Part of Speech (POS) Tagging

• Annotate each word in a sentence with its POS:
  – noun, verb, adjective, adverb, pronoun, preposition, interjection, ...

  PRP  VBD  TO  VB  TO  DT  NN  IN  NN  VBD  VBG

They used to object to the use of object oriented programming

  obJECT  OBject

• Useful for many other NLP tasks:
  – speech recognition and synthesis
  – syntactic parsing
  – word sense disambiguation
  – information retrieval, …
Parts of Speech

• Lexical categories that are defined based on:
  – **Syntactic function**:
    • nouns can occur with determiners: a *goat*.
    • nouns can take possessives: IBM’s annual revenue.
    • most nouns can occur in the plural: *goats*.
  – **Morphological function**:
    • many verbs can be composed with the prefix “un”.

• There are tendencies toward **semantic coherence**:
  – nouns often refer to “people, places, or things”.
  – adjectives often refer to properties.
POS: Closed Class vs. Open Class

• **Closed Class:**
  – relatively fixed membership.
  – usually *function words*:
    • short common words which have a structuring role in grammar.
  – **Prepositions**: of, in, by, on, under, over, …
  – **Auxiliaries**: may, can, will, had, been, should, …
  – **Pronouns**: I, you, she, mine, his, them, …
  – **Determiners**: a, an, the, which, that, …
  – **Conjunctions**: and, but, or (coord.), as, if, when, (subord.), …
  – **Particles**: up, down, on, off, …
  – **Numerals**: one, two, three, third, …
POS: Open Class vs. Closed Class

• **Open Class:**
  – new members are continually added.
    • *to fax, to google, futon, …*
  – English has 4: **Nouns, Verbs, Adjectives, Adverbs.**
    • Many languages have these 4, but not all (e.g. Korean).
  – **Nouns**: people, places, or things
  – **Verbs**: actions and processes
  – **Adjectives**: properties or qualities
  – **Adverbs**: a hodge-podge
    • *Unfortunately, John walked home extremely slowly yesterday.*
    • directional, locative, temporal, degree, manner, …
POS: Open vs. Closed Classes

- **Open Class:**
  - new members are continually added.

1. Love is too weak a word for what I feel... I *lurve* you. Y'know, I *loove* you, I, I *luff* you. There are two f's. I have to invent... Of course I love you.

   *(Woody Allen in *Annie Hall*, Woody Allen)*

2. 'Twas brillig, and the slithy toves
   Did gyre and gimble in the wabe;
   All mimsy were the borogoves,
   And the mome raths outgrabe.

   "Beware the Jabberwock, my son!
   The jaws that bite, the claws that catch!
   Beware the Jubjub bird, and shun
   The frumious Bandersnatch!"

   *(Jabberwocky, Lewis Carroll)*
Parts of Speech: Granularity

- Grammatical sketch of Greek [Dionysius Thrax, c. 100 B.C.]:
  - 8 tags: noun, verb, pronoun, preposition, adjective, conjunction, participle, and article.

- Brown corpus [Francis, 1979]:
  - 87 tags.

- Penn Treebank [Marcus et al., 1993]:
  - 45 tags.

- British National Corpus (BNC) [Garside et al., 1997]:
  - C5 tagset: 61 tags.
  - C7 tagset: 146 tags.

