CS344: Introduction to Artificial Intelligence
(associated lab: CS386)

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 12, 13: Sequence Labeling using HMM- POS tagging; Baum Welch
1st Feb, 2011
POS tagging sits between Morphology and Parsing

- Semantics Extraction
- Parsing: Syntactic Processing
- **POS tagging**
- Morphological Processing

Rules and Resources
Because of this sequence, at the *level* of POS tagging the only information available is the word, its constituents, its properties and its neighbouring words and their properties.

\[ W_0 \quad W_1 \quad W_2 \quad \ldots \quad W_i \quad W_{i+1} \quad W_{n-1} \quad W_n \]

Word of interest
Cannot assume parsing and semantic processing

- Parsing identifies long distance dependencies
- Needs POS tagging which must finish earlier
- Semantic processing needs parsing and POS tagging
Example

- *Vaha ladakaa so rahaai hai*  
  *(that boy is sleeping)*
- *Vaha cricket khel rahaai hai*  
  *(he plays cricket)*
- The fact that “*vaha*” is demonstrative in the first sentence and pronoun in the second sentence, needs deeper levels of information
“vaha cricket” is not that simple!

- Vaha cricket jisme bhrastaachaar ho, hame nahii chaahiye
- (that cricket which has corruption in it is not acceptable to us)
- Here “vaha” is demonstrative
- Needs deeper level of processing
Syntactic processing also cannot be assumed

- `raam kaa yaha baar baar shyaam kaa ghar binaa bataayeJAANAA
  mujhe bilkul pasand nahii haai`

- *(I do not at all like the fact that Ram goes to Shyam's house repeatedly without informing (anybody))*

- "Ram-GENITIVE this again and again Shyam-GENITIVE house any not saying GOING I-dative at all like not VCOP"

- `JAANAA can be VINF (verb infinitive) or VN (verb nominal, i.e., gerundial)`
Syntactic processing also cannot be assumed (cntd.)

- *raam kaa yaha baar baar shyaam kaa ghar binaa bataaye JAANAA mujhe bilkul pasand nahii haai*

- The correct clue for disambiguation here is 'raam kaa', and this word group is far apart

- One needs to determine the structure of intervening constituents

- This needs parsing which in turn needs correct tags

- Thus there is a circularity which can be broken only by retaining ONE of VINF and VN.
Fundamental principle of POS tagset design

- IN THE TAGSET DO NOT HAVE TAGS THAT ARE POTENTIAL COMPETITORS AND TIE BETWEEN WHICH CAN BE BROKEN ONLY BY NLP PROCESSES COMING AFTER THE PARTICULAR TAGGING TASK.
Computation of POS tags
Process

- List all possible tag for each word in sentence.
- Choose best suitable tag sequence.
Example

- “People jump high”.
- People : Noun/Verb
- jump : Noun/Verb
- high : Noun/Adjective
- We can start with probabilities.
This model is called Generative model. Here words are observed from tags as states. This is similar to HMM.
Example of Calculation from Actual Data

- Corpus
  - ^Ram got many NLP books. He found them all very interesting.

- Pos Tagged
Recording numbers (bigram assumption)

\[
\begin{array}{cccccc}
\wedge & N & V & A & R & . \\
\wedge & 0 & 2 & 0 & 0 & 0 & 0 \\
N & 0 & 1 & 2 & 1 & 0 & 1 \\
V & 0 & 1 & 0 & 1 & 0 & 0 \\
A & 0 & 1 & 0 & 0 & 1 & 1 \\
R & 0 & 0 & 0 & 1 & 0 & 0 \\
. & 1 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

\^ Ram got many NLP books. He found them all very interesting.
Pos Tagged
\^ N V A N N . N V N A R A .
^ Ram got many NLP books. He found them all very interesting.
^ Pos Tagged
To find

- \( T^* = \text{argmax } (P(T) P(W/T)) \)
- \( P(T).P(W/T) = \prod_{i=1}^{n+1} P( t_i / t_{i-1} ).P(w_i /t_i) \)
- \( P( t_i / t_{i-1} ) : \text{Bigram probability} \)
- \( P(w_i /t_i) : \text{Lexical probability} \)

Note: \( P(w_i/t_i) = 1 \) for \( i=0 \) (\(^, \) sentence beginner)) and \( i=(n+1) \) (\( ., \) fullstop)
### Bigram probabilities

