Computational Linguistics 1
CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland
1 September 2011

Agenda
• Administrivia
• Introduction to Computational Linguistics & applications
• Rule-based & statistical NLP

Computational Linguistics 1

Administrivia
• Course webpage: www.umiacs.umd.edu/~hollingk/classes/CompLing1-f11.html
• Course mailing list: umd-cmsc723-fall-2011@googlegroups.com
• Textbook
  • Speech and Language Processing
    Daniel Jurafsky and James H. Martin
• Teaching Assistant: Alex Ecins
• Office hours

Course Policies
• Policies
  • Attendance
  • Homework
  • Submitted by e-mail to: complying723.fall2011@gmail.com
  • Computer access?
  • Late/incomplete work
• Exams
• Grading
  • Exams
  • Homeworks
  • In-class participation
  • Readings

Pre-requisites
• Must have strong computational background
• Be a competent programmer
  • Depth-first search
  • Programming language: recommend Python/NLTK
• Be interested in linguistics
  • "The aged bottle flies fast"
• Enrollment/waitlist
• Machine Learning students?
What is Computational Linguistics?

- Computer processing of naturally-occurring language
- What humans do when processing language
- (vs) What linguists do when processing language
- Various names
  - Computational linguistics
  - Natural language processing (NLP)
  - Speech/language/text processing
  - Human language technology
- Interdisciplinary field
  - Roots in linguistics and computer science (specifically, AI)
  - Influenced by electrical engineering, cognitive science, psychology, and other fields
  - Dominated today by machine learning and statistics

Applications

- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)
- Optical Character Recognition (OCR)
  - Image processing, e.g., captchas
- Parsing: syntax & semantics
  - "The aged bottle flies fast"

Syntactic Analysis

- Parsing: the process of assigning syntactic structure
Semantics

- Different structures, same meaning:
  - I saw the man.
  - The man was seen by me.
  - The man was who I saw.
  - …

- Semantic representations attempt to abstract “meaning”
  - First-order predicate logic:
    \[ \exists x, \text{MAN}(x) \land \text{SEE}(x, I) \land \text{TENSE} \text{(past)} \]
  - Semantic frames and roles:
    \( (\text{PREDICATE} = \text{see}, \text{EXPERIENCER} = I, \text{PATIENT} = \text{man}) \)

Lexical Semantics

- Any verb can add “able” to form an adjective.
  - I taught the class. The class is teachable.
  - I loved that bear. The bear is loveable.
  - I rejected the idea. The idea is rejectable.

- Association of words with specific semantic forms
  - John: noun, masculine, proper
  - the boys: noun, masculine, plural, human
  - load/smear verbs: specific restrictions on subjects and objects

Applications

- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)
- Optical Character Recognition (OCR)
  - Image processing, e.g., captchas
- Parsing: syntax & semantics
  - “The aged bottle flies fast”
- Machine translation
  - “Maria no daba una bofetada a la bruja verde”
- Information extraction (Watson)
- Automatic essay grading
- Spell checking, grammar checking

Why is NLP hard?

- We do it all the time, practically without thinking about it!
- Garbled input
  - Noisy waveforms input to speech recognition
  - Distorted images for OCR
  - “Cascaded” errors
  - Cascades in NLP
- Ambiguity

At the word level

- Homophones
  - “It’s hard to wreck a nice beach”
- Part of speech
  - Duck!
    \[ \text{[VB Duck]} \]
  - Duck is delicious for dinner.
    \[ \text{[NN Duck]} \text{ is delicious for dinner.} \]
- Word sense
  - I went to the \text{bank} to deposit my check.
  - I went to the \text{bank} of the river to fish.
  - I went to the \text{bank} of windows and chose the one for “complaints”.

What’s a word?

- Break up by spaces, right?
  - \text{Ebay Sells Most of Skype to Private Investors}
  - Swine flu isn’t something to be feared
- What about these?
  - टाटा कहा, घाटा प(रा करो

(What’s a sentence…?)
At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope
- Structural ambiguity
  - I cooked her duck.
  - Visiting relatives can be annoying.
  - Time flies like an arrow.

Pragmatics and World Knowledge

- Interpretation of sentences requires context, world knowledge, speaker intention/goals, etc.
  - Example 1:
    - Could you turn in your assignments now? (command)
    - Could you finish the assignment? (question, command)
  - Example 2:
    - I couldn’t decide how to catch the thief. Then I decided to spy on the thief with binoculars.
    - To my surprise, I found out he had them too. Then I knew to just follow the thief with binoculars.

Difficult cases...

- Requires world knowledge:
  - The city council denied the demonstrators the permit because they advocated violence
  - The city council denied the demonstrators the permit because they feared violence
- Requires context:
  - John hit the man. He had stolen his bicycle.

Agenda

- Administrivia
- Introduction to Computational Linguistics & applications
- Rule-based & statistical NLP

Application Goals

- Science vs Engineering
  - Understanding the phenomenon of human language
  - Building better applications
- Accurate; minimize errors (false positives/negatives)
- Maximize coverage
- Robust, degrades gracefully
- Fast, scalable

Rule-Based Approaches

- Prevalent through the 80’s
  - Rationalism as the dominant approach
- Manually-encoded rules for various aspects of NLP
  - E.g., swallow is a verb of ingestion, taking an animate subject and a physical object that is edible, …
What’s the problem?

• Rule engineering is time-consuming and error-prone
  • Natural language is full of exceptions
• Rule engineering requires knowledge
  • Is this a bad thing?
• Rule engineering is expensive
  • Experts cost a lot of money
• Coverage is limited
  • Knowledge often limited to specific domains

More problems...

