

Spelling Checker for Bangla

Mumit Khan
BRAC University
Dhaka, Bangladesh

Spelling Error Patterns

- Real word error
- Non word error

Non word errors

- 80% of all misspelled words (non-words) are single error misspellings: a single one of the following:
 - Substitution error: mistyping *the* as *ther*
 - Deletion error: mistyping *the* as *th*
 - Insertion error: mistyping *the* as *thw*
 - Transposition error: mistyping *the* as *hte*

[Jurafsky and Martin]

Typing errors

- Kurich (1992) breaks down human typing errors into two classes.
 - Typographic error (misspelling *spell* as *speel*)
 - Cognitive error (misspelling *separate* as *seperate*)

Cognitive error

- Two types
 - Phonetic error (substituting a phonetically equivalent sequence of letters, for example *separate* and *seperate*).
 - Homonym error (substituting *piece* and *peace*)
- Homonym error is a part of real word error, which has not been covered for Bangla yet.
- Phonetic error is handled using phonetic encoding.

Solution for Typographic Error

- Approximate string matching algorithm
 - Levenshtein Edit Distance
 - It suits most for string matching algorithms. Its generic to every languages. Not specific to Bangla.

Levenshtein Edit Distance

- The edit distance of two strings, s_1 and s_2 , is defined as the minimum number of point mutations required to change s_1 into s_2 , where a point mutation is one of:
 - Replace a letter,
 - Insert a letter,
 - Delete a letter,
 - Transpose consecutive letters

Example

- *Example:*
- $e(\text{"Virginia"}, \text{"Vermont"}) = 5$
- Virginia
- Verginia
- Verminia
- Vermonia
- Vermonta
- Vermont

Edit Distance in a Bangla Spelling Checker

- Lexicon: কথা, কাক, কলা, মালা
- Our misspelled word is কল
- Hence, our ranked suggestion for কল will be কলা, কাক, কথা, মালা

Dictionary word	Edit Distance with word কল
কথা	2
কাক	2
কলা	1
মালা	3

The Problem with Edit Distance

- Time complexity $O(n)$.
- If lexicon has $> 100,000$ words, misspelled word may need to be checked against each entry.

Solution of Edit Distance Problem

- Damerau (1964) found that 80% of all misspelled words (non-word errors) were caused by single-error misspellings.
- In another corpus, Peterson (1986) found that single error misspellings accounted for an even higher percentage of all misspelled words (93-95%).

– *Daniel Jurafsky and James H. Martin*

Single Error Misspelling

- Assume single error misspelled word.
- Only find ED with those in the lexicon that will have at most 1 error.
- Assuming
 - misLen = length of misspelled word
 - Length = length of the word being checked against
- So only consider the words with
 - (length == misLen ||
 - length - 1 == misLen ||
 - length + 1 == misLen)

Phonetic
error

Challenges for phonetic errors

- Bangla has many consonant clusters or juktakkhor with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.): let us consider ক্ষ. ক্ষ = ক+্+ষ; ক্ষত /kʰɔ̃t̪o/ is pronounced as খত /kʰɔ̃t̪o/, where ষ does not have any sound.
- Bangla has different uses of *Phalaa's*, such as BA, MA, YA, RA and LA phalaa.

