Advanced Natural Language Processing:
Background and Overview

Regina Barzilay and Michael Collins
EECS/CSAIL

September 7, 2005
## Course Logistics

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Regina Barzilay, Michael Collins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td><a href="mailto:regina@csail.mit.edu">regina@csail.mit.edu</a>, <a href="mailto:mcollins@csail.mit.edu">mcollins@csail.mit.edu</a></td>
</tr>
<tr>
<td>Classes</td>
<td>Tues&amp;Thurs 13:00–14:30</td>
</tr>
<tr>
<td>Location</td>
<td>Room 32-155</td>
</tr>
</tbody>
</table>
Questions that today’s class will answer

- What is Natural Language Processing (NLP)?
- Why is NLP hard?
- Can we build programs that learn from text?
- What will this course be about?
What is Natural Language Processing?

computers using natural language as input and/or output

Diagram:

language  →  computer  →  language

understanding (NLU)  ←  generation (NLG)
is pushing for the resolution.

the strength and solidarity of the official Lebanese position.

coastlines. However, Minister of Information, said that the French forces and those of other international control because of the request by the Lebanese government of West Lebanon.

Israel lifted the siege today and monitor heets.

Thursday, August 1, 14:42 New York Time (the latest update)

This page has been automatically translated from Arabic. BETA.

Google Translation
10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automotive account. Experience in online marketing, automotive and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables . . . Compensation: $50,000-$80,000 Hiring Organization: 10TH DEGREE

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>Advertising</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITION</td>
<td>Assistant Account Manager</td>
</tr>
<tr>
<td>LOCATION</td>
<td>Irvine, CA</td>
</tr>
<tr>
<td>COMPANY</td>
<td>10TH DEGREE</td>
</tr>
<tr>
<td>SALARY</td>
<td>$50,000-$80,000</td>
</tr>
</tbody>
</table>
Information Extraction

• Goal: Map a document collection to structured database

• Motivation:
  – Complex searches (“Find me all the jobs in advertising paying at least $50,000 in Boston”)
  – Statistical queries (“Does the number of jobs in accounting increases over the years?”)
Transcript Segmentation
Advanced Natural Language Processing: Background and Overview

Text Summarization

Summary from the U.S.

Agency Suspends Smallpox Vaccines for People with Heart Disease

Second Smallpox Vaccine Death Occurred After Second Smallpox Vaccination (cnn.com, 03/28/2003, 358 words)

2. Second worker dies after smallpox vaccination (cnn.com, 03/28/2003, 499 words)

3. Second worker dies after smallpox vaccination (dallasmorning.com, 03/27/2003, 559 words)

4. Smallpox vaccine is revealed after second smallpox vaccination (dallasmorning.com, 03/28/2003, 732 words)

5. Smallpox vaccine is revealed after second smallpox vaccine (portland.com, 03/28/2003, 857 words)

Source articles

Vaccine, Heart, Smallpox, Vaccinated, Disease

Story Keywords

Responders, (1)

Summary: The U.S. Centers for Disease Control and Prevention (1) has announced that a second health care worker has died of a heart attack (3) after receiving a smallpox vaccination. The vaccine has been associated with heart trouble, but a cardiac problem (2) has been identified. The vaccine has been suspended for people with heart disease.
User: I need a flight from Boston to Washington, arriving by 10 pm.
System: What day are you flying on?
User: Tomorrow
System: Returns a list of flights
Why is NLP Hard?
[example from L.Lee]

“At last, a computer that understands you like your mother”
Ambiguity

“At last, a computer that understands you like your mother”

1. (*) It understands you as well as your mother understands you
2. It understands (that) you like your mother
3. It understands you as well as it understands your mother

1 and 3: Does this mean well, or poorly?
Ambiguity at Many Levels

At the **acoustic** level (speech recognition):

1. “... a computer that understands you *like your mother*”

2. “... a computer that understands you *lie cured mother*”
Ambiguity at Many Levels

At the *syntactic* level:

Different structures lead to different interpretations.
More Syntactic Ambiguity

Advanced Natural Language Processing: Background and Overview 14/48
Ambiguity at Many Levels

At the **semantic** (meaning) level:

Two definitions of “mother”

- a woman who has given birth to a child
- a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

This is an instance of **word sense ambiguity**
More Word Sense Ambiguity

At the **semantic** (meaning) level:

- They put money in the **bank**
  = buried in mud?

- I saw her duck with a telescope
Ambiguity at Many Levels

At the **discourse** (multi-clause) level:

- Alice says they’ve built a computer that understands you like your mother

- But *she* . . .
  
