MaxEnt taggers

• These models all potentially suffer from the *label bias problem*

• All probability going into state is passed on to successors (since $\sum_s P(s|o, s') = 1$)

• Also, states with fewer outgoing transitions will be preferred to those with more

• One solution: Conditional Random Fields (Lafferty et al. 2001) assign a probability to an entire sequence at one shot:

$$P(t_1, \ldots, t_n|w_1, \ldots, w_n) = \frac{1}{Z} \exp \left( \lambda_j \sum_j f_j(w_1, \ldots, w_n, t_1, \ldots, t_n) \right)$$
Perceptron

- Another solution: discriminative training (Collins 2002) using a variant of the ‘perceptron’ algorithm

- Early machine learning algorithm (Rosenblatt 1958) based on a model neuron

- Iteratively adjust decision boundary to correct errors

- If classes are linearly separable, perceptron algorithm will find an optimal solution

- Many modern descendants: neural nets, bagging, boosting, support vector machines
Parts of speech

- More fundamental problem: what is a part of speech?

- Traditional view of parts of speech: around eight universal, non-overlapping categories

- This is too simple, but we can classify words as:
  - Open class: nouns, verbs, adjectives, adverbs
  - Closed class: prepositions, determiners, conjunctions, particles, etc.

- Some categories overlap (e.g., pronouns are closed class nouns)

- Some words (*ago, enough*) are members of their own category
### Parts of speech

<table>
<thead>
<tr>
<th></th>
<th>CLAWS5</th>
<th>Brown</th>
<th>Penn</th>
<th>ICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>PNP</td>
<td>PPS</td>
<td>PRP</td>
<td>PRON(pers,sing)</td>
</tr>
<tr>
<td>was</td>
<td>VBD</td>
<td>BEDZ</td>
<td>VBD</td>
<td>AUX(pass,past)</td>
</tr>
<tr>
<td>told</td>
<td>VVN</td>
<td>VBN</td>
<td>VBN</td>
<td>V(ditr,edp)</td>
</tr>
<tr>
<td>that</td>
<td>CJT</td>
<td>CS</td>
<td>IN</td>
<td>CONJUNC(subord)</td>
</tr>
<tr>
<td>the</td>
<td>AT0</td>
<td>AT</td>
<td>DT</td>
<td>ART(def)</td>
</tr>
<tr>
<td>journey</td>
<td>NN1</td>
<td>NN</td>
<td>NN</td>
<td>N(com,sing)</td>
</tr>
<tr>
<td>might</td>
<td>VM0</td>
<td>MD</td>
<td>MD</td>
<td>AUX(modal,past)</td>
</tr>
<tr>
<td>kill</td>
<td>VVI</td>
<td>VB</td>
<td>VB</td>
<td>V(montr,infin)</td>
</tr>
<tr>
<td>her</td>
<td>PNP</td>
<td>PPO</td>
<td>PRP</td>
<td>PRON(poss,sing)</td>
</tr>
<tr>
<td>.</td>
<td>PUN</td>
<td>.</td>
<td>.</td>
<td>PUNC(per)</td>
</tr>
</tbody>
</table>
Parts of speech

• Tag sets differ greatly in the number and kind of distinctions they make.

  Brown      179
  Penn treebank  45
  CLAWS1       132
  CLAWS2       166
  CLAWS5       65
  London-Lund  197

• Tag sets are language and application dependent.
Some distinctions cannot be made reliably:

Pat worked through the problem.
Pat worked the problem through.
Pat walked through the door.
*Pat walked the door through.

More detailed tagsets can, paradoxically, be easier to apply

- auxiliaries (be, do, have)
- modals (can, will)
- gerunds (Pat’s constantly humming showtunes is annoying.)
Syntax

- An even deeper problem: part of speech tags are usually not what we care about (Streetlight Principle)

- Tagging techniques can be extended to capture more syntactic structure (chunking)

- We can only get so far without recognizing hierarchical structures:
  
  The velocity of the seismic waves rises to . . .

- We can more accurately capture the structure of natural language by moving from regular grammars (HMMs) to context-free grammars