We will focus on the Penn Treebank POS tags.
# Penn Treebank POS Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordin. conjunction</td>
<td><em>and, but, or</em></td>
<td>SYM</td>
<td>symbol</td>
<td><em>+,%,&amp;</em></td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td><em>one, two, three</em></td>
<td>TO</td>
<td>“to”</td>
<td><em>to</em></td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td><em>a, the</em></td>
<td>UH</td>
<td>interjection</td>
<td><em>ah, oops</em></td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td><em>there</em></td>
<td>VB</td>
<td>verb, base form</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td><em>mea culpa</em></td>
<td>VBD</td>
<td>verb, past tense</td>
<td><em>ate</em></td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td><em>of, in, by</em></td>
<td>VBG</td>
<td>verb, gerund</td>
<td><em>eating</em></td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td><em>yellow</em></td>
<td>VBN</td>
<td>verb, past participle</td>
<td><em>eaten</em></td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td><em>bigger</em></td>
<td>VBP</td>
<td>verb, non-3sg pres</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td><em>wildest</em></td>
<td>VBZ</td>
<td>verb, 3sg pres</td>
<td><em>eats</em></td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td><em>1, 2, One</em></td>
<td>WDT</td>
<td>wh-determiner</td>
<td><em>which, that</em></td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td><em>can, should</em></td>
<td>WP</td>
<td>wh-pronoun</td>
<td><em>what, who</em></td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing. or mass</td>
<td><em>llama</em></td>
<td>WP$</td>
<td>possessive wh-</td>
<td><em>whose</em></td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td><em>llamas</em></td>
<td>WRB</td>
<td>wh-adverb</td>
<td><em>how, where</em></td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td><em>IBM</em></td>
<td>$</td>
<td>dollar sign</td>
<td><em>$</em></td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td><em>Carolinans</em></td>
<td>#</td>
<td>pound sign</td>
<td><em>#</em></td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td><em>all, both</em></td>
<td>“</td>
<td>left quote</td>
<td>‘ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td><em>’s</em></td>
<td>”</td>
<td>right quote</td>
<td>’ or ”</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td><em>I, you, he</em></td>
<td>(</td>
<td>left parenthesis</td>
<td>[, (, {, &lt;</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
<td><em>your, one’s</em></td>
<td>)</td>
<td>right parenthesis</td>
<td>], ), }, &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td><em>quickly, never</em></td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td><em>faster</em></td>
<td>.</td>
<td>sentence-final punc</td>
<td>. ! ?</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td><em>fastest</em></td>
<td>:</td>
<td>mid-sentence punc</td>
<td>: ; ... --</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td><em>up, off</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Penn Treebank POS tags

• Selected from the original 87 tags of the Brown corpus:
  ⇒ lost finer distinctions between lexical categories.

1) Prepositions and subordinating conjunctions:
   – after/CS spending/VBG a/AT day/NN at/IN the/AT palace/NN
   – after/IN a/AT wedding/NN trip/NN to/IN Hawaii/NNP ./.

2) Infinitive to and prepositional to:
   – to/TO give/VB priority/NN to/IN teachers/NNS

3) Adverbial nouns:
   – Brown: Monday/NR, home/NR, west/NR, tomorrow/NR
   – PTB: Monday/NNP, (home, tomorrow, west)/(NN, RB)
POS Tagging ≡ POS Disambiguation

• Words often have more than one POS tag, e.g. back:
  – the back/JJ door
  – on my back/NN
  – win the voters back/RB
  – promised to back/VB the bill

• Brown corpus statistics [DeRose, 1998]:
  – 11.5% ambiguous English word types.
  – 40% of all word occurrences are ambiguous.
    • most are easy to disambiguate
      – the tags are not equaly likely, i.e. low tag entropy: table
## POS Tag Ambiguity

<table>
<thead>
<tr>
<th></th>
<th>87-tag Original Brown</th>
<th>45-tag Treebank Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unambiguous (1 tag)</strong></td>
<td>44,019</td>
<td>38,857</td>
</tr>
<tr>
<td><strong>Ambiguous (2–7 tags)</strong></td>
<td>5,490</td>
<td>8844</td>
</tr>
<tr>
<td>Details:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 tags</td>
<td>4,967</td>
<td>6,731</td>
</tr>
<tr>
<td>3 tags</td>
<td>411</td>
<td>1621</td>
</tr>
<tr>
<td>4 tags</td>
<td>91</td>
<td>357</td>
</tr>
<tr>
<td>5 tags</td>
<td>17</td>
<td>90</td>
</tr>
<tr>
<td>6 tags</td>
<td>2 (well, beat)</td>
<td>32</td>
</tr>
<tr>
<td>7 tags</td>
<td>2 (still, down)</td>
<td></td>
</tr>
<tr>
<td>8 tags</td>
<td></td>
<td>4 (’s, half, back, a)</td>
</tr>
<tr>
<td>9 tags</td>
<td></td>
<td>3 (that, more, in)</td>
</tr>
</tbody>
</table>
POS Disambiguation: Context

“Here's a movie where you forgive the **preposterous** because it takes you to the **perplexing**.”