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>A</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>0.15</td>
<td>0.7</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>V</td>
<td>0.6</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>A</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>R</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Lexical Probability

<table>
<thead>
<tr>
<th></th>
<th>People</th>
<th>jump</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>$10^{-5}$</td>
<td>$0.4 \times 10^{-7}$</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>V</td>
<td>$10^{-7}$</td>
<td>$10^{-2}$</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>$10^{-1}$</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

values in cell are P(col-heading/row-heading)
Some notable text corpora of English

- American National Corpus
- Bank of English
- British National Corpus
- Corpus Juris Secundum
- Corpus of Contemporary American English (COCA)
  400+ million words, 1990-present. Freely searchable online.
- Brown Corpus, forming part of the "Brown Family" of corpora, together with LOB, Frown and F-LOB.
- International Corpus of English
- Oxford English Corpus
- Scottish Corpus of Texts & Speech
Accuracy measurement in POS tagging
Standard Bar chart: Per Part of Speech Accuracy

Per-POS Accuracy Distribution Using MF

POS tags: NN, NST, PRP, DEM, VM, VAUX, JJ, RB, PSP, RP, CC, QW, QF, QC, QQ, INTF, INJ, NEG, UT, RDP

Accuracy range: 0 to 100
### Standard Data: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>NN</th>
<th>NST</th>
<th>PRP</th>
<th>DEM</th>
<th>VM</th>
<th>VAUX</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>49988</td>
<td>18</td>
<td>92</td>
<td>2</td>
<td>167</td>
<td>4</td>
</tr>
<tr>
<td>NST</td>
<td>33</td>
<td>507</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>PRP</td>
<td>145</td>
<td>3</td>
<td>8071</td>
<td>312</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>DEM</td>
<td>3</td>
<td>0</td>
<td>231</td>
<td>3002</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>VM</td>
<td>225</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>17078</td>
<td>347</td>
</tr>
<tr>
<td>VAUX</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>257</td>
<td>6025</td>
</tr>
</tbody>
</table>

**Table:** POS Confusion Matrix with MF
How to check quality of tagging (P, R, F)

- Three parameters
  - Precision $P = \frac{|A \cap O|}{|O|}$
  - Recall $R = \frac{|A \cap O|}{|A|}$
  - F-score $= \frac{2PR}{P+R}$
    - Harmonic mean
Relation between P & R

Precision P

Recall R

P is inversely related to R (unless additional knowledge is given)
HMM Training

Baum Welch or Forward Backward Algorithm
Key Intuition

Given: Training sequence
Initialization: Probability values
Compute: $Pr(\text{state seq} | \text{training seq})$
  - get expected count of transition
  - compute rule probabilities
Approach: Initialize the probabilities and recompute them...
  - EM like approach
Baum-Welch algorithm: counts

String = abb aaa bbb aaa

Sequence of states with respect to input symbols

\[ o/p \text{ seq} \rightarrow q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r \]
Calculating probabilities from table

\[ P(q \xrightarrow{a} r) = \frac{5}{8} \]

\[ P(q \xrightarrow{b} r) = \frac{3}{8} \]

\[ P(s^i \xrightarrow{w_k} s^j) = \frac{\sum_{l=1}^{T} \sum_{m=1}^{A} c(s^i \xrightarrow{w_m} s^l)}{T} \]

\[ T = \#\text{states} \]

\[ A = \#\text{alphabet symbols} \]

Now if we have a non-deterministic transitions then multiple state seq possible for the given o/p seq (ref. to previous slide’s feature). Our aim is to find expected count through this.
Interplay Between Two Equations

\[ P(s^i \xrightarrow{W_k} s^j) = \frac{C(s^i \xrightarrow{W_k} s^j)}{\sum_{l=0}^{T} \sum_{m=0}^{A} C(s^i \xrightarrow{W_m} s^l)} \]

\[ C(s^i \xrightarrow{W_k} s^j) = \sum_{s_{0,n+1}} P(S_{0,n+1} \mid W_{0,n}) \times n(s^i \xrightarrow{W_k} s^j, S_{0,n+1}, W_{0,n}) \]

No. of times the transitions \( s^i \rightarrow s^j \) occurs in the string
Illustration

