Unsupervised Tagging?

- AKA part-of-speech induction
- Task:
  - Raw sentences in
  - Tagged sentences out
- Obvious thing to do:
  - Start with a (mostly) uniform HMM
  - Run EM
  - Inspect results

Forward Recurrence

Backward Recurrence

Fractional Transitions

EM for HMMs: Quantities

- Cache total path values:
  \[ \alpha_i(s) = P(w_0 \ldots w_i | s_i) = \sum_{s_{i-1}} P(s_{i-1} | s_i) P(w_i | s_i) \alpha_{i-1}(s_{i-1}), \]
  \[ \beta_i(s) = P(w_{i+1} \ldots w_n | s_i) = \sum_{s_{i+1}} P(s_{i+1} | s_i) P(w_{i+1} | s_{i+1}) \beta_{i+1}(s_{i+1}). \]
- Can calculate in \(O(s^2 n)\) time (why?)
EM for HMMs: Process

- From these quantities, we can re-estimate transitions:
\[
\text{count}(s \rightarrow s') = \frac{\sum_{i} \alpha_i(s) P(s'|s) P(w_i|s) \beta_{i+1}(s')}{P(w)}
\]
- And emissions:
\[
\text{count}(w_i, s) = \frac{\sum_{i} \alpha_i(s) \beta_{i+1}(s)}{P(w)}
\]
- If you don’t get these formulas immediately, just think about hard EM instead, where we re-estimate from the Viterbi sequences.

Merialdo: Setup

- Some (discouraging) experiments [Merialdo 94]
- Setup:
  - You know the set of allowable tags for each word
  - Fix k training examples to their true labels
  - Learn P(w|t) on these examples
  - Learn P(t|t-1, t-2) on these examples
  - On n examples, re-estimate with EM
- Note: we know allowed tags but not frequencies

Merialdo: Results

<table>
<thead>
<tr>
<th>Number of tagged sentences used for (the initial model)</th>
<th>0</th>
<th>100</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
<th>20000</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct tags (%) words after ML on LM words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>68.5</td>
<td>90.0</td>
<td>95.4</td>
<td>96.2</td>
<td>96.0</td>
<td>97.0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>85.5</td>
<td>93.6</td>
<td>95.8</td>
<td>96.3</td>
<td>96.2</td>
<td>96.9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>81.8</td>
<td>93.0</td>
<td>95.7</td>
<td>96.3</td>
<td>96.4</td>
<td>96.8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>83.0</td>
<td>93.1</td>
<td>95.6</td>
<td>96.4</td>
<td>96.2</td>
<td>96.8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>84.0</td>
<td>93.0</td>
<td>95.2</td>
<td>95.3</td>
<td>95.8</td>
<td>96.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>84.8</td>
<td>92.9</td>
<td>95.1</td>
<td>94.8</td>
<td>95.6</td>
<td>95.8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>85.3</td>
<td>92.8</td>
<td>94.9</td>
<td>95.2</td>
<td>95.5</td>
<td>95.7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>85.8</td>
<td>92.8</td>
<td>94.7</td>
<td>95.1</td>
<td>95.3</td>
<td>95.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>86.1</td>
<td>92.7</td>
<td>94.6</td>
<td>95.0</td>
<td>95.2</td>
<td>95.4</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>86.3</td>
<td>92.6</td>
<td>94.5</td>
<td>94.9</td>
<td>95.1</td>
<td>95.3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>86.6</td>
<td>92.6</td>
<td>94.4</td>
<td>94.8</td>
<td>95.0</td>
<td>95.2</td>
<td></td>
</tr>
</tbody>
</table>

Distributional Clustering

- Three main variants on the same idea:
  - Pairwise similarities and heuristic clustering
    - E.g. [Finch and Chater 92]
  - Produces dendrograms
  - Vector space methods
    - E.g. [Shuetze 93]
  - Models of ambiguity
  - Probabilistic methods
    - Various formulations, e.g. [Lee and Pereira 99]

Distributional Clustering

- Nearest Neighbors
Dendrograms

Vector Space Version

- [Shuetze 93] clusters words as points in $\mathbb{R}^n$
- Context counts

A Probabilistic Version?

$$P(S, C) = \prod_i P(c_j) P(w_j | c_j) P(w_{j+1}, w_{j+2} | c_j)$$

What Else?