*[Source Code, by Roger Ebert, March 31, 2011]*

“The **good**, the **bad**, and the **ugly**”

“The **young** and the **restless**”

“The **bold** and the **beautiful**”
POS Tagging \equiv POS Disambiguation

- Some distinctions are difficult even for humans:
  - Mrs. Shaefer never got \textit{around} to joining
    \texttt{NNP NNP RB VBD RP TO VBG}
  - All we gotta do is go \textit{around} the corner
    \texttt{DT PRP VBN VB VBZ VB IN DT NN}
  - Chateau Petrus costs \textit{around} 250
    \texttt{NNP NNP VBZ RB CD}

- Use heuristics [Santorini, 1990]:
  - She told \texttt{off/RP} her friends
  - She told her friends \texttt{off/RP}
    \texttt{off/RP}
  - She stepped \texttt{off/IN} the train
    \texttt{off/IN}
  - \*She stepped the train \texttt{off/IN}
How Difficult is POS Tagging?

• Most current tagging algorithms: $\sim$ 96% - 97% accuracy for Penn Treebank tagsets.
  – Current SofA 97.55% tagging accuracy. How good is this?
    • Bidirectional LSTM-CRF Models for Sequence Tagging [Huang, Xu, Yu, 2015].
  – **Human Ceiling**: how well humans do?
    • human annotators: about 96% - 97% [Marcus et al., 1993].
    • when allowed to discuss tags, consensus is 100% [Voutilainen, 95]
  – **Most Frequent Class Baseline**:
    • 90% - 91% on the 87-tag Brown tagset [Charniak et al., 1993].
    • 93.69% on the 45-tag Penn Treebank, with unknown word model [Toutanova et al., 2003].
POS Tagging Methods

• **Rule Based:**
  – Rules are designed by human experts based on linguistic knowledge.

• **Machine Learning:**
  – Trained on data that has been manually labeled by humans.
  – Rule learning:
    • **Transformation Based Learning (TBL).**
  – **Sequence tagging:**
    • **Hidden Markov Models (HMM).**
    • **Maximum Entropy (MaxEnt).**
    • **Sequential Conditional Random Fields (CRF).**
    • **Recurrent Neural Networks (RNN):**
      – bidirectional, with a CRF layer (BI-LSTM-CRF).
POS Tagging: Rule Based

1) Start with a dictionary.

2) Assign all possible tags to words from the dictionary.

3) Write rules by hand to selectively remove tags, leaving the correct tag for each word.
1) Start with a dictionary:

- **she:** PRP
- **promised:** VBN, VBD
- **to:** TO
- **back:** VB, JJ, RB, NN
- **the:** DT
- **bill:** NN, VB

… for the ~100,000 words of English.
POS Tagging: Rule Based

2) Assign every possible tag:

NN
RB
VBN JJ VB
PRP VBD TO VB DT NN
She promised to back the bill
3) Write rules to eliminate incorrect tags.
   – Eliminate VBN if VBD is an option when VBN|VBD follows "<S> PRP"

She promised to back the bill
POS Tagging as Sequence Labeling

• **Sequence Labeling:**
  – Tokenization and Sentence Segmentation.
  – Part of Speech Tagging.
  – Information Extraction.
  – Shallow Parsing.
  – Semantic Role Labeling.
  – DNA Analysis.
  – Music Analysis.

• Solved using **Machine Learning**.
Sequence Labeling

- **Sentence Segmentation:**
  
  Mr. Burns is a Homer Simpson’s boss. He is very rich.

- **Tokenization:**
  
  Mr. Burns is a Homer Simpson’s boss. He is very rich.
Drug giant **Pfizer Inc.** has reached an agreement to buy the private biotechnology firm **Rinat Neuroscience Corp.**
Sequence Labeling

- **Information Extraction:**
  - segmenting classifieds into topical sections.

