A Cross-Language Approach to Historic Document Retrieval

Marijn Koolen, Frans Adriaans, Jaap Kamps, Maarten de Rijke

University of Amsterdam and Utrecht (2006)

http://staff.science.uva.nl/~kamps/publications/2006/kool:cros06.pdf

Context:

Seminar "Text Mining for Historical Documents" (WS 2009/10)

http://www.coli.uni-saarland.de/courses/tm-hist10/

Presenter: Johannes Braunias, 22 February 2010

Non-standard Orthography

- Many historical texts are available, but not *accessible*:
- Historic language differs from modern language
  - in spelling:

darme man	die arme man	→ de arme man
tien tiden	te dien tiden	→ op die tijd
har**e**n**t**are	hare ende dare	→ her en der
hi cuss**e**se	hi cussede se	→ hij kuste ze
gae**d**i	gaet ghi	→ gaat u
kind**i**ne	kinde hi hem	→ kende hij hem

These examples involve clitics (agglutinated and phonetically dependant pre- or suffixes [= affixes] in the first column) http://en.wikipedia.org/wiki/Proclitic

  - and meaning

Credits to http://s2.ned.univie.ac.at/Publicaties/taalgeschiedenis/nl/mnlortho.htm

Non-standard Orthography

- → Disappointing results
  with modern-language queries
  because of shift in spelling and meaning:
  Search terms don't match historical terms.
- This paper deals with Dutch

Non-standard Orthography

- Goal:
  Make texts accessible to speakers
  of modern language
- Challenge:
  Bridge the gap between historical
  and modern language
- Historic Document Retrieval (HDR):
  The retrieval of relevant historic documents given a modern query.

Approaches to HDR

- Use spelling correction
- Rewrite rules (our approach)
- → Treat historic language as a separate language
- 1. Automatically construct translation resources (rewrite rules)
- 2. Evaluate these rules experimentally: Retrieve documents using CLIR techniques (Cross-language Information Retrieval) and stemming

Material we use for evaluation

… of the effeciency of rules:

393 documents (in 17th century historic Dutch)
25 topics (in modern Dutch)

Used format: TREC

- TREC = Text Retrieval Conference and format used by the the conference for experimental data
- Combines many documents into one file, separated by `<doc><docno></docno></doc>` tags

More on TREC

- Example TREC document file (containing 8 documents):

<DOC> And the sons of Noah, that went forth of the ark, were Shem, and Ham, and Japheth: and Ham is the father of Canaan. </DOC>
<DOC> genesis </DOC>
<DOC> These are the three sons of Noah: and of them was the whole earth overspread.</DOC>
<DOC> genesis </DOC>
<DOC> And Noah began to be an husbandman, and he planted a vineyard:</DOC>
<DOC> genesis </DOC>
<DOC> And he drank of the wine, and was drunken; and he was uncovered within his tent.</DOC>
<DOC> genesis </DOC>

- Example TREC title file:

<*TOP*>
<NUM>123<NUM>
<TITLE>title
<DESC>description
<NARR>narrative
</TOP>

Credits to http://www.seg.rmit.edu.au/zettair/doc/Build.html and http://terrier.org/docs/current/configure_retrieval.html

1. Construct translation resources

- Rewrite rules (algorithms), which map several spelling variants to one modern word
  - Phonetic similarity (PSS)
  - Orthographic similarity (RSF, RNF)

PSS RSF RNF
Phonetic Sequence Similarity

- Compares phonetic transcriptions (NeXTeNS):
  veeghen (historic) → *v e g @ n* *(phonetic transcr.)*
  vegen (modern) → *v e g @ n*
- Words are split into sequences of vowels and consonants and then compared:

historic:	v	ee	gh	e	n
modern:	v	e	g	e	n

Resulting rewrite rules:
ee → e
gh → g

- More matches/generations of a rule increase probability for correctness

PSS **RSF** RNF

Relative Sequence Frequency

- Split historic and modern words
  into vowel and consonant sequences
  v | o | lck *(count sequences in historic corpus)*
  Determine frequency of each sequence *(e.g. "lck")*
  in the corpus (separately for historic and modern)
  v | o | rk *(count sequences in modern corpus)*

- Calculate RSF:

$$RSF(S_i) = \frac{RF(S_i^{hist})}{RF(S_i^{mod})}$$

RSF(Si) > 1 means: Typical historic sequence

PSS **RSF** RNF

Relative Sequence Frequency

- v o **lck** *historic*
  v o **C** *historic wildcard word*
  v o l *words matched in the modern corpus*
  v o lk
  v o rk

- Created rules:
  lck → l 1
  lck → lk 1
  lck → rk 1

→ Each time a rule is generated by a wildcard word, its score is increased. Most probable rule has highest score.