- Various newer ideas:
  - Context distributional clustering [Clark 00]
  - Morphology-driven models [Clark 03]
  - Contrastive estimation [Smith and Eisner 05]

- Also:
  - What about ambiguous words?
  - Using wider context signatures has been used for learning synonyms (what’s wrong with this approach?)
  - Can extend these ideas for grammar induction (later)
**Places of articulation**

- **Labial place**
  - Bilabial: p, b, m
  - Labiodental: f, v

- **Coronal place**
  - Dental: th/dh
  - Alveolar: t/d/s/z/l
  - Post: sh/zh/y

- **Dorsal Place**
  - Velar: k/g/ng

**Manner of Articulation**

- **Stop**: complete closure of articulators, so no air escapes through mouth
- **Oral stop**: palate is raised, no air escapes through nose. Air pressure builds up behind closure, explodes when released
  - p, t, k, b, d, g
- **Nasal stop**: oral closure, but palate is lowered, air escapes through nose.
  - m, n, ng

**Oral vs. Nasal Sounds**

Thanks to Jong-huk Kim for this figure!
Vowels

Simple Periodic Waves

- Characterized by:
  - period: T
  - amplitude A
  - phase ϕ
- Fundamental frequency in cycles per second, or Hz
  - $F_0 = 1/T$

Simple periodic waves of sound

- Y axis: Amplitude = amount of air pressure at that point in time
- Zero is normal air pressure, negative is rarefaction
- X axis: time. Frequency = number of cycles per second.
- Frequency = 1/Period
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex waves: 100Hz+1000Hz

Spectrum

- Frequency components (100 and 1000 Hz) on x-axis

Spectrum of an actual soundwave
Waveforms for speech

- Waveform of the vowel [iy]

- Frequency: repetitions/second of a wave
- Above vowel has 28 reps in .11 secs
- So freq is 28/.11 = 255 Hz
- This is speed that vocal folds move, hence voicing
- Amplitude: y axis: amount of air pressure at that point in time
- Zero is normal air pressure, negative is rarefaction

She just had a baby

- Vowels are voiced, long, loud
- Length in time = length in space in waveform picture
- Voicing: regular peaks in amplitude
- When stops closed: no peaks: silence.
- Peaks = voicing: .46 to .58 (vowel [iy]), from second .65 to .74 (vowel [ax]) and so on
- Silence of stop closure (1.06 to 1.08 for first [b], or 1.26 to 1.28 for second [b])
- Fricatives like [sh] intense irregular pattern: see .33 to .46

Examples from Ladefoged

- [pad]
- [bad]
- [spat]

Part of [ae] waveform from “had”

- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

Back to Spectra

- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.

- X-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

Why these Peaks?

- Articulatory facts:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others
Resonances of the vocal tract

- The human vocal tract as an open tube
- Air in a tube of a given length will tend to vibrate at resonance frequency of the tube.
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

Length 17.5 cm.

Computing the 3 Formants of Schwa

- Let the length of the tube be L
  - $F_1 = \frac{c}{2L} = \frac{35,000}{4 \times 17.5} = 500$ Hz
  - $F_2 = \frac{3c}{4L} = \frac{3 \times 35,000}{4 \times 17.5} = 1500$ Hz
  - $F_3 = \frac{5c}{4L} = \frac{5 \times 35,000}{4 \times 17.5} = 2500$ Hz

- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called formants

Seeing formants: the spectrogram

American English Vowel Space

Figure from W. Barry Speech Science slides
Figure from Sundberg
Figure from Jennifer Venditti
Dialect Issues

- Speech varies from dialect to dialect (examples are American vs. British English)
  - Syntactic ("I could" vs. "I could do")
  - Lexical ("elevator" vs. "lift")
  - Phonological (butter: [bʌ:tə] vs. [bʌ:ter])
  - Phonetic

- Mismatch between training and testing dialects can cause a large increase in error rate

How to read spectrograms

- bab: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- dad: first formant increases, but F2 and F3 slight fall
- gag: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

She came back and started again

- lots of high-freq energy
- closure for k
- burst of aspiration for k
- ey vowel; faint 1100 Hz formant is nasalization
- bilabial nasal
- short b closure, voicing barely visible.
- ae; note upward transitions after bilabial stop at beginning
- note F2 and F3 coming together for "k"