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

Project for CS5244 (Digital Library)

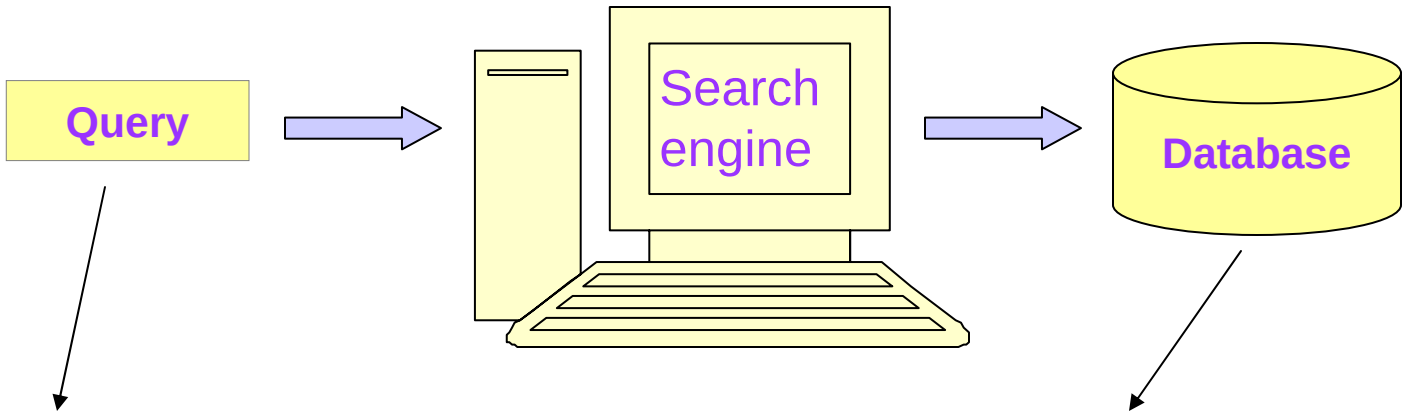
By Chen Ding, He Cong, Zhan Jiaming,
National University of Singapore

Nov 19th, 2004

Introduction

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

Misspellings hinder the information seeking process.



