

# Spell Checking



Dr. Sarmad Hussain  
Tahira Naseem  
CRULP, NUCES, PAKISTAN  
[www.crupl.org](http://www.crupl.org)



# Outline

- ◆ **Spelling Errors**
- ◆ **Error Detection**
- ◆ **Error Correction**
  - **Non-word error correction**
  - **Real-word error correction**

# Spelling Errors

- ◆ Typographic errors vs. Cognitive errors
  - Taht (that) vs. seprate (separate)
- ◆ Non-Word errors vs. Real-Word errors
  - Non-Word error the → teh
  - Real-Word error form → from  
(Near 40% of the errors are real word errors.)

# Spelling Errors

## ◆ Basic types of errors

- Insertion; e.g. typing acress for cress
- Deletion, e.g. typing acress for actress
- Substitution, e.g. typing acress for across
- Transposition, e.g. typing acress for caress

	GPO	Web7
<b>Transposition</b>	<b>4 (2.6%)</b>	<b>47 (13.1%)</b>
<b>Insertion</b>	<b>29 (18.7%)</b>	<b>73 (20.3)</b>
<b>Deletion</b>	<b>49 (31.6%)</b>	<b>124 (34.4%)</b>
<b>Substitution</b>	<b>62 (40.0%)</b>	<b>97 (26.9%)</b>
<b>Total</b>	<b>144 (92.9%)</b>	<b>341 (94.7%)</b>

**GPO:** Government  
Printing Office

**Web7:** Webster  
dictionary

# Spelling Errors

- ◆ Factors affecting error patterns
  - Keyboard adjacencies (cause basic errors)
  - Shift key characters
  - Sound similarity      receive → recieve
  - Shape similarity      انتخاب → انتحاب



---

# Spell Checkers

---

- ◆ Functions of Spell checker
  - Error detection
  - Error correction



# Error Detection

- ◆ Identify word boundaries in text  
(generally spaces and punctuations, mark words' boundaries)
- ◆ Given a word, look it up in the dictionary for validation.  
(this works only for Non-Word errors.  
Context based error detection is used for real word errors.)



---

# Error Detection

---

- ◆ What if there is no word boundary delimiter.

for example in Thai and Chinese language there are no spaces between words.

- Use Dictionary + Words' frequencies  
[Sproat 1996]

# Spell Checkers

## Types of Error Correction

- ◆ Automatic

Error is automatically replaced with correction with out user intervention

Spellchecker should be able to provide one best correction suggestion.

- ◆ Interactive

User can interactively select one of the suggested corrections for replacement.

spellchecker can suggest multiple corrections.



---

# Correction Techniques

---

## Types of Correction Techniques

- ◆ Isolated Word Error Correction  
for non-word errors
- ◆ Context based Error Correction  
for ranking corrections  
for real word errors detection

# Non-Word error correction techniques

- ◆ Soundex and its variants
- ◆ Edit distance
- ◆ N-Grams
- ◆ Probabilistic error correction techniques

# Soundex

- ◆ Soundex algorithm tries to assign common codes to similar sounding words and names.

# Soundex Algorithm

1. Retain the first letter of the name, and drop all occurrences of a,e,h,i,o,u,w,y in other positions.
2. Assign the following numbers to the remaining letters after the first:  
b, f, p, v  $\rightarrow$  1      d, t  $\rightarrow$  3      m, n  $\rightarrow$  5  
c, g, j, k, q, s, x, z  $\rightarrow$  2      l  $\rightarrow$  4      r  $\rightarrow$  6
3. If two or more letters with the same code were adjacent in the original name (before step 1), omit all but the first.
4. Convert to the form "letter, digit, digit, digit" by adding trailing zeros (if there are less than three digits), or by dropping rightmost digits (if there are more than three).

