Introduction to

Information Retrieval

Spelling Correction and the Noisy Channel
Spelling Correction and the Noisy Channel

The Spelling Correction Task
Applications for spelling correction

Word processing

Web search

Phones

Showing results for natural language processing
Search instead for natural langage processing
Spelling Tasks

- Spelling Error Detection
- Spelling Error Correction:
  - Autocorrect
    - hte → the
  - Suggest a correction
  - Suggestion lists
Types of spelling errors

- Non-word Errors
  - graffe → giraffe

- Real-word Errors
  - Typographical errors
    - three → there
  - Cognitive Errors (homophones)
    - piece → peace,
    - too → two
Rates of spelling errors

26%: Web queries  *Wang et al. 2003*

13%: Retyping, no backspace:  *Whitelaw et al. English&German*

7%: Words corrected retyping on phone-sized organizer

2%: Words uncorrected on organizer  *Soukoreff &MacKenzie 2003*

1-2%: Retyping:  *Kane and Wobbrock 2007, Gruden et al. 1983*
Non-word spelling errors

- Non-word spelling error detection:
  - Any word not in a *dictionary* is an error
  - The larger the dictionary the better

- Non-word spelling error correction:
  - Generate *candidates*: real words that are similar to error
  - Choose the one which is best:
    - Shortest weighted edit distance
    - Highest noisy channel probability
Real word spelling errors

- For each word $w$, generate candidate set:
  - Find candidate words with similar *pronunciations*
  - Find candidate words with similar *spelling*
  - Include $w$ in candidate set

- Choose best candidate
  - Noisy Channel
  - Classifier
Spelling Correction
and the Noisy
Channel

The Noisy Channel Model of Spelling
Noisy Channel Intuition
Noisy Channel

- We see an observation $x$ of a misspelled word
- Find the correct word $w$

\[
\hat{w} = \arg\max_{w \in V} P(w \mid x)
\]

\[
= \arg\max_{w \in V} \frac{P(x \mid w)P(w)}{P(x)}
\]

\[
= \arg\max_{w \in V} P(x \mid w)P(w)
\]
History: Noisy channel for spelling proposed around 1990

- **IBM**

- **AT&T Bell Labs**
Non-word spelling error example

acress
1. Candidate (V) generation

- Words with similar spelling
  - Small edit distance to error
- Words with similar pronunciation
  - Small edit distance of pronunciation to error
Damerau-Levenshtein edit distance

- Minimal edit distance between two strings, where edits are:
  - Insertion
  - Deletion
  - Substitution
  - Transposition of two adjacent letters
Words within 1 of \textit{acress}

<table>
<thead>
<tr>
<th>Error</th>
<th>Candidate Correction</th>
<th>Correct Letter</th>
<th>Error Letter</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>across</td>
<td>actress</td>
<td>t</td>
<td>ε</td>
<td>deletion</td>
</tr>
<tr>
<td>across</td>
<td>cress</td>
<td>ε</td>
<td>a</td>
<td>insertion</td>
</tr>
<tr>
<td>across</td>
<td>caress</td>
<td>ca</td>
<td>ac</td>
<td>transposition</td>
</tr>
<tr>
<td>across</td>
<td>access</td>
<td>c</td>
<td>r</td>
<td>substitution</td>
</tr>
<tr>
<td>across</td>
<td>across</td>
<td>o</td>
<td>e</td>
<td>substitution</td>
</tr>
<tr>
<td>across</td>
<td>acres</td>
<td>ε</td>
<td>s</td>
<td>insertion</td>
</tr>
<tr>
<td>across</td>
<td>acres</td>
<td>ε</td>
<td>s</td>
<td>insertion</td>
</tr>
</tbody>
</table>
Candidate generation

- 80% of errors are within edit distance 1
- Almost all errors within edit distance 2

- Also allow insertion of **space** or **hyphen**
  - thisidea → this idea
  - inlaw → in-law
2. Prior Probability $P(w)$: Language Model

- Use any of the language modeling algorithms we’ve learned
- Unigram, bigram, trigram
- Web-scale spelling correction
  - Stupid backoff
## Unigram Prior probability

Counts from 404,253,213 words in Corpus of Contemporary English (COCA)