Actual (Desired) HMM

Initial guess
One run of Baum-Welch algorithm: *string ababb*

<table>
<thead>
<tr>
<th>$\epsilon \rightarrow a$</th>
<th>$a \rightarrow b$</th>
<th>$b \rightarrow a$</th>
<th>$a \rightarrow b$</th>
<th>$b \rightarrow b$</th>
<th>$b \rightarrow \epsilon$</th>
<th>$P(\text{path})$</th>
<th>$q \rightarrow r$</th>
<th>$r \rightarrow q$</th>
<th>$q \rightarrow \epsilon$</th>
<th>$q \rightarrow b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>$r$</td>
<td>$q$</td>
<td>$r$</td>
<td>$q$</td>
<td>$q$</td>
<td>0.00077</td>
<td>0.00154</td>
<td>0.00154</td>
<td>0</td>
<td>0.00077</td>
</tr>
<tr>
<td>$q$</td>
<td>$r$</td>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00884</td>
</tr>
<tr>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>$r$</td>
<td>$q$</td>
<td>$q$</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00442</td>
<td>0.00884</td>
</tr>
<tr>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>$q$</td>
<td>0.02548</td>
<td>0.0</td>
<td>0.000</td>
<td>0.05096</td>
<td>0.07644</td>
</tr>
</tbody>
</table>

| Rounded Total ->       | 0.035            | 0.01             | 0.01             | 0.06             | 0.095           |
| New Probabilities (P) -> | 0.06 = (0.01/(0.01+0.06+0.095)) | 1.0             | 0.36             | 0.581            |

* $\epsilon$ is considered as starting and ending symbol of the input sequence string. Through multiple iterations the probability values will converge.*
Computational part (1/2)

\[ C(s^i \xrightarrow{W_k} s^j) = \sum_{s_{0,n+1}} [P(S_{0,n+1} \mid W_{0,n}) \times n(s^i \xrightarrow{W_k} s^j, S_{0,n+1}, W_{0,n})] \]

\[ = \frac{1}{P(W_{0,n})} \sum_{s_{0,n+1}} [P(S_{0,n+1}, W_{0,n}) \times n(s^i \xrightarrow{W_k} s^j, S_{0,n+1}, W_{0,n})] \]

\[ = \frac{1}{P(W_{0,n})} \sum_{t=0,n} \sum_{s_{0,n+1}} [P(S_t = s^i, W_t = w_k, S_{t+1} = s^j, S_{0,n+1}, W_{0,n})] \]

\[ = \frac{1}{P(W_{0,n})} \sum_{t=0,n} [P(S_t = s^i, W_t = w_k, S_{t+1} = s^j, W_{0,n})] \]

\[ S0 \xrightarrow{w_0} S1 \xrightarrow{w_1} S1 \xrightarrow{w_2} ... S_i \xrightarrow{w_k} S_j ... \xrightarrow{w_{n-1}} Sn-1 \xrightarrow{w_n} Sn \xrightarrow{w_{n+1}} Sn+1 \]
Computational part (2/2)

\[
\sum_{t=0}^{n} P(S_t = s^i, S_{t+1} = s^j, W_t = w_k, W_{0,n})
\]

\[
= \sum_{t=0}^{n} P(W_{0,t-1}, S_t = s^i, S_{t+1} = s^j, W_t = w_k, W_{t+1,n})
\]

\[
= \sum_{t=0}^{n} P(W_{0,t-1}, S_t = s^i) P(S_{t+1} = s^j, W_t = w_k \mid W_{0,t-1}, S_t = s^i) P(W_{t+1,n} \mid S_{t+1} = s^j)
\]

\[
= \sum_{t=0}^{n} F(t-1, i) P(S_{t+1} = s^j, W_t = w_k \mid S_t = s^i) B(t+1, j)
\]

\[
= \sum_{t=0}^{n} F(t-1, i) P(S_{t+1} = s^j, W_t = w_k \mid S_t = s^i) B(t+1, j)
\]

\[
= \sum_{t=0}^{n} F(t-1, i) P(s^i \xrightarrow{w_k} s^j) B(t+1, j)
\]

\[
S_0 \rightarrow S_1 \rightarrow S_1 \rightarrow \ldots \rightarrow S_i \rightarrow S_j \ldots \rightarrow S_{n-1} \rightarrow S_n \rightarrow S_{n+1}
\]
Discussions

1. Symmetry breaking:
   Example: Symmetry breaking leads to no change in initial values

   Desired

   Initialized

2. Struck in Local maxima

3. Label bias problem
   Probabilities have to sum to 1.
   Values can rise at the cost of fall of values for others.