Character Confusion based approach to the correction of the spelling variants in OCR-ed, historical corpora

Martin REYNAERT

Tilburg centre for Creative Computing

Political Mashup developers’ meeting, UvA, Amsterdam.
19 October, 2010
CORPORA AND NON-WORD ERRORS

REUTERS RCV1 CORPUS: exhaustively checked

33,333 erroneous word forms (collected manually)
≈ 21% of word types in error

TWENTE NEWS CORPUS 2002

16,500 erroneous word forms (identified so far)

STATEN-GENERALEAL DIGITAAL 1989-1995

99% OCR-accuracy – 5,047 typos and OCR-errors

DIGITALE DATABANK DAGBLADEN: Het Volk 1918

75% OCR-accuracy – 3,799 typos and OCR-errors
TEXT-INDUCED CORPUS CLEAN-UP: INTRODUCTION

TICCL for TYPOS and OCR-errors

- Tool to perform large scale, unsupervised spelling correction of corpora

- Spelling correction = reduction of lexical variation caused by typos, OCR-errors, historical orthographical changes...

- Linking of variants to their most likely canonical form. Enrichment of input corpus with variants.

- Production version developed according to KB specifications, second half 2008

- New, ‘release version’ described in: ‘Character confusion versus focus word based correction of spelling and OCR variants in corpora’ accepted for publication in: International Journal of Document Analysis and Recognition
Represent identical bags of characters (i.e. word strings sharing the same set of characters) by an identifying numerical value, use this value as the index key to the word strings in a database and perform simple calculations to retrieve variants from the database.
ANAGRAM HASHING

\[
Key(w) = \sum_{i=1}^{\lvert w \rvert} f(c_i)^n
\]

- A bad hashing function: produces collisions
- lines up ANAGRAMS: strings consisting of the same bag of characters
- In practice, we use the code value of each character in the string raised to the power 5.
- Values obtained for the string are summed
ANAGRAM HASHING II

- CAT = anagram of ACT and TAC
- $A + C + T = 65^5 + 67^5 + 84^5 = 6,692,535,156$
- $C + A + T = 67^5 + 65^5 + 84^5 = 6,692,535,156$
  - ALLOWS FOR ADDITION AND SUBTRACTION
  - SAME APPLIES FOR WORD COMBINATIONS, PHRASES, SENTENCES...
  - BASIS FOR TISC: Text-Induced Spelling Correction
Given ANAGRAM VALUE (AV): 6,692,535,156

- $\text{AV(\text{ACT})} + 84^5$ (plus T) = TACT
- $\text{AV(\text{ACT})} - 67^5$ (minus C) = AT, TA
- $\text{AV(\text{ACT})} - 84^5 + 82^5$ (minus T, plus R) = CAR
- $\text{AV(\text{ACT})} - 84^5 + 78^5 + 83^5$ (minus T, plus N, plus S) = CANS/SCAN

Search = non-local within string: efficient + effective

Linear cost: ‘alphabet’ and word length dependent

Sequential: take a word and systematically search for its variants, then take the next word..., etc.
Parallel identification of the spelling variants

- **TICCL = FOCUS WORD BASED APPROACH**: sequential + local variant look-up
  - Focus on a single word and retrieve all likely variants.
  - A bit like informed groping in the dark! You know where to look and when you find one that that one is there. You have no idea what else is there or how many like that are out there.

- **TICCL = CHARACTER CONFUSION BASED APPROACH**: global variant look-up
  - Focus on a particular character confusion, e.g. ‘IN’ misrecognized as ‘M’ and retrieve all word pairs displaying this confusion.
  - Repeat this for all the character confusions you are interested in.
  - Easily distributed over several processors/computers.
  - Boils down to a systematic gathering of facts and useful statistics about the facts.
OCR misrecognition: example from ‘Het Volk 1918’

OCR: vurkenskvieescjiachaarsclite

Historical spelling: varkensvleeschtschaarschte
Confusions: a - u, i - l, h - ji, s - a, h - li
Levenshtein distance = 8
Current spelling : varkensvleesschaarste
English: pork meat shortage
Table 5 SGD 1989-1995: Performance results on word types for the full system run in both FW and CC modes. N-best ranking scores are on ranks 1, 2, 3, 5 and 10-best. Note that the FW approach requires an order of magnitude more time to achieve similar results as the CC approach.

<table>
<thead>
<tr>
<th>rank</th>
<th>Focus Word</th>
<th>Character Confusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>1</td>
<td>0.923</td>
<td>0.955</td>
</tr>
<tr>
<td>2</td>
<td>0.963</td>
<td>0.948</td>
</tr>
<tr>
<td>3</td>
<td>0.963</td>
<td>0.946</td>
</tr>
<tr>
<td>5</td>
<td>0.963</td>
<td>0.946</td>
</tr>
<tr>
<td>10</td>
<td>0.963</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>Perl processing time</td>
<td>1328.9</td>
</tr>
</tbody>
</table>
### Table 7: \textit{hetvolk1918}: Performance results on n-best first ranking on types and tokens for the \textit{cc} approach

<table>
<thead>
<tr>
<th>rank</th>
<th>types</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>1</td>
<td>0.937</td>
<td>0.929</td>
<td>0.933</td>
<td>0.877</td>
<td>0.960</td>
</tr>
<tr>
<td>2</td>
<td>0.957</td>
<td>0.928</td>
<td>0.942</td>
<td>0.881</td>
<td>0.959</td>
</tr>
<tr>
<td>3</td>
<td>0.961</td>
<td>0.924</td>
<td>0.942</td>
<td>0.882</td>
<td>0.959</td>
</tr>
<tr>
<td>4</td>
<td>0.963</td>
<td>0.923</td>
<td>0.942</td>
<td>0.883</td>
<td>0.958</td>
</tr>
<tr>
<td>5/10</td>
<td>0.963</td>
<td>0.923</td>
<td>0.942</td>
<td>0.883</td>
<td>0.958</td>
</tr>
</tbody>
</table>
Future work in Political Mashup

- Go beyond single words, go ’context-sensitive’
- Deal with (non- and less-alphabetical) non-words
- Perhaps: look back at the image (Template Matching)
Thanks!!

Thanks for your attention!

Papers about TICCL are available at:
http://ilk.uvt.nl/

Character Confusion based approach to the correction of the spelling variants in OCR-ed, historical corpora

Martin REYNAERT

Tilburg centre for Creative Computing