<table>
<thead>
<tr>
<th>word</th>
<th>Frequency of word</th>
<th>$P(\text{word})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>actress</td>
<td>9,321</td>
<td>0.0000230573</td>
</tr>
<tr>
<td>cress</td>
<td>220</td>
<td>0.0000005442</td>
</tr>
<tr>
<td>caress</td>
<td>686</td>
<td>0.0000016969</td>
</tr>
<tr>
<td>access</td>
<td>37,038</td>
<td>0.0000916207</td>
</tr>
<tr>
<td>across</td>
<td>120,844</td>
<td>0.0002989314</td>
</tr>
<tr>
<td>acres</td>
<td>12,874</td>
<td>0.0000318463</td>
</tr>
</tbody>
</table>
3. Likelihood $\Pr(x|w)$: Channel model probability

- **Error model probability, Edit probability**
- *Kernighan, Church, Gale* 1990

- Misspelled word $x = x_1, x_2, x_3, \ldots, x_m$
- Correct word $w = w_1, w_2, w_3, \ldots, w_n$

- $\Pr(x|w) = \text{probability of the edit}$
  - (deletion/insertion/substitution/transposition)

Only consider edit distance = 1 case here
Computing error probability: confusion matrix

\[
\begin{align*}
del[x,y] & : \text{count(xy typed as x)} \\
ins[x,y] & : \text{count(x typed as xy)} \\
sub[x,y] & : \text{count(x typed as y)} \\
trans[x,y] & : \text{count(xy typed as yx)}
\end{align*}
\]

Insertion and deletion conditioned on previous character
Confusion matrix $M$

$$\hat{P}(w \mid c) = \frac{M[w, c]}{\sum_{w'} M[w', c]}$$

**Substitution of X (incorrect) for Y (correct)**

<table>
<thead>
<tr>
<th>X</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
<th>m</th>
<th>n</th>
<th>o</th>
<th>p</th>
<th>q</th>
<th>r</th>
<th>s</th>
<th>t</th>
<th>u</th>
<th>v</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>342</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>118</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>39</td>
<td>40</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
<td>10</td>
<td>13</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>43</td>
<td>30</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>388</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>93</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>12</td>
<td>6</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>13</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>14</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>103</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>146</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>j</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>11</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>91</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>116</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>30</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>20</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>12</td>
<td>22</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>11</td>
<td>8</td>
<td>27</td>
<td>33</td>
<td>35</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>42</td>
<td>7</td>
<td>5</td>
<td>19</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>19</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>36</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
Generating the confusion matrix

- Peter Norvig’s list of errors
- Peter Norvig’s list of counts of single-edit errors

https://norvig.com/ngrams/
Channel model

\[ P(x|w) = \begin{cases} 
\frac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1}w_i]}, & \text{if deletion} \\
\frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\
\frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\
\frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_iw_{i+1}]}, & \text{if transposition}
\end{cases} \]

Kernighan, Church, Gale 1990
# Channel model for *across*

| Candidate Word | Correct Letter | Error Letter | x|w | P(x|word) |
|----------------|----------------|--------------|-----|----------|
| actress        | t              | -            | c|ct | .000117  |
| cress          | -              | a            | a|# | .00000144|
| caress         | ca             | ac           | ac|ca | .00000164|
| access         | c              | r            | r|c  | .000000209|
| across         | o              | e            | e|o  | .0000093 |
| acres          | -              | s            | es|e  | .0000321 |
| acres          | -              | s            | ss|s  | .0000342 |
Noisy channel probability for *acress*

| Candidate Word | Correct Letter | Error Letter | x|w | P(x|word) | P(word) | 10^9*P(x|w)P(w) |
|----------------|----------------|--------------|----------------|-----------|----------|----------------|
| actress        | t              | -            | c|ct | .000117  | .0000231 | 2.7           |
| cress          | -              | a            | a|#  | .00000144 | .000000544 | .00078        |
| caress         | ca             | ac           | ac|ca | .00000164 | .00000170 | .0028         |
| access         | c              | r            | r|c  | .000000209 | .0000916   | .019          |
| across         | o              | e            | e|o  | .0000093  | .000299   | 2.8           |
| acres          | -              | s            | es|e  | .0000321  | .0000318   | 1.0           |
| acres          | -              | s            | ss|s  | .0000342  | .0000318   | 1.0           |
Noisy channel probability for \textit{acress}

| Candidate Word | Correct Letter | Error Letter | $x|w$       | $P(x|\text{word})$ | $P(\text{word})$ | $10^9 \cdot P(x|w)P(w)$ |
|----------------|----------------|--------------|-------------|---------------------|------------------|------------------------|
| actress       | t              | -            | c|ct              | .000117            | .0000231          | 2.7                    |
| cress         | -              | a            | a|#               | .00000144          | .000000544        | .00078                 |
| caress        | ca             | ac           | ac|ca              | .00000164          | .00000170         | .0028                  |
| access        | c              | r            | r|c               | .000000209         | .00000916         | .019                   |
| across        | o              | e            | e|o               | .0000093           | .000299           | 2.8                    |
| acres         | -              | s            | es|e              | .0000321           | .0000318          | 1.0                    |
| acres         | -              | s            | ss|s              | .0000342           | .0000318          | 1.0                    |
Using a bigram language model