  Vine covered cottage, near Contra Costa Hills. 2 bedroom house, modern kitchen and dishwasher. No pets allowed. $1050 / month

[Haghighi & Klein, NAACL ‘06]
Sequence Labeling

• Information Extraction:
  – segmenting classifieds into topical sections.

Vine covered cottage, near Contra Costa Hills. 2 bedroom house,

modern kitchen and dishwasher. No pets allowed. $1050 / month

[Haghighi & Klein, NAACL ‘06]

– Features
– Neighborhood
– Size
– Restrictions
– Rent
Sequence Labeling

• **Semantic Role Labeling**: For each clause, determine the semantic role played by each noun phrase that is an argument to the verb:

  John drove Mary from Athens to Columbus in his Toyota Prius.  
The hammer broke the window.

  • agent
  • patient
  • source
  • destination
  • instrument
Sequence Labeling

• DNA Analysis:
  – transcription factor binding sites.
  – promoters.
  – introns, exons, …

AATGCGCTAACGTTCGATACGAGATAAGTCA
Sequence Labeling

• Music Analysis:
  – segmentation into “musical phrases”

[Romeo & Juliet, Nino Rota]
1) **Classify** each token *individually* into one of a number of classes:

- token represented as a vector of features extracted from context.
- to build classification model, use general ML algorithms:
  - Maximum Entropy (i.e. Logistic Regression)
  - Support Vector Machines (SVMs)  
    [Ratnaparkhi, EMNLP’96]
  - Neural Networks (Perceptrons, Multilayer Perceptrons).
  - Winnow.
  - Naïve Bayes, Bayesian Networks.
  - Decision Trees.
  - k-Nearest Neighbor, …
A Maximum Entropy Model for POS Tagging

[Ratnaparkhi, EMNLP’96]

• Represent each position \( i \) in text as \( \phi(t_i, h_i) = \{ \phi_k(t_i, h_i) \} \):
  
  - \( t_i \) is the identity of the POS tag at position \( i \).
  - \( h_i \) is the history/context of position \( i \).

  \[
  h_i = \{ w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2} \}
  \]

  - \( \phi(t_i, h_i) \) is a vector of features \( \phi_k(t_i, h_i) \), for \( k = 1..K \).

  \[
  \phi_k(t_i, h_i) = \begin{cases} 
  1 & \text{if suffix}(w_i) = "\text{ing}" \& \ t_i = \text{VBG} \\
  0 & \text{otherwise}
  \end{cases}
  \]

• Represent the “unnormalized” score of a tag as:

  \[
  \text{score}(t_i, h_i) = w^T \phi(t_i, h_i) = \sum_{k=1}^{K} w_k \phi_k(t_i, h_i)
  \]

  want \( w_k \) to be large here
A Maximum Entropy Model for POS Tagging [Ratnaparkhi, EMNLP’96]

<table>
<thead>
<tr>
<th>Condition</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_i$ is not rare</td>
<td>$w_i = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td>$w_i$ is rare</td>
<td>$X$ is prefix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$X$ is suffix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains number &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains uppercase character &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains hyphen &amp; $t_i = T$</td>
</tr>
<tr>
<td>$\forall w_i$</td>
<td>$t_{i-1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$t_{i-2}t_{i-1} = XY$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-2} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+2} = X$ &amp; $t_i = T$</td>
</tr>
</tbody>
</table>

Table 1: Features on the current history $h_i$

<table>
<thead>
<tr>
<th>Word:</th>
<th>the</th>
<th>stories</th>
<th>about</th>
<th>well-heeled</th>
<th>communities</th>
<th>and</th>
<th>developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag:</td>
<td>DT</td>
<td>NNS</td>
<td>IN</td>
<td>JJ</td>
<td>NNS</td>
<td>CC</td>
<td>NNS</td>
</tr>
<tr>
<td>Position:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Sample Data
A Maximum Entropy Model for POS Tagging

[Clips, EMNLP’96]