# Soundex Example

- ◆ Robert → R163 Robin → R150
- ◆ Smith → S530 Smyth → S530

1	b, f, p, v
2	c, g, j, k, q, s, x, z
3	d, t
4	l
5	m, n
6	R

**Soundex Codes**  
[Erikson 1997]

# Soundex Performance

- ◆ Low Hit Rate

25% of the relevant words go undetected

- ◆ Low Precision

Large size words sets are detected

only 33% of detected words are actually relevant

[Stanier 1990]

# Reasons of Shortcomings

- ◆ Large size of group 2.  
decreasing its size will reduce the size of detected result set of words.
- ◆ Retaining initial letter as it is.
- ◆ Ignoring the possibility of sounds produced by more than one letters.
- ◆ Ignoring the omission and change in sounds of letters in consonants clusters.       $dg \rightarrow j$ ,  $tch \rightarrow ch$
- ◆ Ignoring the ending characters of long words

# Possible Enhancements

- ◆ Use smaller groups of characters
- ◆ Assign code to leading letter
- ◆ N- Grams Substitution
- ◆ Multiple length codes
- ◆ Multiple codes for one word
- ◆ Overlapping groups of letters
- ◆ Position Information

Phonix is a soundex variant that makes some of these enhancements.

# Phonix

- ◆ Phonix is a soundex variant [Erikson 1997]
- ◆ Codes assigned to letters are different
  - 1 → b, p
  - 2 → c, g, j, k, q
  - 3 → d,
  - 4 → l
  - 5 → m, n
  - 6 → r
  - 7 → f, v
  - 8 → s, x, z
- ◆ prior to code generation, string is normalized by applying some letter groups' substitutions also called N-gram substitution. For example the sequence *tch* is mapped to *ch*, *ph* is mapped to *f*.

# Edit Distance Techniques

- ◆ Single Edit Distance
- ◆ Multiple Edits Distance
- ◆ Editex

# Single Error Technique

- ◆ Generate all those strings from which the erroneous word can be formed by applying any of the single error operations. (insertion, deletion, substitution or transposition)
- ◆ Check the strings in the dictionary
- ◆ Filter out those strings which are not valid words in the language. (letter-N-Gram validity filter is also applied for efficiency [Kann 1998])
- ◆ Present remaining words as suggestions

# Problems

- ◆ 20% of the errors (multiple errors) are being completely ignored.
- ◆ No weight is given to phonetic similarity. For example according to the algorithm, for the misspelling 'regect', both 'reject' and 'regent' are equally likely corrections.

# Edit Distance

- ◆ Edit distance between two strings is the minimum number of editing operations (insertion, deletion, substitution and transposition) required to convert one string to the other.

Edit distance between “acress” and “actress” is 1

Edit distance between “atometric” and “automatic” is 2

# The Algorithm

- ◆ Compute edit distance between erroneous word and all dictionary words.  
(generally using a dynamic programming algorithm)
- ◆ Select those dictionary words whose edit distance is within a pre specified threshold value.
- ◆ Present these words as suggestions

# Problem

- ◆ Again, phonetic similarities are not being considered.

# Editex

- ◆ A mixture of soundex and edit distance
- ◆ If a letter from a soundex group is substituted by another letter from the same group then the contribution of this substitution operation to the edit distance is half of the any other single error operation.

# Problem

- ◆ Though phonetic similarity is being taken into consideration but other factor causing spelling mistakes, like keyboard adjacencies, visual similarity etc. are still being ignored.

# N-Gram Based Technique

- ◆ N-Grams

An N-gram is a sequence of N adjacent letters in a word

The more N-grams, two strings, share the more similar they are.

similarity coefficient  $\delta$

$$\delta = |\text{common N-grams}| / |\text{Total N-grams}|$$

# N-Gram Example

- ◆ **N-Gram similarity example:**

fact vs. fract

Bigrams in fact : -f fa ac ct t- 5 bigrams

Bigrams in fract : -f fr ra ac ct t- 6 bigrams

Union : -f fa fr ra ac ct t- 7 bigrams

Common : -f ac ct t- 4 bigrams

$$\delta = 4/7 = 0.57$$

# Problems

- ◆ N-gram technique doesn't show good performance on short words. For example when using trigrams, the words of length 3 will share no trigram with themselves just after introduction of single error.
- ◆ N-gram similarity measure works best for insertion and deletion errors, well for substitution errors, but very poor for transposition errors.

# Probabilistic Techniques

- ◆ Noisy Channel Model



# Noisy Channel Model

- ◆ Consider the phenomenon of making spelling mistakes as the process of sending text through a noisy communication channel, which introduces errors in the text.

Find the most probable transmitted word (correct dictionary word) for a received erroneous string (misspelling).