- “a stellar and versatile \textit{across} whose combination of sass and glamour…”

- Counts from the Corpus of Contemporary American English with add-1 smoothing

- \[ P(\text{actress} | \text{versatile}) = 0.000021 \]
- \[ P(\text{whose} | \text{actress}) = 0.0010 \]
- \[ P(\text{across} | \text{versatile}) = 0.000021 \]
- \[ P(\text{whose} | \text{across}) = 0.000006 \]

- \[ P(“\text{versatile actress whose}”) = 0.000021 \times 0.0010 = 210 \times 10^{-10} \]
- \[ P(“\text{versatile across whose}”) = 0.000021 \times 0.000006 = 1 \times 10^{-10} \]
Using a bigram language model

"a stellar and versatile across whose combination of sass and glamour..."

Counts from the Corpus of Contemporary American English with add-1 smoothing

- \( P(\text{actress}|\text{versatile}) = .000021 \)
- \( P(\text{whose}|\text{actress}) = .0010 \)
- \( P(\text{across}|\text{versatile}) = .000021 \)
- \( P(\text{whose}|\text{across}) = .000006 \)

- \( P(\text{"versatile actress whose"}) = .000021 \times .0010 = 210 \times 10^{-10} \)
- \( P(\text{"versatile across whose"}) = .000021 \times .000006 = 1 \times 10^{-10} \)
Evaluation

- Some spelling error test sets
  - Wikipedia’s list of common English misspelling
  - Aspell filtered version of that list
  - Birkbeck spelling error corpus
  - Peter Norvig’s list of errors (includes Wikipedia and Birkbeck, for training or testing)
Spelling Correction and the Noisy Channel

Real-Word Spelling Correction
Real-word spelling errors

- ...leaving in about fifteen minutes to go to her house.
- The design *an* construction of the system...
- Can they *lave* him my messages?
- The study was conducted mainly *be* John Black.

- 25-40% of spelling errors are real words  
  Kukich 1992
Solving real-world spelling errors

- For each word in sentence
  - Generate candidate set
    - the word itself
    - all single-letter edits that are English words
    - words that are homophones

- Choose best candidates
  - Noisy channel model
  - Task-specific classifier
Noisy channel for real-word spell correction

- Given a sentence $w_1, w_2, w_3, ..., w_n$
- Generate a set of candidates for each word $w_i$
  - Candidate($w_1$) = \{w_1, w'_1, w''_1, w'''_1, ...\}
  - Candidate($w_2$) = \{w_2, w'_2, w''_2, w'''_2, ...\}
  - Candidate($w_n$) = \{w_n, w'_n, w''_n, w'''_n, ...\}
- Choose the sequence $W$ that maximizes $P(W)$
Noisy channel for real-word spell correction
Noisy channel for real-word spell correction
Simplification: One error per sentence

- Out of all possible sentences with one word replaced:
  - \( w_1, w''_2, w_3, w_4 \) two off thew
  - \( w_1, w_2, w'_3, w_4 \) two of the
  - \( w''''_1, w_2, w_3, w_4 \) too of thew
  - ...

- Choose the sequence \( W \) that maximizes \( P(W) \)
Where to get the probabilities

- Language model
  - Unigram
  - Bigram
  - Etc

- Channel model
  - Same as for non-word spelling correction
  - Plus need probability for no error, \( P(w|w) \)
Probability of no error

- What is the channel probability for a correctly typed word?
- $P(\text{“the”} \mid \text{“the”})$

- Obviously this depends on the application
  - .90 (1 error in 10 words)
  - .95 (1 error in 20 words)
  - .99 (1 error in 100 words)
  - .995 (1 error in 200 words)
## Peter Norvig’s “thew” example

| x   | w    | x|w | P(x|w)   | P(w)       | 10^9 P(x|w)P(w) |
|-----|------|----|-----|---------|------------|-----------------|
| thew | the  | ew|e  | 0.000007 | 0.02       | 144             |
| thew | thew |    |    | 0.95     | 0.00000009 | 90              |
| thew | thaw  | e| a | 0.001    | 0.0000007 | 0.7             |
| thew | threw | h| hr | 0.000008 | 0.000004  | 0.03            |
| thew | thwe  | ew|we | 0.000003 | 0.0000004 | 0.0001          |