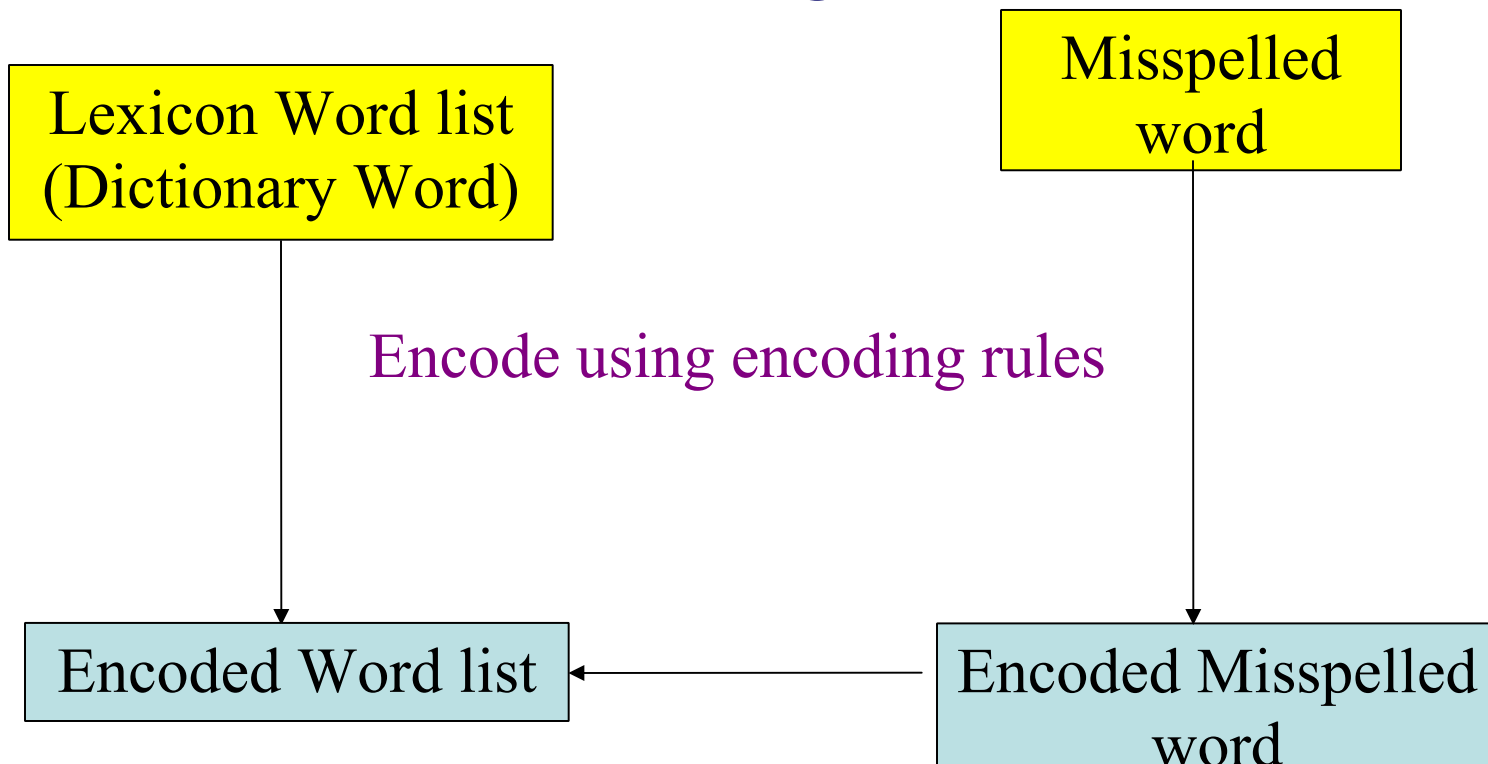
Challenges for phonetic errors

- Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ.
 - At the beginning of word
 - (ক্ষত → খত /kʰɔt̪o/);
 - In the middle or at the end of a word
 - (দক্ষ → দকখ /dɔkkʰo/).
- Multiple pronunciations of some letters in the same context, such as হ with ব:
 - আহ্বান → আওভান /aovan/.
 - আহ্বান → আহভান /aɦobɦian/

Phonetic Encoding

- Encode a word based on how it is pronounced.
- **realise** and **realize** in English, and ক্ষত and খত in Bangla should get the same code.

Use of Phonetic Encoding in a Spelling Checker



Search the encoded misspelled word in the encoded word list rather than searching the misspelled word in the Dictionary word list

Example of Spell Checking Using Encoding

Lexicon Word List	Encoded Word List
অকালপক্ক	"okalpkk"
সকাল	"skal"
চাঁদ	"cad"
দগ্ধ	"dgd"

Encoded Misspelled word	Misspelled Word
"skal"	শকাল

Search the encoded misspelled word in the encoded word list rather than searching the misspelled word in the Dictionary word list

Phonetic Encoding in English

- Soundex
- Metaphone
- Phonix
- Double Metaphone

Soundex Table

<i>Code</i>	<i>Letters</i>
0 (not coded)	A, E, I, O, U, H, W, Y
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Example of Soundex