PSS RSF **RNF**

Relative N-gram Frequency

- Split words into n-grams ("*n* letters in sequence")
  Example with n = 3:
  volck → #vo vol olc lck ck# *(# = word boundary)*
- Algorithm similar to RSF,
  with restriction of maximal edit distance 2
  to not overproduce matches
  (like vo**lck** → vo**orrijkosten**)

Select the best rules

- Select highest scoring rules ("pruning"):

$$S(R_i) = \sum_{j=0}^{N} \left(D(W_j^{hist}, W_j^{mod}) - D(W_j^{rewr}, W_j^{mod}) \right)$$

evaluated on 1600 word pairs.
the more positive, the more closer the spelling is.

- Compare PSS, RSF, and RNF:
  Feed the algorithms with historic words and compare them to modern equivalents (next page)
- … test rules on small test set
  of historic word and their modern counterparts

Results of evaluating the different sets of rewrite rules

Method	number of rules	total rewrites	perfect rewrites	new distance
none	–	–	–	2.38 –
PSS	104	253	101	1.66 (−0.72)
RSF	62	252	140	1.33 (−1.05)
RNF-2	12	271	152	1.29 (−1.09)
RNF-3	127	274	162	1.19 (−1.19)
RNF-4	276	269	166	1.20 (−1.18)
RNF-5	276	153	97	1.79 (−0.59)
RNF-all	691	315	207	0.97 (−1.41)
RNF-all + RSF + PSS	753	337	224	0.86 (−1.52)

- The best option: combine all 3 allgorithms
- Edit distance and perfect rewrites:
  Which measure performs better in retrieval?

2. Evaluation in Document Retrieval (HDR)

1. Do translation tools help?
2. *Document* translation or *query* translation?
3. Long or short topic statements?

- Measure: MRR, Mean Reciprocal Rank
- Parameters:
  - Monolinguality ("baseline")
  - Use short or long title
  - Using a stemmer or not

MRR – Mean Reciprocal Rank

Query	Results	Correct response	Rank	Reciprocal rank
cat	catten, cati, **cats**	cats	3	1/3
torus	torii, **tori**, toruses	tori	2	1/2
virus	**viruses**, virii, viri	viruses	1	1

Given those three samples, we could calculate the mean reciprocal rank as (1/3 + 1/2 + 1)/3 = 11/18 or about 0.61

http://en.wikipedia.org/wiki/Mean_reciprocal_rank

2. Evaluation in Document Retrieval (HDR)

- Evaluating translation effectiveness, using the title of the topic statement (top half) or its description field (bottom)

Method	MRR	% Change
Baseline (titles)	0.1316	–
Soundex7	0.2600*	+97.6
PSS	0.2397*	+82.1
RSF	0.1299	-1.3
RNF-all	0.2114*	+60.6
RNF-all + RSF + PSS	0.2780**	+111.2
Baseline (descriptions)	0.1840	–
Soundex7	0.1890	+2.7
PSS	0.2556	+38.9
RSF	0.1861	+1.1
RNF-all	0.2025	+10.1
RNF-all + RSF + PSS	0.2842*	+54.5

2. Evaluation in Document Retrieval (HDR)

- Does the stemming of modern translations further improve retrieval?
  Using the title of the topic statement (top half) or its description field (bottom)

Method	MRR	% Change
Baseline (titles)	0.1316	–
Stemming	0.1539	+16.9
RNF-all + RSF + PSS	0.2780**	+111.2
RNF-all + RSF + PSS + Stemming	0.2766**	+110.2
Baseline (descriptions)	0.1840	–
Stemming	0.1870	+1.6
RNF-all + RSF + PSS	0.2842*	+54.5
RNF-all + RSF + PSS + Stemming	0.3410**	+85.3

Conclusion

- Approach:
  Automatic construction of translation resources,
  Retrieval of historic documents with CLIR
- Findings:
  - Can build translation resources
    with help of PSS, RSF, RNF
  - Modern queries alone are not satisfying →
    document translation with algorithms,
    and with modern-language stemmer
    performs well

Further remarks: Bottlenecks

- Spelling bottleneck
- Vocabulary bottleneck
  - new words and disappearing words (over time)
  - shift of meaning
  - → vocabulary bottleneck is harder. Approaches:
    - indirect (query expansion)
    - direct (mining annotations to historic texts on the web)