

Urdu Spell Checking

Dr. Sarmad Hussain

Tahira Naseem

CRULP, NUCES, PAKISTAN

www.crulp.org



PAN
Localization

www.PANL10n.net

Outline

- # **Spelling errors trends in Urdu**
 - # **Solution**
 - **Correction techniques**
 - **Ranking techniques**
 - # **Results**
 - # **References**
-

Spelling errors trends

Space Omission Errors (above 70%)

آ جا رہے → آ جا رہے

Shape similarity based errors

انتخاب → انتخاب

Sound similarity based errors

’لحاض‘ → ’لحاظ‘

Spelling errors trends

Effect of Shift keys

مان → ماں

25% of the errors are real-word errors

سلامی → اسلامي

Solution

Urdu Spell Checking

Error Correction

- Soundex variations for Urdu
- Edit Distance
- Space omission errors handling

Ranking (based on)

- Frequency
 - Soundex
 - Shapex
-

Soundex

Urdu Spell Checking

Variations

- Encoding first letter
- Increasing Number of letter groups
 - Using Hexadecimal (0-F) codes (instead of decimal codes)

Ranking (based on)

- Frequency
-

Soundex 0-F Codes

Methodology

Code	Alphabets	Code	Alphabets
0	ث س ص ش	A	ر ژ
1	ت ط ة	B	ژ ی
2	ز ض ظ ذ	C	آ ع
3	ج چ	D	ف
4	ح ه	E	ل
5	خ ک ق	F	و
6	د		
7	ب پ		
8	ن م		
9	گ		

Normalization

- ۵ → ا Word finally
- ه → ۵ With un aspirated characters
- ا, و and ى become consonants, **Syllable** initially

Soundex

Methodology

Results

	Top One %age	Top Five %age	Top Ten %age	Above Ten %age	Total Recall %age	avg size of result set
Soundex 0-F (with 1st letter encoded)	26.2	44.4	49.5	5.4	54.8	21
Soundex 0-9 (with 1st letter encoded)	20.1	35.5	43.0	17.9	60.9	92
Soundex 0-F-L (with 1st letter not encoded)	23.7	36.2	39.1	0.7	39.8	7
Soundex 0-9-L (with 1st letter not encoded)	22.6	40.9	46.2	3.2	49.5	21

Edit Distance

Methodology

- # Generate corrections by applying Single Edit in Reverse
- # Check the validity of generated correction
 - Bigram validity test
 - Dictionary Look-up
- # Ranking
 - Frequency based ranking
 - Sound similarity based ranking
 - Shape similarity based ranking

Ranking

Urdu Spell Checking

- Frequency based ranking
 - Sound similarity based ranking
 - Shape similarity based ranking
 - Ranking for space omission errors
-

Ranking

Urdu Spell Checking

Sound similarity based ranking

corrections whose Soundex code matched the Soundex code of error were considered more likely for being actual intended word.

Example:

لحاظ is a more likely correction for لحاض compared to لحاف .

Ranking

Urdu Spell Checking

Frequency based ranking

If a word is more frequent in the language, it is ranked higher.

Example:

کروٹ is a more likely correction for کھوٹ compared to کوٹ .

Ranking

Urdu Spell Checking

Shape similarity based ranking

- 35% of the substitution are shape based in Urdu so a substitution with shape similarity is ranked higher.

کنگن is a more likely correction for کنکن compared to کندن .

- Shape codes (Shapex)

Code **Alphabets**

0	آ آل
2	ت ٹ ث ن س ش
4	د ڈ ذ ر ژ ر ژ وؤ
6	ط ظ
8	ک گ

Code **Alphabets**

1	ب ہ پ ی ئی
3	ح خ ج چ
5	ص ض
7	ع غ ف ق م
9	ے ئے

Edit Distance

Methodology

Results

	Top One %age	Top Five %age	Top Ten %age	Above %age	Total Recall %age	avg size of result set
SE + FR	58.1	90.0	93.5	1.1	94.6	8.5
SE + SXR + FR	64.2	89.6	93.9	0.7	94.6	8.5
SE + SPR + FR	64.5	89.6	93.5	1.1	94.6	8.5
SE + SXR + SPR + FR	71.7	87.1	93.5	1.1	94.6	8.5