  . . . doesn’t know any details
  
  . . . doesn’t understand me at all

This is an instance of **anaphora**, where she co-referees to some other discourse entity
Knowledge Bottleneck in NLP

We need:

- Knowledge about language
- Knowledge about the world

Possible solutions:

- Symbolic approach: Encode all the required information into computer
- Statistical approach: Infer language properties from language samples
Case study: Determiner Placement

Task: Automatically place determiners (*a, the, null*) in a text

Scientists in United States have found way of turning lazy monkeys into workaholics using gene therapy. Usually monkeys work hard only when they know reward is coming, but animals given this treatment did their best all time. Researchers at National Institute of Mental Health near Washington DC, led by Dr Barry Richmond, have now developed genetic treatment which changes their work ethic markedly. "Monkeys under influence of treatment don’t procrastinate," Dr Richmond says. Treatment consists of anti-sense DNA - mirror image of piece of one of our genes - and basically prevents that gene from working. But for rest of us, day when such treatments fall into hands of our bosses may be one we would prefer to put off.
Relevant Grammar Rules

- Determiner placement is largely determined by:
  - Type of noun (countable, uncountable)
  - Reference (specific, generic)
  - Information value (given, new)
  - Number (singular, plural)

- However, many exceptions and special cases play a role:
  - The definite article is used with newspaper titles (*The Times*), but zero article in names of magazines and journals (*Time*)
Symbolic Approach: Determiner Placement

What categories of knowledge do we need:

- **Linguistic knowledge:**
  - Static knowledge: number, countability, . . .

- **World knowledge:**
  - Uniqueness of reference (*the current president of the US*), type of noun (*newspaper vs. magazine*), situational associativity between nouns (*the score of the football game*), . . .

Hard to manually encode this information!
Statistical Approach: Determiner Placement

Naive approach:

- Collect a large collection of texts relevant to your domain (e.g., newspaper text)

- For each noun seen during training, compute its probability to take a certain determiner

\[
p(determiner|noun) = \frac{freq(noun,determiner)}{freq(noun)}
\]

(assuming \(freq(noun) > 0\))

- Given a new noun, select a determiner with the highest likelihood as estimated on the training corpus
Does it work?

- Implementation
  - Corpus: training — first 21 sections of the Wall Street Journal (WSJ) corpus, testing — the 23th section
  - Prediction accuracy: 71.5%

- The results are not great, but surprisingly high for such a simple method
  - A large fraction of nouns in this corpus always appear with the same determiner
    “the FBI”, “the defendant”, …
Determiner Placement as Classification

- **Prediction:** “the”, “a”, “null”
- **Representation of the problem:**
  - plural? (yes, no)
  - first appearance in text? (yes, no)
  - noun (members of the vocabulary set)

<table>
<thead>
<tr>
<th>Noun</th>
<th>plural?</th>
<th>first appearance</th>
<th>determiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>defendant</td>
<td>no</td>
<td>yes</td>
<td>the</td>
</tr>
<tr>
<td>cars</td>
<td>yes</td>
<td>no</td>
<td>null</td>
</tr>
<tr>
<td>FBI</td>
<td>no</td>
<td>no</td>
<td>the</td>
</tr>
<tr>
<td>concert</td>
<td>no</td>
<td>yes</td>
<td>a</td>
</tr>
</tbody>
</table>

**Goal:** Learn classification function that can predict unseen examples
Classification Approach

- Learn a function from $X \rightarrow Y$ (in the previous example, $Y = \{"the", "a", null\}$)

- Assume there is some distribution $D(x, y)$, where $x \in X$, and $y \in Y$. Our training sample is drawn from $D(x, y)$.

- Attempt to explicitly model the distribution $D(X, Y)$ and $D(X|Y)$
Basic NLP Problem: Tagging

**Task:** Label each word in a sentence with its appropriate part of speech (POS)

*Time/Noun flies/Verb like/Preposition an/Determiner arrow/Noun*

<table>
<thead>
<tr>
<th>Word</th>
<th>Noun</th>
<th>Verb</th>
<th>Preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>flies</td>
<td>21</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>10</td>
<td>30</td>
<td>21</td>
</tr>
</tbody>
</table>
Basic NLP Problem: Tagging

- Naive solution: for each word, determine its tag independently
- Desired alternative: take into account dependencies among different predictions
  - Classification is suboptimal
  - We will model tagging as a mapping from a string to a tagged sequence
Beyond Classification

Many NLP applications can be viewed as a mapping from one complex set to another:

- Parsing: strings to trees
- Machine Translation: strings to strings
- Natural Language Generation: database entries to strings

Classification framework is not suitable in these cases!
Boeing is located in Seattle.
Parsing

- Penn WSJ Treebank = 50,000 sentences with associated trees
- Usual set-up: 40,000 training sentences, 2400 test sentences

Canadian Utilities had 1988 revenue of C$ 1.16 billion, mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers.
<table>
<thead>
<tr>
<th>Он благополучно избежал встречи с своей хозяйкой на лестнице.</th>
</tr>
</thead>
<tbody>
<tr>
<td>He had successfully avoided meeting his landlady on the staircase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Каморка его приходилась под самою кровлей высокого пятиэтажного дома и походила более на шкаф, чем на квартиру.</th>
</tr>
</thead>
<tbody>
<tr>
<td>His garret was under the roof of a high, five-storied house and was more like a cupboard than a room.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Квартирная же хозяйка его, у которой он нанимал эту каморку с обедом и прислугой, помещалась одной лестницей ниже, в отдельной квартире.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The landlady who provided him with garret, dinners, and attendance, lived on the floor below.</td>
</tr>
</tbody>
</table>
What will this Course be about?

