

# A Comprehensive Bangla Spelling Checker



---

**Naushad UzZaman and Mumit Khan**

*Center for Research on Bangla  
Language Processing*

**BRAC University  
Bangladesh**

International Conference on  
Computer Processing of Bangla  
(ICCPB 2006)  
17 February, 2006  
Dhaka, Bangladesh



# Outline

---

- Introduction
- Previous work in each step and propose solutions for each step
- Performance of our proposed solution



# Introduction

---

- Spelling checker
  - Detect misspelled words
  - Generate suggestions for misspelled word
  - Rank the suggestions
- Can be used in
  - Word processors
  - Optical Character Recognition (OCR)
  - Text To Speech (TTS)
  - Automatic Speech Recognition (ASR)
  - Machine Translation
  - Many more...



# Detecting misspelled words

---

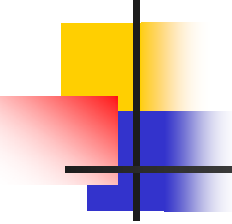
- Kukich (1992) breaks down human typing errors in two classes
  - Typographical error
    - People's mistake while typing
    - E.g. *spell* as *speel*
  - Cognitive error
    - Do not know how to spell the word
    - E.g. *separate* as *seperate*



# Detecting misspelled words

---

- Cognitive error
  - Phonetic error
    - Substituting a phonetically equivalent sequence of letters
    - E.g. *separate* as *seperate*
  - Homonym error
    - Accidentally produce a real word. a.k.a. real word error
    - E.g. *there* as *their*



# Previous work on detecting misspelled word

---

- Typographical error and cognitive phonetic error is trivial
- Cognitive homonym error is not trivial
- BB Choudhury (2001) and Abdullah and Rahman (2004) uses direct dictionary look-up and approximate string matching for detecting misspelled word
- We used dictionary look-up
- Cognitive homonym error is *not solved*



# Generating suggestions for misspelling words

---

- Error patterns of typographical error:
  - Damerau (1964) founds in English that 80% single error misspelling (insertion, deletion, substitution, transposition error)
  - BB Choudhury (2001) founds in more than 15 million Bangla words
    - 41.36% words due to single error misspelling (*error zone = 1*)
    - 32.94% (error zone = 2)
  - *Error zone is subset of edit distance*
  - Solution: Edit distance 2



# Previous work on typographical error

---

- Considers edit distance 2
- BB Choudhury (2001) handles using error zone 2
- Abdullah and Rahman handles using *recursive simulation* method



# Previous work on typographical error

---

- BB Choudhury's (2001) method needs twice the amount of memory for reverse dictionary
- Abdullah and Rahman's (2004) method requires  $m^{(2*n+1)}$  dictionary lookup.
  - $n$  = length of word
  - $m$  = average no of words in the *circular list*



# Our proposal for typographical error

---

- Given a query word with length 'n'
- Take the subset of lexicon, within length  $n+2$  and  $n-2$
- Generate edit distance with the subset

# Error pattern for phonetic error

- There are groups of phonetically similar characters in Bangla;
  - NA (ন) and NNA (ণ)
  - SA (স), SHA (শ) and SSA (ষ)
- Bangla has many consonant clusters or conjuncts with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.):
  - let us consider ক্ষ. ক্ষ = ক+্+ষ; ক্ষত is pronounced as খত, where ষ does not have any sound.

# Error pattern for phonetic error

- Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ.
  - At the beginning of word, it is pronounced as খ. (ক্ষত → খত);
  - In the middle or at the end of a word, it is pronounced as কখ, (দক্ষ → দকখ).
- Multiple pronunciations of some letters in the same context, such as হ with ব:
  - ভ: আহ্বান → আওভান /aovan/
  - আহ্বান is usually pronounced as আহভান /ahobhan/.
  - Both pronunciations are considered correct.

