Midterm Preparation

Adam Meyers
New York University
Administrative Details

• Determines 1/4 of grade
• Time and Place: Wednesday March 9, 2016
• Ask clarification questions during test
  – I especially want to fix errors such as typos
• Open Book, Open Notes
  – You can bring materials
  – Search the web
  – Do simple calculations with a calculator
  – **DO NOT:**
    • communicate with others (texting, email, phone)
    • write/run actual programs
• Put your name and ID number on all test materials
• I will take attendance: please bring your ID
The Purpose of the Midterm?

• Pedagogical purposes
  – Track whether students learned parts of curriculum and what may need further clarification
  – Provide a motivating force for students to study the “important” parts of the curriculum
  – Clarify how to prepare students to do final projects

• Administrative purpose: determine 1/4 of grade

• Possible conflict
  – A difficult test makes pedagogical sense
  – An “acceptable” average grade may make administrative sense

• Current Strategy: Motivate test based on pedagogical objectives, but make it as open book as possible
  – You can bring materials, search the web, etc., but I will want you to solve the problems posed by the test
Outline

• Linguistic Resources & Descriptive Linguistics
  – Especially Corpus Annotation
• Rules used by Automated Procedures
  – Ones covered in Class
• Algorithms Discussed in Class
• How does Evaluation Work
Annotation

• You should be able to write usable specifications
• You should be able to annotate based on specifications
• You should understand some of the mechanics
  – Character offsets
  – A Markup language
  – BIO tags
• You should understand the difference between training and test corpora
Descriptive Linguistics

• The basic parts of speech and phrasal categories.
  – The difference between a determiners and an adjective
  – forms of verbs

• You know how to manually divide sentences into tokens

• You should know how to identify the head of a phrase

• You should be able to draw a phrase structure tree modeling the linguistic analysis of a sentence
Rules: Regular Expressions

• You should know how to write a basic regular expression

• You should know how to write a phrase structure rule including at least:
  – Context free rules
  – Left (or right) regular rules

• For a regular expression, you should be able to identify a set of phrase structure rules that describe the same language (set of strings)
Algorithms: Deterministic Finite State Machine

• Given:
  – Finite State Machine (FSM)
  – Input String

• Would the FSM recognize the string?

• Which sequence of states would be entered before recognition was complete?

• How would the FSM on the next slide process:
  – AababAB
  – AABB
DFSA for Regexp: $A(ab)^*ABB$?
Algorithms: Context-Free Generator

• Show the steps for randomly generating a sentence given:
  – A lexicon and a context-free grammar with start symbol $S$
• The algorithm expands each non-terminal into a randomly chosen right hand side.
• Going left to right, the first non-terminal symbol is always expanded first.
• The mechanism (as discussed in class) is to place each right hand side on top of the stack with the left-most symbol at the top of the stack.
Example of Generator

• Add S to top of empty stack
  – Stack is now: S

• Substitute **NP VP** for S
  – Stack is now: **NP VP**

• Substitute **DT N PP** for NP
  – Stack is now: **DT N PP VP**

• Substitute **DT** with **the**, pop off **the** (terminal)
  – Stack is now: **N PP VP**

• Etc.
Algorithms: The CKY parsing algorithm

• Fill in the triangular chart given a (short) sentence and a set of context free rules

• Remember
  – How the chart encodes start and end positions
  – That each rule is in Chomsky Normal Form
    • i.e., is binary branching

• See the next slide
### 6th Iteration of CKY Algorithm

<table>
<thead>
<tr>
<th>The</th>
<th>clam</th>
<th>'s</th>
<th>group</th>
<th>had</th>
<th>knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>D</td>
<td>NP</td>
<td>POSSP</td>
<td>NP</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>[0,1]</td>
<td>[0,2]</td>
<td>[0,3]</td>
<td>[0,4]</td>
<td>[0,5]</td>
</tr>
<tr>
<td>1</td>
<td>N, NP</td>
<td>POSSP</td>
<td>NP</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>[1,2]</td>
<td>[1,3]</td>
<td>[1,4]</td>
<td>[1,5]</td>
<td>[1,6]</td>
</tr>
<tr>
<td>2</td>
<td>POSS</td>
<td>N,NP</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>[2,3]</td>
<td>[3,4]</td>
<td>[3,5]</td>
<td>[3,6]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>N,NP</td>
<td>S</td>
<td>VP</td>
<td>VP</td>
<td>N,NP</td>
</tr>
<tr>
<td></td>
<td>[3,4]</td>
<td>[3,5]</td>
<td>[4,5]</td>
<td>[4,6]</td>
<td>[5,6]</td>
</tr>
</tbody>
</table>
Viterbi Decoding of HMM for *rose pickles*

