The Spelling Correction Task
Applications for spelling correction

Word processing

Spelling and Grammar: English (US)

Not in dictionary:

Spell checking is a component of

Suggestions:

- component

Add

Change

Change All

Ignore

Ignore All

AutoCorrect

Web search

ploogle

natural language processing

Showing results for natural language processing
Search instead for natural language processing

Phones

New iMessage

To: Dan Jurafsky

Oops, running late

Q W E R T Y U I O P

A S D F G H J K L

Z X C V B N M

space

return
Spelling Tasks

• Spelling Error Detection

• Spelling Error Correction:
  • Autocorrect
    • hte ➔ the
  • Suggest a correction
  • Suggestion lists
Types of spelling errors

• Non-word Errors
  • *graffe* $\rightarrow$ *giraffe*

• Real-word Errors
  • Typographical errors
    • *three* $\rightarrow$ *there*
  • Cognitive Errors (homophones)
    • *piece* $\rightarrow$ *peace*,
    • *too* $\rightarrow$ *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 Spelling Correction Task
Spelling Correction and the Noisy Channel

The Noisy Channel Model of Spelling
Noisy Channel Intuition

original word

noisy channel

noisy word
Noisy Channel

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

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

- **IBM**

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

acress
Candidate generation

- Words with similar spelling
  - Small edit distance to error
- Words with similar pronunciation
  - Small edit distance of pronunciation to error
Dan Jurafsky

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 across

<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>–</td>
<td>deletion</td>
</tr>
<tr>
<td>across</td>
<td>cress</td>
<td>–</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>–</td>
<td>s</td>
<td>insertion</td>
</tr>
<tr>
<td>across</td>
<td>acres</td>
<td>–</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`
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(word)</th>
</tr>
</thead>
<tbody>
<tr>
<td>actress</td>
<td>9,321</td>
<td>0.000230573</td>
</tr>
<tr>
<td>cress</td>
<td>220</td>
<td>0.000005442</td>
</tr>
<tr>
<td>caress</td>
<td>686</td>
<td>0.000016969</td>
</tr>
<tr>
<td>access</td>
<td>37,038</td>
<td>0.000916207</td>
</tr>
<tr>
<td>across</td>
<td>120,844</td>
<td>0.002989314</td>
</tr>
<tr>
<td>acres</td>
<td>12,874</td>
<td>0.000318463</td>
</tr>
</tbody>
</table>
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 \( w = w_1, w_2, w_3, \ldots, w_n \)

- \( P(x|w) = \text{probability of the edit} \)
  - (deletion/insertion/substitution/transposition)
Computing error probability: confusion matrix

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

Insertion and deletion conditioned on previous character
Confusion matrix for spelling errors

\[ \text{sub}[X, Y] = \text{Substitution of } X \text{ (incorrect) for } Y \text{ (correct)} \]

<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>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</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>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>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>6</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>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>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>
<td>0</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>0</td>
<td>1</td>
<td>0</td>
<td>18</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>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g</td>
<td>4</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</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>0</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>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>0</td>
<td>0</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>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>
</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>2</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></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>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</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>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>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>4</td>
<td>4</td>
<td>0</td>
<td>180</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>15</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</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>4</td>
<td>35</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>28</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>o</td>
<td>91</td>
<td>1</td>
<td>13</td>
<td>116</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>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</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>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>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>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>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>3</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>
</tr>
<tr>
<td>s</td>
<td>11</td>
<td>8</td>
<td>27</td>
<td>37</td>
<td>35</td>
<td>4</td>
<td>0</td>
<td>1</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>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>1</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>6</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>0</td>
<td>64</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>0</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>0</td>
<td>7</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>1</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>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</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>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</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>1</td>
<td>0</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>0</td>
<td>7</td>
<td>5</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>0</td>
<td>3</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
Channel model

Kernighan, Church, Gale 1990

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

| Candidate Correction | Correct Letter | Error Letter | x|w | P(x|word) |
|----------------------|----------------|--------------|-----------------|--------|
| actress              | t              | -            | c | ct | 0.000117 |
| cress                | -              | a            | a | #  | 0.00000144 |
| caress               | ca             | ac           | ac | ca | 0.00000164 |
| access               | c              | r            | r | c  | 0.000000209 |
| across               | o              | e            | e | o  | 0.0000093  |
| acres                | -              | s            | es | e  | 0.0000321  |
| acres                | -              | s            | ss | s  | 0.0000342  |
Noisy channel probability for across

| Candidate Correction | 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 acres