- Realize – 6004020 – 642
- Realise – 6004020 – 642

- Knight – 250203 – 2523
- Nite – 5030 – 53

Metaphone Transformations

- B -> B unless at the end of a word after "m" as in "dumb"
- C -> X (sh) if -cia- or -ch-
- S if -ci-, -ce- or -cy-
- K otherwise, including -sch-
- D -> J if in -dge-, -dgy- or -dgi-
- T otherwise
- F -> F
- G -> **silent if in -gh-** and not at end or before a vowel
- in -gn- or -gnd- (also see dge etc. above)
- J if before i or e or y if not double gg
- K otherwise
- H -> silent if after vowel and no vowel follows
- H otherwise
- J -> J
- K -> silent if after "c"
- K otherwise
- L -> L
- M -> M
- N -> N

- P -> F if before "h"
- P otherwise
- Q -> K
- R -> R
- S -> X (sh) if before "h" or in -sio- or -sia-
- S otherwise
- T -> X (sh) if -tia- or -tio-
- 0 (th) if before "h"
- silent if in -tch-
- T otherwise
- V -> F
- W -> silent if not followed by a vowel
- W if followed by a vowel
- X -> KS
- Y -> silent if not followed by a vowel
- Y if followed by a vowel
- Z -> S
-
- Initial Letter Exceptions
-
- Initial **kn-**, gn- pn, ac- or wr- -> **drop first letter**
- Initial x- -> change to "s"
- Initial wh- -> change to "w"

Example of Metaphone

- Knight – NT
- Nite – NT

- Basinger is pronounced in both way as “Basin-gger” or “Basin-ger”.
- Basinger - BSNJR
- Basin-gger - BSNKR
- Basin-ger - BSNJR

Key Concepts From English

- Soundex: groups the letters of same pronunciation.
- Metaphone & Phonix: also considers the context of a letter to encode it.
- Double Metaphone: gives multiple codes to same word, if it is pronounced in more than two ways.

Existing Encoding in Bangla

- Hoque and Kaykobad's Soundex type encoding, 2002
- Zaman and Khan's Soundex type encoding, 2004

Hoque and Kaykobad's Soundex Table

Name	Group Member
1	ক, খ, গ, ঘ, ঙ্গ
2	চ, ছ, জ, ঝ, য
3	ট, ঠ, ড, ঢ
4	ত, থ, দ, ধ, ত্
5	প, ফ, ব, ভ
6	ঙ, ঞ, ং
7	শ, স, ষ
8	র, ড়, ঢ়, ঝ
9	ন, ণ
α	ম
β	ল

Example of Hoque and Kaykobad's Soundex

- For example, কৰ্ম will be converted to a 4 lengthen sound code as “ক8a0”, with zero padding.

Zaman and Khan's Soundex

Code	Group members
0	্, ো, ঁ
“a”	আ, া
“i”	ই, ঈ, ি, িী
“u”	উ, ঊ, ু, ুু
“e”	এ, ে, ঐ, ঐে
“o”	অ, ও, ঔ, ৌ
“k”	ক, খ
“g”	গ, ঘ
“m”	ম, ঙ, ং
“c”	চ, ছ
“j”	য, জ, ঞ

A sample of the
actual table

Example of Zaman Khan's Soundex

Input	Encoding	Suggestion
খুমাড়	kumar	কুমার
পাসান	pasan	পাষণ
দগধ	dgd	দন্ধ (দগ ্ ধ)

Limitation of Existing Encodings

- Bangla has many consonant clusters or juktakkhor with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.): let us consider ক্ষ. ক্ষ = ক+্+ষ; ক্ষত /kʰɔ̃t̪o/ is pronounced as খত /kʰɔ̃t̪o/, where ষ does not have any sound.
- Bangla has different uses of *Phalaa's*, such as BA, MA, YA, RA and LA phalaa.