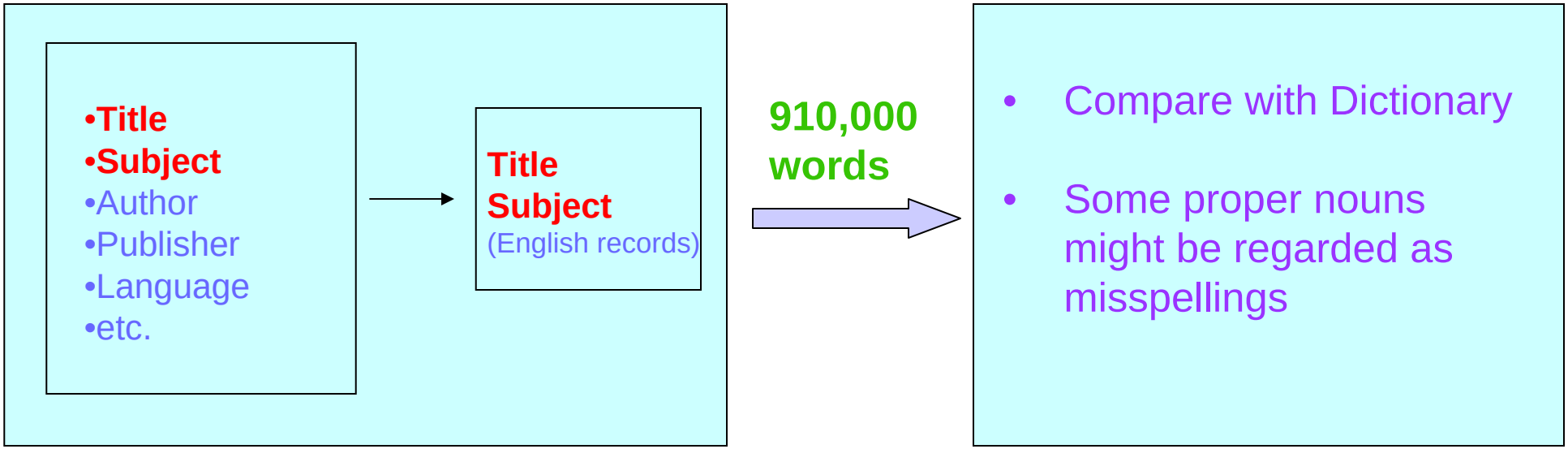
Case1: misspellings from the query

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

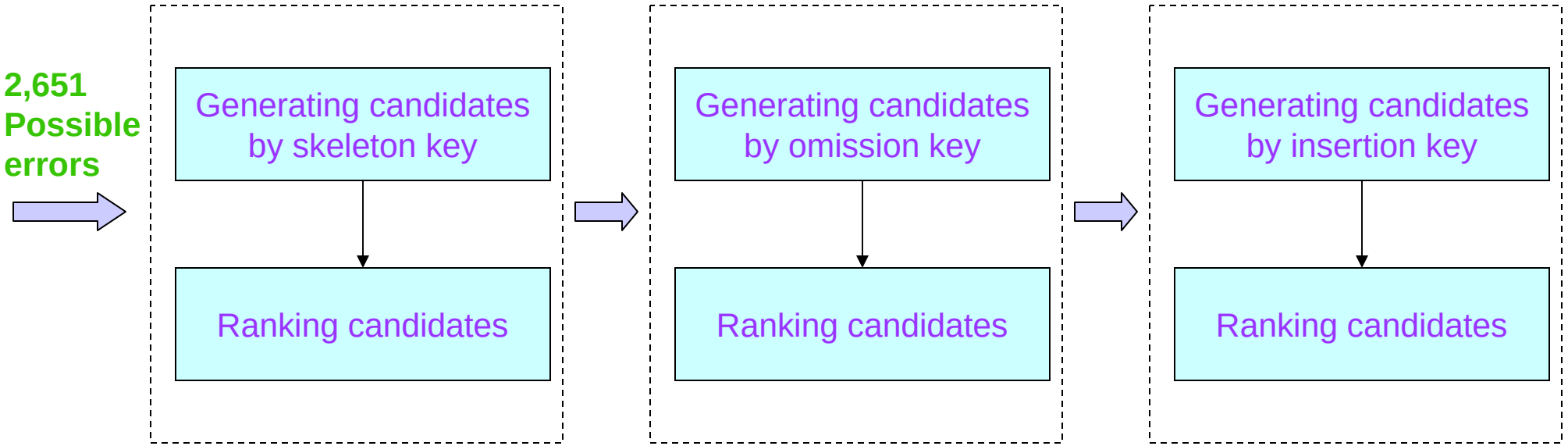
Processes of experiment

1. Extract words from records

2. Detect misspelling



3. Generating candidate corrections



assesment → asmnte

assesment → tmnsae

assesment → nmtsae

4. Ranking candidate corrections

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

- **Ranking**
  - *Minimum edit distance*
    e.g. d ("iniatiative", "initiative") =1 < d ("iniatiative", "inattentive")=4
  - *Keyboard distance*
    e.g. kd ("w", "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

Correction schemes	Could not be corrected	Be corrected	
		Properly	Improperly
Skeleton key	956 (36%)	1444 (54.5%) (48.75~71.73%)	251 (9.5%)
Omission key	498 (18.8%)	1879 (70.9%) (56.25~76.95%)	274 (10.3%)
Insertion key	424 (16%)	1935 (73%) N/A	292 (11%)

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

Key	Spelling1	Spelling2	Correction
mrtaio	Marriot (1)	Marriott (10)	Marriott
mlysai	Malayasia (2) Malalysia (1)	Malaysia (158)	Malaysia
pnyldkeia	Penyelidekan (2)	Penyelidikan (19)	Penyelidikan
krshnmtiau	Krishnamurthi (2)	Krishnamurti (7)	Krishnamurti

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