SE: Single Edit

FR: Frequency based Ranking

SXR: SoundeX based Ranking

SPR: ShaPex based Ranking

Space omission errors

Space deletion

- Following joining character
 - Deal like other deletion errors using reverse single edit.
 - Following non-joiner character
 - Break the words on non-joiner positions (into 2 or more parts)
 - Generate all possible partitions
 - Check the partitions for validity
-

Space omission errors

Ranking

- Frequency based ranking
 - Average of parts' frequencies
 - Frequency of the least frequent part
- Ranking on the basis of number of parts in a valid partition, more parts are less likely.
- Space omission after non-joiner characters is more likely than space omission after joiner characters.

آسی is a more likely correction for آسکی compared to آس کی.

Combined approach

Single Edit + Soundex + Space related error correction

results from all three techniques are gathered and best 10 results are selected.

Combined Approach

Results

	Top One %age	Top Two %age	Top Five %age	Not Found %age	Total Recall %age	avg size of result set
SE+SPD+ SX	82.57	93.91	96.27	3.32	96.68	6
SE+SPD	82.43	93.08	96.13	3.46	96.54	6

References

Alberga, C.N. String Similarity and Misspelling, *In Communications of ACM*, Vol. 10, No. 5, pp. 302-313, May, 1967.

Amir, A. et. al., Indexing and dictionary matching with one error. (Extended Abstract)

Appel, A. and Jacobson, G. 1988. The world's fastest scrabble program. *Communications of the ACM* 31, 5 (May), 572–578.

References

Brill, E. and Moore, R. C. An Improved Error Model for Noisy Channel Spelling Correction. *In proceedings of 38th Annual meeting of Association for Computational Linguistics*, pp. 286-293, 2000.

Brodal, G.S., Approximate Dictionary Queries ,BRICS*** Computer Science Department

Christian, Peter, Soundex – can it be improved?, *Computers in Genealogy* Vol. 6, No. 5, March, 1998.

References

Damerau, F.J. A Technique for Computer Detection and Correction of Spelling Errors. *In Communications of ACM*, Vol. 7, No. 3, pp. 171-177, March, 1964.

Erikson, K. Approximate Swedish Name Matching - Survey And Test Of Different Algorithms, 1997.

Fisher, M.W. A statistical text to phone function using ngrams and rules. *In Proceedings Of The IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 649-662, 1999.

References

Holmes, David and McCabe, M. C., Improving precision and recall for Soundex

Hodge, V. J. & Austin, J., A Comparison of Standard Spell Checking Algorithms and a Novel Binary Neural Approach. *IEEE transactions on knowledge and data engineering*, Vol. 15, No. 5, September 2003.

Kann, Viggo et. al., Implementation Aspects And Applications Of A Spelling Correction Algorithm, May 1998.

References

Kernighan et. al. A Spelling Correction Program Based on Noisy Channel Model, In Proceedings of COLING-90, *The 13th International Conference On Computational Linguistics*, Vol 2. 1990.

Kukich, K. Techniques for automatically correcting words in text, *ACM Computing Survey*, Vol. 14, No. 4, pp 377-439, December 1992.

Peterson, L.J. A Note on Undetected Typing Errors. *In Communications of ACM*, Vol. 29, No. 7, pp. 633-637, July, 1986.

References

Stanier, Alan, How accurate is Soundex matching, *Computers in Genealogy* Vol. 3, No. 7, pp. 286-288. September 1990.

Toutanova, K. and Moore, R. C. Pronunciation Modeling for Improved Spelling Correction, *In proceedings of 40th Annual meeting of Association for Computational Linguistics*, July 2002, pp. 144-151.

Turba, T. N. 1981. Checking for spelling and typographical errors in computer based text. *SIGPLANSIGOA Newslett.* (June), 51-60.

References

Zobel, J., Et. Al. Finding Approximate Matches In Large Lexicons, October 1994.

Zobel, J., Et. Al. Searching Large Lexicon For Partially Specified Terms.

Zobel, J. and Dart, P. Phonetic String Matching: Lessons from Information Retrieval.



Thank You