Spelling Correction for Online Public Access Catalog Records in Library Integrated Catalogue (LINC)

Project for CS5244 (Digital Library)

By Chen Ding, He Cong, Zhan Jiaming, National University of Singapore

Nov 19th, 2004
Introduction

- **Goal**: To correct the misspellings in LINC records of NUS Digital Library

- **Motivation**: Misspellings hinder the information seeking process.

Case 1: misspellings from the query

**Solution**: 1. suggest misspellings for query 2. correct misspellings in the database

Case 2: misspellings from the database
Processes of experiment

1. Extract words from records
   - Title
   - Subject
   - Author
   - Publisher
   - Language
   - etc.

   \[\text{Title} \quad \text{Subject} \quad \text{(English records)}\]

   \[910,000 \text{ words}\]

2. Detect misspelling
   - Compare with Dictionary
   - Some proper nouns might be regarded as misspellings
3. Generating candidate corrections

2,651 Possible errors

Generating candidates by skeleton key

Generating candidates by omission key

Generating candidates by insertion key

Ranking candidates

 assesment → asmnte

 assesment → tmnsae

 assesment → nmtsa
4. Ranking candidate corrections

- **Error**: iniatiative
  
  Candidate Corrections: initiative, inattentive

- **Ranking**
  - *Minimum edit distance*
    
    e.g. \(d(\text{"iniatiative"}, \text{"initiative"}) = 1 < d(\text{"iniatiative"}, \text{"inattentive"}) = 4\)

  - *Keyboard distance*
    
    e.g. \(kd(\text{"w"}, \text{"q"}) = \| (1,1) - (1,0) \| = \sqrt{0^2 + 1^2} = 1\)

  - *Frequency of words:*

    *the more frequently a candidate occurs, the higher its rank is*
## Experiment Results

Performance of error correction (after ranking) by using skeleton, omission and insertion key

<table>
<thead>
<tr>
<th>Correction schemes</th>
<th>Could not be corrected</th>
<th>Be corrected</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Properly</td>
<td>Improperly</td>
</tr>
<tr>
<td><strong>Skeleton key</strong></td>
<td>956 (36%)</td>
<td>1444 (54.5%)</td>
<td>251 (9.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(48.75~71.73%)</td>
<td></td>
</tr>
<tr>
<td><strong>Omission key</strong></td>
<td>498 (18.8%)</td>
<td>1879 (70.9%)</td>
<td>274 (10.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(56.25~76.95%)</td>
<td></td>
</tr>
<tr>
<td><strong>Insertion key</strong></td>
<td>424 (16%)</td>
<td>1935 (73%)</td>
<td>292 (11%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
Frequency approach for proper nouns

- Misspellings of proper nouns which can not be corrected by similarity key
- Similar spellings: the one with higher frequency is more likely the correct form

<table>
<thead>
<tr>
<th>Key</th>
<th>Spelling1</th>
<th>Spelling2</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>mrtaio</td>
<td>Marriot (1)</td>
<td>Marriott (10)</td>
<td>Marriott</td>
</tr>
<tr>
<td>mlysai</td>
<td>Malaysia (2)</td>
<td>Malaysia (158)</td>
<td>Malaysia</td>
</tr>
<tr>
<td></td>
<td>Malaysia (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pnyldkeia</td>
<td>Penyelidekan (2)</td>
<td>Penyelidikan (19)</td>
<td>Penyelidikan</td>
</tr>
<tr>
<td>krshnmtiau</td>
<td>Krishnamurthi (2)</td>
<td>Krishnamurti (7)</td>
<td>Krishnamurti</td>
</tr>
</tbody>
</table>
Conclusion

- Good performance for the data set before considering the proper nouns
- Frequency comparison approach for the proper nouns

Future work

- Context dependent approach
  - Syntactical
- Mutual information
  - Term vs. subject