Limitation of Existing Encodings

- Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ.
 - At the beginning of word
 - (ক্ষত → খত /kʰɔt̪o/);
 - In the middle or at the end of a word
 - (দক্ষ → দকখ /dɔkkʰo/).
- Multiple pronunciations of some letters in the same context, such as হ with ব:
 - আহ্বান → আওভান /aovan/.
 - আহ্বান → আহভান /aɦobɦian/

Double Metaphone Phonetic Encoding

Sample Encoding Rules for য

Soundex Encoding

“j”	য	YA	“09AF”
	জ	JA	“099C”
	ঝ	JHA	“099D”

Double Metaphone Encoding

য	YA as fola	x”09CD”“09AF”	“e”	@ the beginning as YA fola	ব্যথিত, ব্যক্ত, ন্যস্ত
		...xy”09CD”z”09CD”09A F”	Not Coded	@ middle/end with jukhtakhor	সঙ্ক্যা, মর্ত্য
		...xy”09CD”09AF”	Doubles: yy	@ middle/end	অদ্য, মধ্য
য	YA	“09AF”	“j”		
জ	JA	“099C”	“j”		
ঝ	JHA	“099D”	“j”		

Sample Encoding Rules for ক্ষ

Soundex Encoding

"k"	ক	KA	"0995"
0 (zero)	্	Virama/Hasant	"0981"
"s"	ষ	SSA	"09B7"


Double Metaphone Encoding

ক্ষ	"0995""09CD""09B7"	"k"	@the beginning	ক্ষত
ক্ষ	"0995""09CD""09B7"	"kk"	@ middle/end	দক্ষ

Spelling Checker using Phonetic Encoding

Lexicon	Encoded
Word List	Word List
অকালপক্ক	“okalpkk”
সকাল	“skal”
চাঁদ	“cad”
দগ্ধ	“dgd”

Encoded Misspelled word	Misspelled Word
“skal”	শকাল



Search the encoded misspelled word in the encoded word list rather than searching the misspelled word in the Dictionary word list

Encoding Examples

Wrong Word	Correct Word (Error)	Soundex Encoding		Metaphone Encoding	
		Wrong Word	Corrent Word (Error)	Wrong Word	Correct Word (Error)
কসট	কষ্ট (2)	“kst”	“kst” (0)	“kst”	“kst” (0)
দুকথ	দুঃখ (1)	“dukk”	“duhk”(1)	“dukk”	“dukk” (0)
ষামি	স্বামী (3)	“sami”	“sbami”(1)	“sami”	“sami” (0)
রিদয়	হৃদয় (2)	“ridy”	“hrdy” (2)	“ridy”	“ridy” (0)
নেতরি	নেত্ৰ (2)	“netri”	“netr” (1)	“netri”	“netri” (0)

Encoding Examples

Wrong Word	Correct Word (Error)	Soundex Encoding		Metaphone	Encoding
		Wrong Word	Corrent Word (Error)	Wrong Word	Correct Word (Error)
অকালপকক	অকালপক্ক (1)	“okalpkk”	“okalpkk”(0)	“okalpkk”	“okalpkk”(0)
শকাল	সকাল (1)	“skal”	“skal” (0)	“skal”	“skal” (0)
চাদ	চাঁদ (1)	“cad”	“cad” (0)	“cad”	“cad” (0)
দগধ	দগ্ধ (1)	“dgd”	“dgd” (0)	“dgd”	“dgd” (0)
বিসশো	বিশ্ব (2)	“biss”	“bisb” (1)	“biss”	“biss” (0)

Generate Suggestion List

- Include all the words with same phonetic encoding (handle phonetic error).
- Include all the word having edit distance n (handle transposition error).
 - Here, $n = 1$
 - User can handle edit distance of any number 1, 2, 3, 4, etc. So, it can handle any transposition errors but the trade of is time complexity.

How to Rank Suggestions

- For a misspelled word generate a score with all the dictionary words.
- $\text{score} = a * \text{phonetic_edit_distance} + b * \text{normal_edit_distance}$
- where, $a > b$
- Here, $\text{phonetic_edit_distance} = 0$

Encoding Performance

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

Error Distribution

Error	134
Edit Distance 1	107
Edit Distance 2	27

Current Work

- Delivered spelling checker bundled with documentation.
- Studying some well-established and/or well-designed spelling checkers such as Jazzy, JSpell, ASpell, etc, into account to follow the best practices.

Future Work

- Integrate the spelling checker into Office packages like OpenOffice and MS Office.
- Create component-based spelling checker services. J2EE and .NET implementations in progress.