An Improved Error Model for Noisy Channel Spelling Correction

Eric Brill and Robert C. Moore
Outline

• Introduction
  – Noisy Channel Spell Correction

• Methodology
  – Improved Error Model
  – Training the Model

• Results
Ad Hoc vs Statistical Methods

• Specify allowed lists, try to use knowledge of language
  – Confusion sets {to, too, two}, {they, their, they’re}
  – ph -> f, s -> c
  – Static, not easy to port for new languages, not adaptable to new users

• Statistical
  – Automatically updates weights, learns errors
  – Can be trained for specific users
  – Flexible
Noisy Channel Modeling

Source Word (Correct intended word) ---|--- NOISY CHANNEL ---|--- Noisy Word (Misspelling)
Noisy Channel Modeling

- Given a misspelling, $s$, want closest dictionary word, $w$
- Want $\arg\max_w P(w \mid s)$
  - $\arg\max_w P(s \mid w) \cdot P(w)$
- $P(w)$ is called the source model
- $P(s \mid w)$ is called the error model
  - Want $P(\text{teh} \mid \text{the})$ to be high
  - Want $P(\text{hippopotamus} \mid \text{the})$ to be low
Error Model

- Simple, common error model is Levenshtein distance
  - Can be weighted
  - Substituting ‘e’ for ‘M’ vs ‘e’ for ‘i’

- Weaknesses of using Levenshtein distance?
  - Some ideas on next slide
Levenshtein Weaknesses

- No phonetic information
  - Physical vs fisikal is very large distance

- No context information
  - Substituting a for e common
  - But usually -ant vs -ent
  - Transposition i and e more common after c
  - Models $P(a \mid e)$ but not $P(ant \mid ent)$
Improved Error Model

• Allow all edit operations of the form $\alpha \rightarrow \beta$
  – $\alpha, \beta$ are strings, not necessarily single letters
  – More general than MDE
    • $\epsilon \rightarrow D$ is insertion of $D$
    • $Q \rightarrow D$ is substitution of $D$ for $Q$

• Also allow position information
  – Where in the word did the substitution take place?
  – How does position info + generalization help?
Improved Error Model

• Can capture more specific, richer edits
• Substitute -ent for -ant more common at end
  – antler vs entler not common
  – reluctant vs reluctantent is common
– P(ant -> ent | PSN)
– PSN = \{Start, middle, end, morpheme boundary\}
Improved Error Model

• Try partitioning the words

• physically and fisically

• $P(f \mid ph) \cdot P(i \mid y) \cdot P(s \mid s) \cdot P(i \mid i) \cdot P(k \mid c) \cdot P(le \mid al) = P(T_i \mid R_i)$
Training Error Model

• Training set \( <s, w> \)
• Align characters of \( s \) and \( w \) based on a min edit distance of single char
• \( a \rightarrow a, c \rightarrow k, \varepsilon \rightarrow g, t \rightarrow s, u \rightarrow u, a \rightarrow a, l \rightarrow l \)
Training Error Model

• Then, for context=N, include N additional adjacent edits for each non-match substitution

• For example for the non-match c → k
  – c → k
  – ac → ak
  – c → kg
  – ac → akg
  – ct → kgs
Training Error Model

• Then calculate the probability of each $\alpha \rightarrow \beta$
  – $\text{count}(\alpha \rightarrow \beta) / \text{count}(\alpha)$
  – $\text{count}(\alpha \rightarrow \beta)$ see last slide
  – $\text{count}(\alpha)$ count $\alpha$ in a representative corpus

• Should see $\text{count}(\alpha \rightarrow \beta)$ go up if it is a frequent substitution
Results

• 10,000 word corpus of common English errors
  – 80% training, 20% evaluation
• Dictionary of 200,000 words
  – Null language model assigns equal P to all words
• Test with and without position info
• Test with and without language model
  – Trigram model from “large online text”
No Position Information

0 is MDE
CG is equivalent to Church and Gale 1991 paper
MDE + look at letter directly to left for weight

<table>
<thead>
<tr>
<th>Max Window</th>
<th>1-Best</th>
<th>2-Best</th>
<th>3-Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>87.0</td>
<td>93.9</td>
<td>95.9</td>
</tr>
<tr>
<td>CG</td>
<td>89.5</td>
<td>94.9</td>
<td>96.5</td>
</tr>
<tr>
<td>1</td>
<td>90.9</td>
<td>95.6</td>
<td>96.8</td>
</tr>
<tr>
<td>2</td>
<td>92.9</td>
<td>97.1</td>
<td>98.1</td>
</tr>
<tr>
<td>3</td>
<td>93.6</td>
<td>97.4</td>
<td>98.5</td>
</tr>
<tr>
<td>4</td>
<td>93.6</td>
<td>97.4</td>
<td>98.5</td>
</tr>
<tr>
<td>Max Window</td>
<td>1-Best</td>
<td>2-Best</td>
<td>3-Best</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>0</td>
<td>88.7</td>
<td>95.1</td>
<td>96.6</td>
</tr>
<tr>
<td>1</td>
<td>92.8</td>
<td>96.5</td>
<td>97.4</td>
</tr>
<tr>
<td>2</td>
<td>94.6</td>
<td>98.0</td>
<td>98.7</td>
</tr>
<tr>
<td>3</td>
<td>95.0</td>
<td>98.0</td>
<td>98.8</td>
</tr>
<tr>
<td>4</td>
<td>95.0</td>
<td>98.0</td>
<td>98.8</td>
</tr>
<tr>
<td>5</td>
<td>95.1</td>
<td>98.0</td>
<td>98.8</td>
</tr>
</tbody>
</table>
Test w/ Language Model
Church and Gale
Conclusions

• Move toward personalized, high accuracy, flexible spell checking
• Adapt to individual or subpopulation
That’s All

• The End