of the reader. Objective, computerised indexing may miss out on meaning, but it is consistent and independent of human judgment. Another factor is of course the financial aspect; manually adding metadata is a lot more expensive than indexing by computer. Indexing on the other hand costs a fraction in terms of manpower and the software is (in the case of Open Boek) Open Source (GPL) and free to use. The reason for developing open source software is, in our case, that all research funded by NWO needs to be open source, unless good reasons specify otherwise. It seems that the only reason why manually added metadata is still used is that it is a very comprehensible and easy way to describe data.

Improvements

The newest version of Open Boek contains a number of improvements to enhance the functionality. The first of these is that it is now possible to index English and German texts, in addition to Dutch texts. Because these days almost all scientific articles are written in English, there was special need for an English version of the indexing system.

To adapt the chronological system to a new language, three rather simple steps are needed:

- a new parser needs to be written for every language, that converts written cardinal and ordinal numbers to integers (real numbers).
- a sample database of approx. 8000 cases needs to be created and tagged by hand (see section *Technology* and figure 1).
- Another language-dependent adjustment is the addition of geographical-specific chronologies to compare the period names in the texts with.

Currently the accuracy of Open Boek for English and German texts is not as high as with Dutch texts, but work to improve this is being undertaken.

The second improvement is the use of an algorithm to recognise errors in words caused by badly scanned and OCR'd[9] texts. Open Boek

searches for period names in the text and even an error of one single letter in a word would cause the system to not correctly recognise it. To solve this we used a measure known as the Levenshtein Distance, which counts the number of single-letter mutations needed to convert a word to an other word (Levenshtein, 1966).

The value of the Levenshtein Distance between the period name found in our chronology and the words in the text is calculated; if this is two or less, the word in the text will be interpreted as the period name it has been compared with. An example: when OCR'ing English texts a common error is that the word "middle" is read as "middie". The Levenshtein Distance between these two words is 1 (change the second "i" to an "l"). When the words "Middie Pleistocene" are found in the text then these will be recognised as the "Middle Pleistocene", and tagged accordingly.

Another adjustment is the implementation of automated recognition of bibliographic dates and tagging these as non-relevant dates. To accomplish this an extra classification in the example databases needed to be added. Not only can a number be tagged as "timespan" or "other" it can now also be tagged as "bibliographic". This means that publication dates will not be taken into account when doing a chronological search.

The last improvement that has been made is the possibility to link directly to external websites from the query results. This opens possibilities to implement Open Boek in other data archiving systems such as DANS (see section 1), in which Open Boek could be used solely as a query resolver, displaying results with the corresponding links to the documents on an external server (in this case the DANS-server). The data archiving system can then apply certain user rights and protect the copyright of the documents as agreed with the copyright holder. All this allows for the use of Open Boek on copyrighted material, without violating the copyright by making the documents openly available on the Open Boek server.

Problems and Future Solutions

Of course, the system is not yet perfect and some unsolved problems still occur. One of these problems is that the system does not know what a document relates to in a geographical sense, and this information is especially important to decide which chronology should be used. A problem related to this is the ambiguous character of place names; there are 11 places in the United States called "Amsterdam", there is a place called "Tabel" (Dutch for table) in Papua New Guinea and a place called "Paal" (Dutch for stake or post) in 5 different countries.

Because of these anomalies it is not possible to just calculate the mean of all found coordinates and use that as a basis for determining the geographical location. In fact, when this is done on Dutch articles, the mean usually lies somewhere in the South Atlantic Ocean ('Dutch' names in the USA pulling the mean to the west, and Dutch names in South Africa pulling them to the south). In search of a solution to this we have developed a program to automatically calculate a geographical mean of the article. It does this by (1) taking all the place names found in a text, (2) calculating from which country the most place names derive, (3) taking all place names found in the text from that specific country and finally (4) calculating the mean coordinates from these place names. In the latest tests using this method the mean usually lies around 20 kilometers off the actual geographical subject. Such deviations are explained by the occurrence of places like Amsterdam and Utrecht that occur in almost every document affecting the mean. This function is not yet implemented and fully functional, but research on this issue is being undertaken.

