Natural Language Processing
Lecture 1: Introduction

Xiaoli Fern (CS519, Winter 2015)
Oregon State University
This Lecture

• What is Natural Language Processing (NLP)?
• Why is NLP hard?
• What will this course be about?
What is Natural Language Processing?

computers using natural language as input and/or output

language → computer → language

NLU
Natural Language Understanding

NLG
Natural Language Generation
What is NLP?

Fundamental goal: *deep* understand of *broad* language
Not just string processing or keyword matching!
Why Should You Care?

Trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers
Some Simple Applications of NLP

- Word counters (wc in UNIX)
- Spell Checkers, grammar checkers
- Predictive Text on mobile handsets

Bigger Applications of NLP

- Machine translation
- Information retrieval
- Speech recognition
- Intelligent Web searching
- Data mining
- Natural language generation
- Question answering
Local time on January 2, Kentucky, a small plane crash occurred, resulting including pilots, including four people were killed. A 7-year-old girl survived, police used "bizarre" to describe the name of the only survivors. Winter season, Kentucky temperatures about 4 degrees Celsius, the little girl was wearing a short-sleeved shirt and shorts in the cold night, walking barefoot about 1.6 km, during which pass through two dams, a mountain, a river, and finally reached a residents of the home to obtain relief. Police arrived about 10 minutes later, the little girl to the hospital. Police say the child is able to coherently answer police inquiries, in addition to the wrist injury, the body will not cause illness.
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 (Knowledgebase Population)

• 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 ("How has the number of jobs in accounting changed over the years?")
Agency Suspends Smallpox Vaccines for People With Heart Disease

Summary from the U.S.

A second health care worker has died of a heart attack (3) after receiving a smallpox vaccination (9) and officials are investigating whether vaccinations are to blame (3) for cardiac problems. (6) The vaccine never has been associated with heart trouble but as a precaution (3) the U.S. centers for Disease Control and Prevention (14) is advising people with a history of heart disease to be vaccinated (3) until further notice. (14) Strom suggested that the Bush administration reassess whether it necessary and safe to continue with its aggressive plan to inoculate millions of health care workers and emergency responders. (1)

Story keywords

vaccine, Heart, Smallpox, vaccinated, Disease

Source articles

1. Vaccination program in peril after second death (seattletimes.nwsourse.com, 03/28/2003, 319 words)
2. Wired News: Smallpox Shots: Proceed With Care (Wired, 03/27/2003, 559 words)
3. 2nd worker dies after smallpox vaccination (santimes.com, 03/28/2003, 358 words)
4. 2nd worker dies after smallpox vaccine (dallasnews.com, 03/28/2003, 499 words)
5. Smallpox vaccine is reviewed after second fatal heart attack (boston.com, 03/28/2003, 732 words)
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
...Water is split, providing a source of electrons and protons (hydrogen ions, H\(^+\)) and giving off O\(_2\) as a by-product. Light absorbed by chlorophyll drives a transfer of the electrons and hydrogen ions from water to an acceptor called NADP\(+\) ... 

**Question:** what can the split of water lead to?

a. Light absorption  
b. Transfer of ions
Building a computer that understands text: some basic tasks in the NLP pipeline
Task: Tokenization/segmentation

We need to split text into words and sentences
• Languages like Chinese don’t have spaces between words.
• Even in English, this cannot be done deterministically:
  There was an earthquake near D.C. You could even feel it in Philadelphia, New York, etc.

NLP task:
What is the most likely segmentation/tokenization?
Basic NLP Tasks: POS Tagging and Syntactic Parsing

INPUT:

Boeing is located in Seattle.

OUTPUT:
Basic NLP Tasks: Dependency Parsing

*They hid the letter on the shelf*
Basic NLP Tasks: Coreference

But the little prince could not restrain admiration:

"Oh! How beautiful you are!"

"Am I not?" the flower responded, sweetly. "And I was born at the same moment as the sun . . ."

The little prince could guess easily enough that she was not any too modest--but how moving--and exciting--she was!

"I think it is time for breakfast," she added an instant later. "If you would have the kindness to think of my needs--"

And the little prince, completely abashed, went to look for a sprinkling-can of fresh water. So he tended the flower.
Basic NLP Tasks: Relation Extraction

**relation**: a semantic relationship between two entities

<table>
<thead>
<tr>
<th>ACE relation type</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent-Artifact</td>
<td>Rubin Military Design, the <em>makers</em> of the <em>Kursk</em></td>
</tr>
<tr>
<td>Discourse</td>
<td><em>each of whom</em></td>
</tr>
<tr>
<td>Employment/ Membership</td>
<td>Mr. Smith, a senior <em>programmer</em> at <em>Microsoft</em></td>
</tr>
<tr>
<td>Place-Affiliation</td>
<td><em>Salzburg</em> Red Cross <em>officials</em></td>
</tr>
<tr>
<td>Person-Social</td>
<td><em>relatives</em> of the <em>dead</em></td>
</tr>
<tr>
<td>Physical</td>
<td>a <em>town</em> some 50 miles south of <em>Salzburg</em></td>
</tr>
<tr>
<td>Other-Affiliation</td>
<td><em>Republican senators</em></td>
</tr>
</tbody>
</table>
This Lecture

• What is Natural Language Processing (NLP)?
• Why is NLP hard?
• What will this course be about?
Why is NLP Hard?

(Example from L. Lee)

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

VP
  V
    list

  NP
    DET
      all
    N
      flights
  PP
    on Tuesday

VP
  V
  NP
    DET
      all
    N
      flights
  PP
    on Tuesday
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 or a financial institute?

I saw her *duck* with a telescope
  = a water fowl or move out of the way suddenly?
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
Course Coverage

- NLP sub-problems: part-of-speech tagging, parsing, word-sense disambiguation, etc.
- Machine learning techniques: probabilistic context-free grammars, hidden Markov models, estimation/smoothing techniques, the EM algorithm, log-linear models, etc.
- Applications: information extraction, machine translation …
A Tentative Syllabus/Schedule

- Language modeling, smoothed estimation (4 lectures)
- Tagging, hidden Markov models (3 lectures)
- Statistical parsing with (P)CFGs (6 lectures)
- Machine translation (6 lectures)
- Lexical semantics and words senses (3 lectures)
- Log-linear models, discriminative methods (3 lectures)
- Semi-supervised and unsupervised learning for NLP (3 lectures)
How is this course run?

Textbook: no required text book. Will use Michael Collins’s notes on many subjects. I will also post notes on contents not covered by his notes.

Optional but recommended:

  Jurafsky and Martin
  Speech and Language Processing (2nd Edition)

Lectures:
- Mixture of lecturing and problem-solving/ discussion sessions
- Some sessions require students to watch video lectures prior to class

Assignments:
- 4 Programming assignments (team of up to 2 people)
- 3 literature review/reading assignments
- Final project: an in-depth literature survey or empirical exploration of some NLP task(s)

Grading:
- Class participation: 10%
- Programming assignments: 30%
- Reading assignments: 10%
- Midterm: 25%
- Final project: 25%