# Previous work on phonetic error



---

- BB Choudhury (2001),
- Abdullah and Rahman (2004),
- Hoque and Kaykobad's Soundex (2002) and
- UzZaman and Khan's Soundex (2004)
  - Deals phonetic errors in small scale
  - Mostly considers the first case shown before
    - Groups of phonetically similar character in Bangla
      - NA (ন) and NNA (ণ)
      - SA (স), SHA (শ) and SSA (ষ)

# Proposed solution for phonetic error

- অততন্ত - অত্যন্ত
  - দকখ - দক্ষ
  - সনধা - সন্ধ্যা
  - বেবোহার - ব্যবহার
- 
- Solution: Phonetic encoding



# Phonetic encoding

---

- *Code a word based on its pronunciation.*
  - অত্যন্ত - <ottnt> - অততন্ত
  - সন্ধ্যা- <shndha> - সনধা
  - ব্যবহার - <bebhar> - বেবোহার
- Naushad UzZaman and Mumit Khan, *A Double Metaphone Encoding for Bangla and its Application in Spelling Checker*, Proc. IEEE NLP KE, Wuhan, China, 2005

# Example of Spelling Checker Using Phonetic Encoding

Dictionary Word List	Encoded Word List
অকালপক্ক	“okalpkk”
সকাল	“skal”
পাষণ	“pasan”
দগ্ধ	“dgd”

Encoded Test word	Test Word
“skal”	শকাল





# Ranking the suggestion

---

- Solution: Edit distance
- BB Choudhury (2001), Abdullah and Rahman (2004) used edit distance
  - Can't rank suggestions phonetically
- We used combination of edit distance on phonetic encoding
  - Able to rank the suggestions phonetically and typographically



# Performance on 1607 common misspelled Bangla words

No of words	1607*
Correct (Edit Distance 0)	1473
Error	134
Rate of accuracy	91.67%
Rate of error	8.33%

\*Source of words: Bangla Banan Obhidhan, Dr. Khurshid Alam, Mirnava, Dhaka, Bangladesh.

# Performance on 1607 common misspelled Bangla words

No of words	1607
Correct (Edit Distance 0)	1473
Error	134
Rate of accuracy	91.67%
Rate of error	8.33%

Error	134	8.33%
Edit Distance 1	107	6.65%
Edit Distance 2	27	1.68%



# Summary

---

- Showed steps of spelling checker
  - Showed existing solutions in each step
  - Proposed solutions for each step
- For our particular sample set we get a 100 % accuracy by using a combination of phonetic encoding and edit distance-2.



# Acknowledgment

---

- Supported in part by the PAN Localization Project ([www.panl10n.net](http://www.panl10n.net)), grant from the International Development Research Center, Ottawa, Canada and BRAC University.



# References

---

- Karen Kukich, “Techniques for automatically correcting words in text”, ACM Computing Surveys, 24 (4), page 377 - 439”, 1992.
- B. B. Chaudhuri, “Reversed word dictionary and phonetically similar word grouping based spell-checker to Bangla text”, Proc. LESAL Workshop, Mumbai, 2001.
- Arif Billah Al-Mahmud Abdullah and Ashfaq Rahman, “A Different Approach in Spell Checking for South Asian Languages”, Proc. 2nd International Conference on Information Technology for Applications (ICITA), China, 2004.



# References

---

- F.J. Damerau, “A technique for computer detection and correction of spelling errors”, communication of ACM, 7(3), 171-176, 1964.
- Md. Tamjidul Haque and M. Kaykobad, “Use of Phonetic Similarity for Bangla Spell Checker”, Page 182 – 185, Proc. 5th International, Conference on Computer and Information Technology, Dhaka, December, 2002.
- Naushad UzZaman and Mumit Khan, “A Bangla Phonetic Encoding for Better Spelling Suggestion”, Proc. 7th International Conference on Computer and Information Technology, Dhaka, Bangladesh, December, 2004.



# Edit distance

---

- **Definition:** The smallest number of insertions, deletions, and substitutions required to change one *string* into another.