• Systems became overly complex and difficult to debug
  • Unexpected interaction between rules
• Systems were brittle
  • Often broke on unexpected input (e.g., “The machine swallowed my change.” or “She swallowed my story.”)
• Systems were uninformed by prevalence of phenomena
  • Why WordNet thinks congress is a donkey...

Problem isn’t with rule-based approaches per se, it’s with manual knowledge engineering...

The alternative?

• Empirical approach:
  learn by observing language as it’s used, “in the wild”
• Many different names:
  • Statistical NLP
  • Data-driven NLP
  • Empirical NLP
  • Corpus linguistics
  • ...Central tool: statistics
  • Fancy way of saying “counting things”

Advantages

• Generalize patterns as they exist in actual language use
• Little need for knowledge (just count!)
• Systems more robust and adaptable
• Systems degrade more gracefully

It’s all about the corpus!

• Corpus (pl. corpora): a collection of natural language text systematically gathered and organized in some manner
  • Brown Corpus, Wall Street journal, SwitchBoard, ...
• Can we learn how language works from corpora?
  • Look for patterns in the corpus

Features of a Corpus

• Size
• Balanced or domain-specific
• Written or spoken
• Raw or annotated
• Free or pay
• Other special characteristics (e.g., bitext)
Grab a "corpus"

Corpus Characteristics
- Size: ~0.5 MB
- Tokens: 71,370
- Types: 8,018
- Average frequency of a word: # tokens / # types = 8.9
  - But averages lie…

Most Frequent Words (Unigrams)

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq.</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>determiner (article)</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>conjunction</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>determiner</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>preposition, verbal infinitive marker</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>preposition</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>auxiliary verb</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
<td>(personal/epithet) pronoun</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>preposition</td>
</tr>
</tbody>
</table>

What else can we do by counting?

Raw Bigram Collocations

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word 1</th>
<th>Word 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>81871</td>
<td>of</td>
<td>the</td>
</tr>
<tr>
<td>55841</td>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>25430</td>
<td>to</td>
<td>the</td>
</tr>
<tr>
<td>21942</td>
<td>on</td>
<td>the</td>
</tr>
<tr>
<td>21639</td>
<td>for</td>
<td>the</td>
</tr>
<tr>
<td>18522</td>
<td>and</td>
<td>the</td>
</tr>
<tr>
<td>16131</td>
<td>that</td>
<td>the</td>
</tr>
<tr>
<td>15690</td>
<td>at</td>
<td>the</td>
</tr>
<tr>
<td>15449</td>
<td>to</td>
<td>be</td>
</tr>
<tr>
<td>13808</td>
<td>in</td>
<td>a</td>
</tr>
<tr>
<td>13399</td>
<td>of</td>
<td>a</td>
</tr>
<tr>
<td>13261</td>
<td>by</td>
<td>the</td>
</tr>
<tr>
<td>13183</td>
<td>with</td>
<td>the</td>
</tr>
<tr>
<td>12622</td>
<td>from</td>
<td>the</td>
</tr>
<tr>
<td>11428</td>
<td>New</td>
<td>York</td>
</tr>
</tbody>
</table>

Filtered Bigram Collocations

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word 1</th>
<th>Word 2</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>11467</td>
<td>New</td>
<td>York</td>
<td>A-N</td>
</tr>
<tr>
<td>7261</td>
<td>United</td>
<td>States</td>
<td>A-N</td>
</tr>
<tr>
<td>5412</td>
<td>Los</td>
<td>Angeles</td>
<td>N-N</td>
</tr>
<tr>
<td>3201</td>
<td>last</td>
<td>year</td>
<td>A-N</td>
</tr>
<tr>
<td>3191</td>
<td>Saudi</td>
<td>Arabia</td>
<td>N-N</td>
</tr>
<tr>
<td>2659</td>
<td>last</td>
<td>week</td>
<td>A-N</td>
</tr>
<tr>
<td>2014</td>
<td>vice</td>
<td>president</td>
<td>A-N</td>
</tr>
<tr>
<td>2378</td>
<td>Persian</td>
<td>Gulf</td>
<td>A-N</td>
</tr>
<tr>
<td>2161</td>
<td>San</td>
<td>Francisco</td>
<td>N-N</td>
</tr>
<tr>
<td>2136</td>
<td>President</td>
<td>Bush</td>
<td>N-N</td>
</tr>
<tr>
<td>2005</td>
<td>Middle</td>
<td>East</td>
<td>A-N</td>
</tr>
<tr>
<td>1942</td>
<td>Saddam</td>
<td>Hussein</td>
<td>N-N</td>
</tr>
<tr>
<td>1887</td>
<td>Soviet</td>
<td>Union</td>
<td>A-N</td>
</tr>
<tr>
<td>1850</td>
<td>White</td>
<td>House</td>
<td>A-N</td>
</tr>
<tr>
<td>1633</td>
<td>United</td>
<td>Nations</td>
<td>A-N</td>
</tr>
</tbody>
</table>
Learning verb “frames”

How is statistical NLP different?

• No need to think of examples, exceptions, etc.
• Generalizations are guided by prevalence of phenomena
• Resulting systems better capture real language use

Three Pillars of Statistical NLP

• Corpora
• Representations
• Models and algorithms

Agenda

• Administrivia
• Introduction to Computational Linguistics & applications
• Rule-based & statistical NLP

HW0:
• Online tonight, due next Thursday before class

Next time:
• Introduction to finite-state models:
  regular expressions, Chomsky hierarchy, automata and transducers