| Candidate Correction | 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|# | .0000144 | .00000544 | .00078         |
| caress               | ca             | ac           | ac|ca | .0000164 | .0000170 | .0028          |
| access               | c              | r            | r|c  | .0000209 | .0000916 | .019           |
| across               | o              | e            | e|o  | .000093  | .000299  | 2.8            |
| acres                | –              | s            | es|e  | .000321  | .000318  | 1.0            |
| acres                | –              | s            | ss|s  | .000342  | .000318  | 1.0            |
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
  • \( P(\text{actress}|\text{versatile}) = 0.00021 \)
  • \( P(\text{whose}|\text{actress}) = 0.0010 \)
  • \( P(\text{across}|\text{versatile}) = 0.00021 \)
  • \( P(\text{whose}|\text{across}) = 0.000006 \)

\[
\begin{align*}
P(\text{"versatile actress whose"}) &= 0.00021 \times 0.0010 = 210 \times 10^{-10} \\
P(\text{"versatile across whose"}) &= 0.00021 \times 0.000006 = 1 \times 10^{-10}
\end{align*}
\]
Using a bigram language model

- “a stellar and versatile actress whose combination of sass and glamour…”
- Counts from the Corpus of Contemporary American English with add-1 smoothing
  - \( P(\text{actress}|\text{versatile}) = 0.00021 \)
  - \( P(\text{whose}|\text{actress}) = 0.0010 \)
  - \( P(\text{across}|\text{versatile}) = 0.00021 \)
  - \( P(\text{whose}|\text{across}) = 0.000006 \)

- \( P(\text{“versatile actress whose”}) = 0.00021 \times 0.0010 = 210 \times 10^{-10} \)
- \( P(\text{“versatile across whose”}) = 0.00021 \times 0.000006 = 1 \times 10^{-10} \)
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 Model of Spelling
Spelling Correction and the Noisy Channel

Real-Word Spelling Correction
Real-word spelling errors

- ...leaving in about fifteen minuets 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 $w_1, w_2, w_3, \ldots, w_n$
- Generate a set of candidates for each word $w_i$
  - $\text{Candidate}(w_1) = \{w_1, w'_1, w''_1, w'''_1, \ldots\}$
  - $\text{Candidate}(w_2) = \{w_2, w'_2, w''_2, w'''_2, \ldots\}$
  - $\text{Candidate}(w_n) = \{w_n, w'_n, w''_n, w'''_n, \ldots\}$
- Choose the sequence $W$ that maximizes $P(W)$
Noisy channel for real-word spell correction
Noisy channel for real-word spell correction

two  of  thew  ...

to  threw  thaw

tao  off  thaw

too  on  the

two  of  thaw
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 “thaw” 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 | 0.00000004 | 0.0001          |
Spelling Correction and the Noisy Channel

Real-Word Spelling Correction
Spelling Correction and the Noisy Channel
**HCI issues in spelling**

- If very confident in correction
  - Autocorrect
- Less confident
  - Give the best correction
- Less confident
  - Give a correction list
- Unconfident
  - Just flag as an error
State of the art noisy channel

• We never just multiply the prior and the error model
• Independence assumptions $\rightarrow$ probabilities not commensurate
• Instead: Weigh them

$$\hat{w} = \arg\max_{w \in V} P(x \mid w)P(w)^\lambda$$

• Learn $\lambda$ from a development test set
Phonetic error model

- Metaphone, used in GNU aspell
  - Convert misspelling to metaphone pronunciation
    - “Drop duplicate adjacent letters, except for C.”
    - “If the word begins with 'KN', 'GN', 'PN', 'AE', 'WR', drop the first letter.”
    - “Drop 'B' if after 'M' and if it is at the end of the word”
    - ...
  - Find words whose pronunciation is 1-2 edit distance from misspelling’s
  - Score result list
    - Weighted edit distance of candidate to misspelling
    - Edit distance of candidate pronunciation to misspelling pronunciation
Improvements to channel model

• Allow richer edits  (Brill and Moore 2000)
  • ent → ant
  • ph → f
  • le → al

• Incorporate pronunciation into channel  (Toutanova and Moore 2002)
Channel model

- Factors that could influence $p(\text{misspelling} | \text{word})$
  - The source letter
  - The target letter
  - Surrounding letters
  - The position in the word
  - Nearby keys on the keyboard
  - Homology on the keyboard
  - Pronunciations
  - Likely morpheme transformations
Nearby keys
Classifier-based methods for real-word spelling correction

- Instead of just channel model and language model
- Use many features in a classifier (next lecture).
- Build a classifier for a specific pair like:
  
  **whether/weather**
  
  - “cloudy” within +- 10 words
  - ___ to VERB
  - ___ or not
Spelling Correction and the Noisy Channel

Real-Word Spelling Correction