CATCH project SCRATCH - SCRipt Analysis Tools for the Cultural Heritage: statistics on queries and line matching

S. Zinger, J. Nerbonne
Center for Language and Cognition Groningen (CLCG), Groningen University
{s.zinger, j.nerbonne}@rug.nl

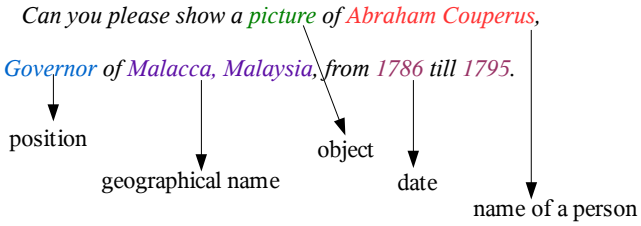
L.R.B. Schomaker
Artificial Intelligence Department, Groningen University
schomaker@ai.rug.nl

Henny van Schie
Nationaal Archief, the Hague
henny.van.schie@nationaalarchief.nl

Introduction

SCRATCH project aims at information retrieval from handwritten documents.
We work with the Kabinet van de Koningin collection of the Nationaal Archief. To determine what kind of information has to be retrieved from historical documents, we calculate statistics on the queries that the Nationaal Archief receives.
Sequence comparison applied to annotations of handwritten lines provides the reference for matching the handwritten lines and can also be useful for user's queries processing in an information retrieval system.

Example of a query to 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 966 queries in English received by the Nationaal Archief during the last 5 years (2001-2006):

Class of a query	Amount of queries	Part (%) from the total number of queries
names of people	268	28%
dates	241	25%
geographical names	212	22%
names of objects (*birth certificate, ship, map, military unit, company*)	152	16%
titles of positions, professions	73	8%
event(s) (*visit, ship crash, arrest*)	20	2%

Analysis of the queries in Dutch to the Nationaal Archief will provide better statistics on the information retrieval needs.

Text line matching

Available data: 3178 handwritten lines segmented from scanned pages and manually annotated.

Example of a handwritten line and its text annotation:

Cursus bij het Kon. Instituut der Marine

Jaccard distance 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.
Average Jaccard distance for the best matches for all lines is 0.49.
Modified Jaccard distance, that does not take into account the unique words of the matching line is 0.64.

Results on line matching:

	Average Jaccard distance		Modified average Jaccard distance	
Similarity measure	on letters	on words	on letters	on words
Levenshtein	0.41	0.44	0.54	0.6
longest common substring	0.40	0.42	0.54	0.6

The best similarity measure for lines is to be determined.

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

NWO
Netherlands Organisation for Scientific Research
NWO project number: 640.002.402