

CATCH project SCRATCH - SCRipt Analysis Tools for the Cultural Heritage:

Text line matching for historical handwritten document retrieval

S. Zinger, J. Nerbonne

Center for Language and Cognition
Groningen (CLCG), Groningen University

{s.zinger, j.nerbonne}@rug.nl

L.R.B. Schomaker

Artificial Intelligence Department,
Groningen University

schomaker@ai.rug.nl

Henny van Schie

Nationaal Archief, the Hague

henny.van.schie@nationalearchief.nl

Introduction

In the historical handwritten document retrieval system that we are currently building, the training data set elements are the images of handwritten lines with the manually made text transcriptions. We apply sequence comparison algorithms to these text transcriptions.

Finding an appropriate method for comparing text lines will allow us to cluster the corresponding images of handwritten lines into training sets. These training sets can then be used for pattern recognition.

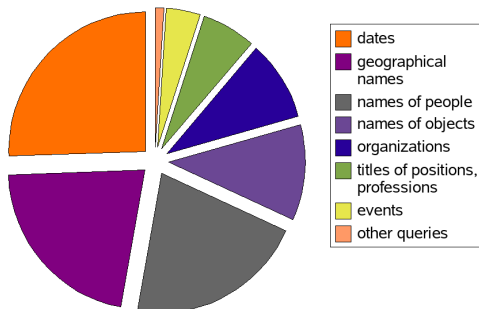
We work with the Kabinet van de Koningin collection of the Nationaal Archief.

Queries to the Nationaal Archief

The goal is to know what kind of information people would like to extract.

Available data: emails to the Nationaal Archief received since the year 2001 till the present time – several thousands of emails, mostly in English and Dutch.

Classification of the 6082 queries in Dutch received by the Nationaal Archief during the last 4 years (2001-2004):



Analysis of the queries shows that most of the queries concern named entities: organizations, names of people, geographical names.

Text line matching

Available data: 5445 handwritten lines segmented from scanned pages and manually transcribed.

Example of a handwritten line and its text annotation:

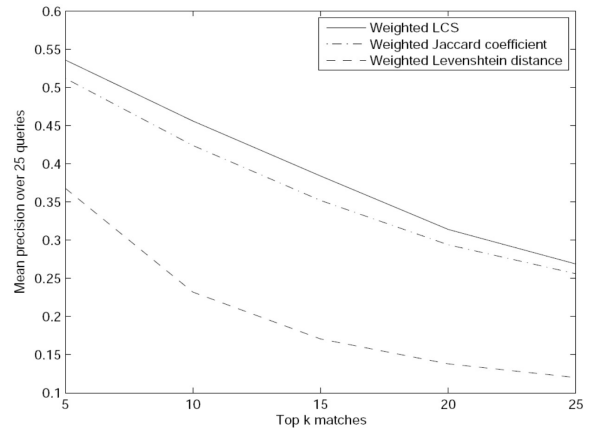
Cursus bij het Kon. Instituut der Marine
Cursus bij het Kon. Instituut der Marine

To match line, we apply the following sequence comparison algorithms: Levenshtein distance, Longest Common Substring (LCS), Jaccard coefficient.

Jaccard coefficient is defined as number of words that are common for the match and the query, divided by the sum of number of unique words for the query, number of common words for the match and query and number of unique words for the match.

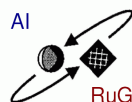
Content-based weighting helps to underline the importance concepts present in lines: low frequency words that often correspond to the classes of user queries [1].

Results on content-based weighted line matching:



[1] S. Zinger, J. Nerbonne, L. Schomaker, H. van Schie, "Content-based text line comparison for historical document retrieval", Computational Phonology workshop, Recent Advances in Natural Language Processing conference, RANLP-2007, pp. 79-84, Borovets (Bulgaria), September 2007.

Web-site of the project: <http://www.ai.rug.nl/alice/nwo-catch-scratch/>



Netherlands Organisation for Scientific Research

NWO project number: 640.002.402