- Computationally suitable and expressive representations of linguistic knowledge at various levels: syntax, semantics, discourse
- Algorithms for learning language properties from text samples: smoothed estimation, log-linear models, probabilistic context free grammars, the EM algorithm, co-training, ...
- Technologies underlying text processing applications: machine translation, text summarization, information retrieval
Syllabus

Estimation techniques, and language modeling (1 lecture)
Parsing and Syntax (5 lectures)
The EM algorithm in NLP (1 lecture)
Stochastic tagging, and log-linear models (2 lectures)
Probabilistic similarity measures and clustering (2 lectures)
Machine Translation (2 lectures)
Discourse Processing: segmentation, anaphora resolution (3 lectures)
Dialogue systems (1 lectures)
Natural Language Generation/Summarization (1 lecture)
Unsupervised methods in NLP (1 lecture)
Books
Prerequisites

- Interest in language and basic knowledge of English
- Some basic linear algebra, probability, algorithms at the level of 6.046
- Some programming skills
Assessment

- Midterm (20%)
- Final (30%)
- 5 homeworks (50%)
Counting Words
What is a Word?

- **English:**
  - “Wash. vs wash”
  - “won’t”, “John’s”
  - “85-year-old grandmother”, “the idea of a child-as-required-yuppie-possession must be motivating them”;

- **East Asian languages:**
  - words are not separated by white spaces
Counting Words

- **Type** — number of distinct words in a corpus (vocabulary size)
- **Token** — total number of words in a corpus

Word Distribution from Tom Sawyer:
word types — 8,018
word tokens — 71,370
average frequency — 9
## Frequency Distribution in Tom Sawyer

<table>
<thead>
<tr>
<th>word</th>
<th>Freq. (f)</th>
<th>Rank (r)</th>
<th>( f \times r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>3</td>
<td>5235</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>8770</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8400</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
</tr>
<tr>
<td>there</td>
<td>222</td>
<td>40</td>
<td>8880</td>
</tr>
<tr>
<td>one</td>
<td>172</td>
<td>50</td>
<td>8600</td>
</tr>
<tr>
<td>about</td>
<td>158</td>
<td>60</td>
<td>9480</td>
</tr>
<tr>
<td>never</td>
<td>124</td>
<td>80</td>
<td>9920</td>
</tr>
<tr>
<td>Oh</td>
<td>116</td>
<td>90</td>
<td>10440</td>
</tr>
</tbody>
</table>
Zipf’s Law

Zipf’s Law captures the relationship between frequency and rank.

If the most frequently occurring word appears in the text with the frequency $P(1)$, the next most frequently occurring word has frequency $P(2)$, and the rank-$r$ word has the frequency $f(r)$, the frequency distribution is:

$$f(r) = \frac{C}{r},$$

with $C$ is a constant.
Zipf's Law
Zipf’s Law and Principle of Least Effort

*Human Behavior and the Principle of Least Effort* (Zipf): “… Zipf argues that he found a unifying principle, the Principle of Least Effort, which underlies essentially the entire human condition (the book even includes some questionable remarks on human sexuality!). The principle argues that people will act so as to minimize their probable average rate of work”. (Manning & Schutze, p.23)
Examples of collections approximately obeying Zipf’s law

- Frequency of accesses to web pages
- Sizes of settlements
- Income distribution amongst individuals
- Size of earthquakes
- Notes in musical performances
Is Zipf’s Law unique to human language?

(Li 1992): randomly generated text exhibits Zipf’s law

- Consider monkey language: a generator that randomly produces characters from the \((M+1)\) letters of the alphabet and the blank.

- A word is a “nonblank” symbol string ended by a blank

- Probability of a word \(w\) is determined by its length
  - if \(M=26\): \(P(a_\_)=P(b_\_) = \ldots = \frac{1}{27^2}\)
  - In general: \(P_i(L) = \frac{1}{(M+1)^L+1}\), where \(P_i(L)\) is the probability of any word of length \(L\)

- There are \(M^L\) words having length \(L\)
Monkey Language (cont.)

- All words with the the length $L$ rank higher than word with the length $L + 1$, because they have larger value of frequency of occurrence.

- If we represent the rank of any word with length $L$ by $r(L)$:

$$
\sum_{l=1}^{L-1} M^l < r(L) \leq \sum_{l=1}^{L} M^l
$$
Intuition behind the Proof

- Probability of any word of length $L$ decreases exponentially in $L$

$$p \approx \frac{1}{M^L}$$

- Rank of a word grows exponentially in the length of a word $L$ (because there are exponentially many words of length $L$)

$$r \approx M^L$$

- If the rates of exponential growth are the same in both cases, we can say that the probability is inversely proportional to the rank

$$p \approx \frac{1}{r}$$
Conclusions

- Zipf’s Law is not a distinctive property of natural language texts
- Most tokens have low frequency even in a large text collection
  - Sparsity is a major problem for statistical language learning

Next time: How to estimate probability of unseen elements?