Spelling Errors Detection and Correction

Nada Naji
najin@ccs.neu.edu

College of Computer and Information Science
Northeastern University

Spell Checking

- Basic approach: suggest corrections for words not found in *spelling dictionary*
- Suggestions found by comparing word to words in dictionary using similarity measure
- Most common similarity measure is *edit distance*
  - number of operations required to transform one word into the other

Edit distance to “Nancy”?

Edit Distance

- *Damerau-Levenshtein* distance
  - counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
  - e.g., Damerau-Levenshtein distance 1

    extenssions $\rightarrow$ extensions(insertion error)
    poiner $\rightarrow$ pointer (deletion error)
    marshmellow $\rightarrow$ marshmallow (substitution error)
    brimingham $\rightarrow$ birmingham (transposition error)

  - distance 2

    doceration $\rightarrow$ deceration
    deceration $\rightarrow$ decoration

Edit Distance

- Number of techniques used to speed up calculation of edit distances
  - restrict to words starting with same character
  - restrict to words of same or similar length
  - restrict to words that sound the same
- Last option uses a *phonetic code* to group words
  - e.g. Soundex

Soundex Code

1. Keep the first letter (in upper case).
2. Replace these letters with hyphens: a,e,i,o,u,y,h,w.
3. Replace the other letters by numbers as follows:

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

4. Delete adjacent repeats of a number.
5. Delete the hyphens.
6. Keep the first three numbers or pad out with zeros.

extenssions $\longrightarrow$ E235; extensions $\longrightarrow$ E235
marshmellow $\longrightarrow$ M625; marshmallow $\longrightarrow$ M625
brimingham $\longrightarrow$ B655; birmingham $\longrightarrow$ B655
poiner $\longrightarrow$ P560; pointer $\longrightarrow$ P536

Spelling Correction Issues

- Ranking corrections
  - "Did you mean..." feature requires accurate ranking of possible corrections
- Context
  - Choosing right suggestion depends on context (other words)
  - e.g., *lawers* → *lowers, lawyers, layers, lasers, lagers* but *trial lawers* → *trial lawyers*
- Run-on errors
  - e.g., "mainscourcebank"
  - missing spaces can be considered another single character error in right framework

Noisy Channel Model

- User chooses word *w* based on probability distribution *P(w)*
  - called the *language model*
  - can capture context information, e.g. $P(w_1|w_2)$
- User writes word, but *noisy channel* causes word *e* to be written instead with probability *P(e|w)*
  - called *error model*
  - represents information about the frequency of spelling errors

Noisy Channel Model

- Need to estimate probability of correction
  - $P(w|e) = P(e|w)P(w)$
- Estimate language model using context
  - e.g., P(w) = $\lambda P(w) + (1 - \lambda)P(w|w_p)$
  - w_p is previous word
- e.g.,
  - “pumpkin spife”
  - “spice” and “spine” both likely corrections,
    but $P($spice$|$pumpkin$) > P($spine$|$pumkin$)$

Noisy Channel Model

- Language model probabilities estimated using corpus (set of documents) and query log
- Both simple and complex methods have been used for estimating error model
  - simple approach: assume all words with same edit distance have same probability, only edit distance 1 and 2 considered
  - more complex approach: incorporate estimates based on common typing errors (e.g. keyboard layout)

References & further readings

- Search Engines: Information Retrieval in Practice. Book by Donald Metzler, Trevor Strohman, and W. Bruce Croft. 2015.

- Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission. Levenshtein, Vladimir I. 1965.

- An improved error model for noisy channel spelling correction. In Proc. ACL, pp. 286-293. Brill, Eric, and Robert C. Moore. 2000