Table 2: Sample Data

<table>
<thead>
<tr>
<th>Word:</th>
<th>the stories about well-heeled communities and developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag:</td>
<td>DT NNS IN JJ NNS CC NNS</td>
</tr>
<tr>
<td>Position:</td>
<td>1 2 3 4 5 6 7</td>
</tr>
</tbody>
</table>

Table 1: Features on the current history $h_i$

- $w_i$ is not rare: $w_i = X \& t_i = T$
- $w_i$ is rare: $X$ is prefix of $w_i$, $|X| \leq 4 \& t_i = T$
- $w_i$ contains uppercase character $\& t_i = T$
- $w_i$ contains number $\& t_i = T$
- $w_i$ contains hyphen $\& t_i = T$
- $t_{i-1} = X \& t_i = T$
- $w_{i-2} = X \& t_i = T$
- $w_{i+1} = X \& t_i = T$
- $w_{i+2} = X \& t_i = T$

$w_i = \text{about} \& t_i = \text{IN}$
$w_{i-1} = \text{stories} \& t_i = \text{IN}$
$w_{i-2} = \text{the} \& t_i = \text{IN}$
$w_{i+1} = \text{well-heeled} \& t_i = \text{IN}$
$w_{i+2} = \text{communities} \& t_i = \text{IN}$
$t_{i-1} = \text{NNS} \& t_i = \text{IN}$
$t_{i-2} = \text{DT NNS} \& t_i = \text{IN}$

the non-zero features for position 3

feature templates
A Maximum Entropy Model for POS Tagging

[Robert Ratnaparkhi, EMNLP’96]

---

<table>
<thead>
<tr>
<th>Word:</th>
<th>the stories about well-heeled communities and developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag:</td>
<td>DT     NNS   IN    JJ     NNS   CC   NNS</td>
</tr>
<tr>
<td>Position:</td>
<td>1      2     3     4      5      6     7</td>
</tr>
</tbody>
</table>

Table 2: Sample Data

<table>
<thead>
<tr>
<th>Condition</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_i$ is not rare</td>
<td>$w_i = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td>$w_i$ is rare</td>
<td>$X$ is prefix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$X$ is suffix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains number &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains uppercase character &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains hyphen &amp; $t_i = T$</td>
</tr>
<tr>
<td>$\forall w_i$</td>
<td>$t_{i-1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$t_{i-2}t_{i-1} = XY$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-2} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+1} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+2} = X$ &amp; $t_i = T$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-1} = \text{about}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$w_{i-2} = \text{stories}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+1} = \text{communities}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$w_{i+2} = \text{and}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$t_{i-1} = \text{IN}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$t_{i-2}t_{i-1} = \text{NNS IN}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{prefix}(w_i) = \text{w}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{prefix}(w_i) = \text{we}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{prefix}(w_i) = \text{wel}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{prefix}(w_i) = \text{well}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{suffix}(w_i) = \text{d}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{suffix}(w_i) = \text{ed}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{suffix}(w_i) = \text{led}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$\text{suffix}(w_i) = \text{eled}$ &amp; $t_i = \text{JJ}$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains hyphen &amp; $t_i = \text{JJ}$</td>
</tr>
</tbody>
</table>

Table 1: Features on the current history $h_i$

---

**the non-zero features for position 4**
A Maximum Entropy Model for POS Tagging

How do we learn the weights $w$?
- Train on manually annotated data (supervised learning).

What does it mean “train $w$ on annotated corpus”?
- Probabilistic Discriminative Models:
  - **Maximum Entropy (MaxEnt).**
- Distribution Free Methods:
  - (Voted) Perceptrons. [Collins, ACL 2002]
  - Support Vector Machines (SVMs).
A Maximum Entropy Model for POS Tagging

[ Ratnaparkhi, EMNLP’96 ]

- Probabilistic Discriminative Model:
  \[ p(t_i | h_i) = \frac{\exp(w^T \varphi(t_i, h_i))}{\sum_{t'_i} \exp(w^T \varphi(t'_i, h_i))} \]

- Training using:
  - Maximum Likelihood (ML).
  - Maximum A Posteriori (MAP) with a Gaussian prior on \( w \).

