A Cross-Language Approach to Historic Document Retrieval

Marijn Koolen [ab] Frans Adriaans [c] Jaap Kamps [ab] Maarten de Rijke [a]

[a] *ISLA, University of Amsterdam*
[b] *Archives and Information Studies, University of Amsterdam*
[c] *Utrecht Institute of Linguistics OTS, Utrecht University*

Abstract

The full paper appeared as: M. Koolen, F. Adriaans, J. Kamps and M. de Rijke. A Cross-Language Approach to Historic Document Retrieval. In: M. Lalmas et al. (eds.), *Advances in Information Retrieval: 28th European Conference on Information Retrieval (ECIR 2006)*, LNCS 3936, pp. 407–419, Springer Verlag, 2006.

Introduction Our cultural heritage, as preserved in libraries, archives and museums, is made up of documents written many centuries ago. Large-scale digitization initiatives, like DigiCULT [2], make these documents available to non-expert users through digital libraries and vertical search engines. For a user, querying a historic document collection may be a disappointing experience. Natural languages evolve over time, changing in pronunciation and spelling, and new words are introduced continuously, while older words may disappear out of everyday use. For these reasons, queries involving modern words may not be very effective for retrieving documents that contain many historic terms. Although reading a 300-year-old document might not be problematic because the words are still recognizable, the changes in vocabulary and spelling can make it difficult to use a search engine to find relevant documents. To illustrate this, consider the following example from our collection of 17th century Dutch law texts. Looking for information on the tasks of a lawyer (modern Dutch: *advocaat*) in these texts, the modern spelling will not lead you to documents containing the 17th century Dutch spelling variant *advocaet*. Since spelling rules were not introduced until the 19th century, 17th century Dutch spelling is inconsistent. Being based mainly on pronunciation, words were often spelled in several different variants, which poses a problem for standard retrieval engines. We therefore define Historic Document Retrieval (HDR) as the retrieval of relevant historic documents for a modern query. Our approach to this problem is to treat the historic and modern languages as different languages, and use cross-language information retrieval (CLIR) techniques to translate one language into the other.

Earlier research has seen similar approaches. In 1992, Robertson & Willett [4] used spelling correction techniques and phonetic substitutions for retrieving 17th century English spelling variants of modern words. The phonetic substitutions were constructed manually. They found that phonetic substitutions (i.e., replacing a typically historic sequence of characters by a modern sequence with the same pronunciation) have very little effect, while spelling correction techniques increased performance in finding spelling variants. Braun [1] used (again, manually constructed) rewrite rules, similar to phonetic substitutions, to rewrite historic character sequences to modern character sequences. In this case, they turned out to be very effective in an IR experiment, making a modern Dutch stemmer [3] more effective after rewriting the historic documents. However, the manual construction of rewrite rules is very time-consuming, and each set of rules that is created, is language dependent. Rules created for 17th century Dutch will probably not work for 17th century English, nor for 14th century Dutch. Constructing rule sets automatically would save a lot of time and effort. But is it possible to construct such translation resources automatically? Furthermore, is this cross-language approach (translating historic language into modern language or vice versa) a suitable approach to HDR?

Table 1: Evaluating translation and stemming effectiveness, using the title field of the topic statement (top half) or its description field (bottom). Best scores are in boldface, significance $^\star = p < .05$, $^{\star\star} = p < .01$.

Method	**MRR**	% Change
Baseline (titles)	0.1316	–
RNF-all + RSF + PSS	**0.2780****	+111.2
RNF-all + RSF + PSS + Stemming	0.2766**	+110.2
Baseline (descriptions)	0.1840	–
RNF-all + RSF + PSS	0.2842*	+54.5
RNF-all + RSF + PSS + Stemming	**0.3410****	+85.3

Results We have developed algorithms to construct rewrite rules which replace historical character combinations with modern variants. Using N-gram frequencies, typical historic character combinations are selected for rewriting (like *ae* in *advocaet*). A character combination is typically historic if it is much more frequent in a historic corpus than in a modern corpus. By replacing these N-grams in the historic words with a wildcard *, the resulting word (*advoc*t*) is matched with modern words. The character combination in a modern word (*advocaat*) that matches the wildcard is a possible substitute for the historical character combination (resulting in $ae \rightarrow aa$). Since the rules are based on the relative frequency of n-grams, we call this algorithm the Relative N-gram Frequency (RNF) algorithm. (A variant of this is the RSF algorithm, based on the relative frequency of vowel sequences and consonant sequences.) Another approach to construct rules is to use phonetic transcriptions of historic and modern words, and match those words that are pronounced the same. By aligning sequences of vowels, and sequences of consonants, differences in spelling can be transformed into rewrite rules. (*ae* is aligned with *aa*, resulting in $ae \rightarrow aa$.) This method is the Phonetic Sequence Similarity (PSS) algorithm. Multiple modern character combinations can be found for a historic character combination. The modern combination that is found most often, is used. The rules are applied on the historic collection to construct a translation resource, which is used in the retrieval experiments.

Our experimental evidence is based on a collection of 17th century Dutch documents and a set of 25 known-item topics in modern Dutch. We have experimented with both query translation (adding historic variants to modern words) and document translation (modernizing the historic documents), and found that document translation outperforms query translation. Table 1 shows the retrieval results for our most successful run, which uses a combination of the RNF, RSF, and PSS rules. We see that, after rewriting, we can use a modern stemming algorithm. This improves the performance on the description fields.

Conclusions Our main findings are as follows: First, we are able to automatically construct rules for modernizing the historic language using algorithms that compare historic and modern words on the phonetic and orthographic level, and use statistics to bridge the gap. Second, modern queries are not very effective for retrieving historic documents, but the historic language tools lead to a substantial improvement in retrieval effectiveness. The improvements are above and beyond the improvement due to using a modern stemming algorithm (whose effectiveness actually goes up when the historic language is modernized). However, our approach only addresses the spelling gap. The problems caused by changes in vocabulary are unresolved. We're working on this specific problem as well. Currently, we are investigating the possibilities of mining annotations from 17th century literary texts to construct a translation dictionary.

Acknowledgments Thanks to Margariet Moelands (National Library of the Netherlands) for drawing our attention to historic document retrieval. Thanks to Loes Braun for making available the *Antwerpse Compilatae* and *Gelders staden landrecht* corpus. This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 016.054.616, 639.072.601, 017.001.190, 220-80-001, 264-70-050, 365-20-005, 612.000.106, 612.000.207, 612.013.-001, 612.066.302, 612.066.513, 612.069.006, 640.001.501, and 640.002.501), and the E.U.'s 6th FP for RTD - project MultiMATCH contract IST-033104.

References

[1] L. Braun. Information retrieval from Dutch historical corpora. Master's thesis, Maastricht University, 2002.

[2] DigiCULT. Technology challenges for digital culture, 2005. `http://www.digicult.info/`.

[3] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[4] A.M. Robertson and P. Willett. Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods. In *Proceedings ACM SIGIR '92*, pages 256–265, New York, NY, USA, 1992. ACM Press.