• **Likelihood:**
  – *rose*: NNP .01, NN .02, VBD .05
  – *pickles*: NNP .001, NNS .03, VBZ .05

• **Transition Probabilities:**
## Rose Pickles

- **Likelihood:**
  - *rose*: NNP .01, NN .02, VBD .05
  - *pickles*: NNP .001, NNS .03, VBZ .05
- **Fill in:** max (previous X transition X likelihood)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1: Rose</th>
<th>2: Pickles</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td></td>
<td>.42 * .01</td>
<td>* .3 * .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* 0 * .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* .23 * .001</td>
<td></td>
</tr>
<tr>
<td>NNS</td>
<td></td>
<td></td>
<td>* 0 * .03</td>
<td>*.2 * .03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* .5 * .03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* .24 * .03</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>.20 * .02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VBZ</td>
<td></td>
<td></td>
<td>* .3 * .05</td>
<td>*.3 * .05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* .1 * .05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* 0 * .05</td>
<td></td>
</tr>
<tr>
<td>VBD</td>
<td></td>
<td>.05 * .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
<td></td>
<td></td>
<td>*.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*.45</td>
</tr>
</tbody>
</table>
# Rose/NNP Pickles/VBZ

- **Likelihood:**
  - *rose:* NNP .01, NN .02, VBD .05
  - *pickles:* NNP .001, NNS .03, VBZ .05

- Fill in: max (previous X transition X likelihood)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1: Rose</th>
<th>2: Pickles</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td></td>
<td>.42 * .01</td>
<td></td>
<td><em>.3</em>.001=1.26*10⁻⁶</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>0</em>.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>0</em>.23*.001</td>
</tr>
<tr>
<td>NNS</td>
<td></td>
<td></td>
<td><em>.0</em>.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>.5</em>.03=6*10⁻⁵</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>0</em>.24*.03</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>.20 * .02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VBZ</td>
<td></td>
<td></td>
<td><em>.3</em>.05=6.3*10⁻⁵</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>0</em>.1*.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>0</em>.05</td>
<td></td>
</tr>
<tr>
<td>VBD</td>
<td></td>
<td>.05 * .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
<td></td>
<td></td>
<td><em>.2=2.52</em>10⁻⁷</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>.15 = 9.0</em>10⁻⁶</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>.45 = 2.85</em>10⁻⁵</td>
</tr>
</tbody>
</table>
Common Evaluation Metrics

• If all instances are classified
  – Accuracy = Correct/All-Instances

• If only some instances are classified
  – Precision = Correct/Instances in System Output
  – Recall = Correct/Instances in Answer Key
  – F-measure = Mean of Precision and Recall
    • Actually Harmonic Mean of Precision and Recall
      – $\frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$
Sample Precision and Recall

- System for finding holiday names
- Exactly 10 correct holiday names in hand-coded corpus (the answer key)
- The system marks 12 holiday names, 8 of which match the ones in the answer key.
  - Precision = $8/12 = .67$
  - Recall = $8/10 = .80$
  - F-measure = $2/(.80+.67) = .73$
TFIDF

• TFIDF – Property of Term with respect to a document
  – keyword suitability, representativeness of a topic, etc.
  – Uses: Doc Retrieval, Term Extraction, etc.
• TF = frequency in a document
• IDF = number of documents in sample divided by number of documents containing word
• TFIDF = TF * log(IDF)
• Example: “rock” occurs 10 times in document X. It occurs in 100 out of 3000 documents in collection. TFIDF = 10*log(3000/100) = 34.01
Cosine Similarity Between Query and Document

\[
\text{Similarity}(A, B) = \frac{\sum_i a_i \times b_i}{\sqrt{\sum_i a_i^2 \times \sum_i b_i^2}}
\]

• Example:
  – the terms in the vectors include: animal, vegetable, mineral, monkey, golf enthusiast
  – The vector for the query is: [0,0,0,34,.8]
  – The vector for a given document is: [1,2,3,4,5]
  – What is the similarity?
Other Helpful Info

• There is a sample midterm online in the box next to these slides
• Some additional questions can be found on midterms from my previous NLP classes (see my website)