- Inference (i.e. Testing):
  \[ \hat{t}_i = \arg \max_{t_i \in T} p(t_i | h_i) = \arg \max_{t_i \in T} \exp(w^T \varphi(t_i, h_i)) = \arg \max_{t_i \in T} w^T \varphi(t_i, h_i) \]
A Maximum Entropy Model for POS Tagging

[Animation by Ray Mooney, UT Austin]

John saw the saw and decided to take it to the table.

Inference, need to do Forward traversal of input sequence:

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

  John saw the saw and decided to take it to the table.

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

- Inference, need to do Forward traversal of input sequence:

`NNP VBD
John saw the saw and decided to take it to the table.`

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

John saw the saw and decided to take it to the table.
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

```
John saw the saw and decided to take it to the table.
```

Raster Parkhi, EMNLP’96

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

[Animation by Ray Mooney, UT Austin]

• Inference, need to do Forward traversal of input sequence:

NNP VBD DT NN CC
John saw the saw and decided to take it to the table.
A Maximum Entropy Model for POS Tagging [Ratnaparkhi, EMNLP’96]

- Inference, need to do Forward traversal of input sequence:

NNP VBD DT NN CC VBD
John saw the saw and decided to take it to the table.

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

NNP VBD DT NN CC VBD TO
John saw the saw and decided to take it to the table.

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

- Inference, need to do Forward traversal of input sequence:

  NNP VBD DT NN CC VBD TO VB
  John saw the saw and decided to take it to the table.

[Animation by Ray Mooney, UT Austin]

[Animation by Ray Mooney, UT Austin]

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

- Inference, need to do Forward traversal of input sequence:

```
NNP VBD DT NN CC VBD TO VB PRP
John saw the saw and decided to take it to the table.
```

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

```
NNP VBD DT NN CC VBD TO VB PRP IN
John saw the saw and decided to take it to the table.
```

[Animation by Ray Mooney, UT Austin]
A Maximum Entropy Model for POS Tagging

• Inference, need to do Forward traversal of input sequence:

NNP VBD DT NN CC VBD TO VB PRP IN DT
John saw the saw and decided to take it to the table.
A Maximum Entropy Model for POS Tagging

Inference, need to do Forward traversal of input sequence:

Disambiguating “to” would be easier backward, what can we do?

- use backward traversal, with backward features … but lose forward info.

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

1) **Classify** each token **individually** into one of a number of classes:

2) **Classify** all tokens **jointly** into one of a number of classes:

\[
\hat{t}_1 \ldots \hat{t}_n = \arg \max_{\hat{t}_1, \ldots, \hat{t}_n} \lambda^T \varphi(t_1, \ldots, t_n, w_1, \ldots, w_n)
\]

- Hidden Markov Models.
- Conditional Random Fields.
- Structural SVMs.
- Discriminatively Trained HMMs.
- Bi-directional RNNs.
Hidden Markov Models

- **Probabilistic Generative Models:**

\[
\hat{t}_1 \ldots \hat{t}_n = \arg \max_{t_1, \ldots, t_n} p(t_1, \ldots, t_n \mid w_1, \ldots, w_n)
\]

\[
= \arg \max_{t_1, \ldots, t_n} p(w_1, \ldots, w_n \mid t_1, \ldots, t_n) p(t_1, \ldots, t_n)
\]

- **Likelihood**
- **Prior**
Hidden Markov Models: Assumptions

1) A word event depends only on its POS tag:

\[ p(w_1,...,w_n \mid t_1,...,t_n) = \prod_{i=1}^{n} p(w_i \mid t_i) \]

2) A tag event depends only on the previous tag:

\[ p(t_1,...,t_n) = \prod_{i=1}^{n} p(t_i \mid t_{i-1}) \]

⇒ POS tagging is \( \hat{t}_1...\hat{t}_n = \arg \max_{t_1,...,t_n} \prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1}) \)
Interlude

Tales of HMMs