# Generic Algorithm

- ◆ The model assigns a probability to each correct dictionary word for being a possible correction of the misspelling. The word with highest probability is considered the closest match (or the actual intended word).

# Formal Description

Let  $D$  be a dictionary and  $w_i$  be any word in  $D$ , for a misspelled string  $s$  not present in  $D$  find such word  $w \in D$  for which  $P(w|s)$  is maximum

Hence

$$w = \operatorname{argmax}_{w_i} P(w_i|s) \quad (1)$$

Applying Bayes' rule we can rewrite this probability expression as

$$P(w_i|s) = (P(s|w_i)P(w_i)) / P(s) \quad (2)$$

(1) becomes

$$w = \operatorname{argmax}_{w_i} (P(s | w_i) P(w_i)) \quad (3)$$

The first of these terms  $P(s | w_i)$  is the probability of typing string  $s$  when word  $w_i$  was intended. (channel model)

$P(w_i)$  is the probability that a writer will type  $w_i$  from among all the dictionary words.

(source-model or language-model)



# Noisy Channel Model

- ◆ Channel Model
  - Channel is generally modeled using
    - letter-to-letter confusion probabilities. [Kernighan 1990]
    - string-to-string confusion probabilities. [Brill 2000]
- ◆ Language Model
  - Language can be modeled using
    - Letter to letter transition probabilities [Kukich 1990]
    - Word N-grams probabilities (N is b/w 1 and 3) [Brill 2000]

# Comparison

TECHNIQUE	PERFORMANCE
Soundex (Original)	Precision=33% Recall=75%
Phonix	
Single Error (Damerau)	84%
Levenshtein Edit Distance (grope)	<b>60%*</b>
Probabilistic Weighted Edit Distance	87%, <b>78%*</b>
N-Gram Vectors	<b>52%-74%*</b>
Probabilistic N-Grams	75%-78%, <b>75%*</b>
Noisy Channel (string to string confusion probabilities)	99%

\* [K. Kukich 1990]

# Real-Word Error Correction

- ◆ Real-word spelling errors can be viewed as violation of NLP constraints, there are mainly five types of constraints in NLP, [K. Kukich 1990]
  - Syntactic
    - The study was conducted mainly **be** John Black.
  - Semantic
    - He is trying to **fine** out.
  - Discourse
    - Iris flowers have **four** parts: standards, petals, and falls.
  - Pragmatic
    - Has Mary registered for Intro to **Communication** Science yet?
  
- ◆ context based correction techniques generally make use of syntactic and semantic knowledge to identify real word errors.

# Real-Word Error Correction

- ◆ Statistical context based correction techniques make use of word bigram and trigram probabilities in order to capture collocation trends.
  - Require large corpora and the gaps are still there
  - Computationally heavy

# Real-Word Error Correction

- ◆ POS N-Gram Probabilities
  - Relatively manageable
  - Can be trained on relatively small corpora
  - Errors within same POS go undetected.  
(pan → pain)

# Real-Word Error Correction

- ◆ Co-occurrence frequencies + Confusion sets
  - Confusion set is a set of words that can possibly be confused for a given word. For example  
(Pink → Pin, Pick, Ink)
  - Find frequent co-occurrences for each word of a confusion set
  - If an ambiguous word occurs in the text, find its best replacement (possibly itself) on the basis of neighboring words. [Tong]

# Real-Word Error Correction

## ◆ Summary

- Real word error correction techniques require
  - Large Corpora or/and
  - Matured language analysis

therefore....

for the languages for which these recourses are not available, real word error correction is a very difficult goal.



---

# 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 13<sup>th</sup> 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.
- Sproat, R et. Al. A stochastic finite-state word-segmentation algorithm for Chinese Volume 22 , Issue 3 (September 1996) table of contents, p 377 – 404, 1996, MIT Press



# 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.

A decorative graphic on the left side of the slide consists of a vertical stack of thin, light-colored horizontal lines. To the right of this stack, there are two solid olive-green vertical bars. A dark blue horizontal line spans across the top of the slide, starting from the left edge and ending before the first olive-green bar. A second dark blue horizontal line is positioned below the first, starting from the right edge and ending before the second olive-green bar.

Thank You