A final, and very important problem that can lead to a decrease in accuracy are regional variations in chronologies: the timespan of a period can differ from place to place in a single country. For example, the Dutch neolithic starts earlier on the fertile loessial soil to the south than on the less fertile clay in the north. A solution to this problem may be found in the solution to the geographical problem stated above. Open Boek would firstly calculate where the document is about geographically, and then use the corresponding local chronology. That still leaves us with the problem of how those local chronologies are defined chronologically and geographically. We do not propose any solution to this problem yet and Open Boek will only use chronologies defined per country.

© PalArch Foundation

Conclusions

Even though Open Boek is developed for archaeological texts, its use also extends to other fields of research such as history and art history. The system can help to make large amounts of text searchable in an effortless, cheap and relatively easy way. Indexing using MBL (or other forms of artificial intelligence) will never be a 100% accurate but it is a method that can index on page-level (as opposed to metadata) and can even recognise the meaning of words (as opposed to regular indexes). These advantages make Open Boek the preferable index-and search engine for (large) collections of scientific texts. With the latest improvements as described in section 4, even more functionality has been added to the system and further developments will increase the usefulness of this system. The importance of Open Boek for making archaeological texts available to search could even be compared to the importance of the shovel for making finds available for research.

Acknowledgements

This work was supported by NWO (Nederlandse organisatie voor Wetenschappelijk Onderzoek) and CATCH (Continuous Access To Cultural Heritage) under grant 640.002.401. No Microsoft software was used for the experiments mentioned in the paper or for the preparation of the paper itself.

Endnotes

1 www.cultureelerfgoed.nl
2 www.referentiecollectie.nl/Openboek
3 www.dans.knaw.nl
4 www.dans.knaw.nl/nl/over_dans/, accessed 22-12-2009
5 www.nwo.nl/catch
6 www.nwo.nl
7 www.rich.unimaas.nl/
8 www.cultureelerfgoed.nl
9 OCR stands for Optical Character Recognition: reading scanned texts by computer and converting them to textfiles.

Cited Literature

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch. 2004. Timbl: Tilburg Memory Based Learner, Version 5.1, Reference Guide. Ilk Technical Report 04-02. Technical Report, Tilburg University.

Paijmans, H. & G. Lange. 2008. Het Zoeken en Interpreteren van Jaartallen en Plaatsnamen in Nederlandstalige Retrieval Systemen. - Vitruvius: 14–17.

Pouliquen, B., M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A. Forslund & C. Best . 2006. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. - Arxiv Preprint cs/0609065.

Levenshtein, V.I. 1966. Binary Codes Capable of correcting Deletions, Insertions, and Reversals. - Soviet Physics Doklady 10: 707–710.

Submitted: 21 January 2009
Published: 6 March 2010

Copyright © 2003-2010 PalArch Foundation

The author retains the copyright, but agrees that the PalArch Foundation has the exclusive right to publish the work in electronic or other formats. The author also agrees that the Foundation has the right to distribute copies (electronic and/or hard copies), to include the work in archives and compile volumes. The Foundation will use the original work as first published at www.PalArch.nl.

The author is responsible for obtaining the permission of the use of illustrations (drawings, photographs or other visual images) made by others than the author. The author can be requested to submit proof of this permission to the PalArch Foundation. Pdf texts (papers and proceedings) are free to download on the conditions that each copy is complete and contains the PalArch copyright statement; no changes are made to the contents and no charge is made. The downloaded (and/or printed) versions of PalArch publications may not be duplicated in hard copy or machine readable form or reproduced photographically, and they may not be redistributed, transmitted, translated or stored on microfilm, nor in electronic databases other than for single use by the person that obtained the file. Commercial use or redistribution can only be realised after consultation with and with written permission of the PalArch Foundation.