Overview

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

Course Logistics

Instructor  Michael Collins  
Email    mcollins@csail.mit.edu  
Classes  Tues&Thurs 13:00–14:30  
Location  Room 32-144  
Webpage  http://people.csail.mit.edu/mcollins/6864  

TA  Igor Malioutov  
Email    igorm@csail.mit.edu  

What is Natural Language Processing?

computers using natural language as input and/or output

language \[\rightarrow\] computer \[\rightarrow\] language

understanding (NLU) \[\leftarrow\] generation (NLG)
Machine Translation:

*e.g., Google Translation from Arabic*

Stock prices retreated in the stock markets again with increasing concern about the circumstances surrounding the credit markets in the world, due mostly to the problems it faces American mortgage lending market, which raised concern among investors.

The index retreated Vuciji / 100 on the London Stock Exchange at the beginning of a percentage point in the dealings of up to 6082 points, while the Nikkei index retreated / 225 Japanese rate of 2.2% to close at the lowest level in eight months.

The American Jones index has lost about 1.6 points Tuesday to reach 13029 points, the Nasdaq index had lost 1.7% of its value.

These declines came despite statements by the American Federal Reserve Bank (Central Bank), in which he said that the process of pumping more funds into capital markets when necessary.

The American Federal Reserve Board, for the purposes of relaxation of tension in global financial markets, resulting in the Gaza backtracks American real estate lending, have pumped billions of dollars of emergency funds allocation to the banking sector during the past few days, on Friday and Monday. As the European Central Bank did the same.

---

**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 (“How has the number of jobs in accounting changed over the years?”)

---

**Text Summarization**

**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 after receiving a smallpox vaccination, and officials are investigating whether vaccinations are to blame for cardiac problems. The vaccine has never been associated with heart trouble but as a precaution the U.S. centers for Disease Control and Prevention is advising people with a history of heart disease to be vaccinated until further notice. (4) Stamos suggested that the Bush administration reassures whether it is necessary and safe to continue with an aggressive plan to inoculate millions of health care workers and emergency responders.

**Story keywords**

vaccine, Heart, Smallpox, vaccinated, Disease

**Source articles**

1. Vaccination program in peril after second death (sandpoint MORE.com, 03/18/2003, 319 words)
2. Wired News: Smallpox Shots: Proceed With Care (Wisconsin, 03/18/2003, 559 words)
3. ND worker dies after smallpox vaccination (sandpoint MORE.com, 03/18/2003, 358 words)
4. ND worker dies after smallpox vaccinate (sandpoint MORE.com, 03/18/2003, 499 words)
5. Smallpox vaccine is reviewed after second fatal heart attack (sandpoint MORE.com, 03/18/2003, 732 words)
6. Second Smallpox Vaccine Death: (sandpoint MORE.com, 03/18/2003, 864 words)
**Dialogue Systems**

**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

---

**Basic NLP Problems: Tagging**

**TAGGING:** Strings to Tagged Sequences

\[ a b e e a f h j \Rightarrow \text{a/C b/D e/C e/C a/D f/C h/D j/C} \]

**Example 1:** Part-of-speech tagging

Profits/N soared/V at/P Boeing/N Co./N,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ./.  

**Example 2:** Named Entity Recognition

Profits/NA soared/NA at/NA Boeing/SC Co./CC,/, NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ./.  

---

**Basic NLP Problems: Parsing**

**INPUT:**

Boeing is located in Seattle.

**OUTPUT:**

```
S
  NP
    N Boeing
  VP
    V is
    VP
      V located
      PP
        P in
          NP
            N Seattle
```

---

**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 **syntactic** level:

![Syntactic Levels Diagram]

Different structures lead to different interpretations.

---

**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”

---

**More Syntactic Ambiguity**

![Syntactic Levels Diagram]
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-references 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, natural language interfaces...
## A Syllabus

- Language modeling, smoothed estimation (1 lecture)
- Statistical parsing (4 lectures)
- Log-linear models (1 lecture)
- Tagging (1 lecture)
- History-based models (1 lecture)
- The EM algorithm in NLP (2 lectures)
- Machine translation (3 lectures)
- Global linear models (2 lectures)

## Prerequisites

- Basic linear algebra, probability, algorithms at the level of 6.046
- Programming skills

## Assessment

- Midterm (20%)
- Final (30%)
- 4 homeworks (25%)
- Final project (25%)

---

*Advanced Natural Language Processing: Background and Overview 20*

*Advanced Natural Language Processing: Background and Overview 21*

*Advanced Natural Language Processing: Background and Overview 22*

*Advanced Natural Language Processing: Background and Overview 23*
Advanced Natural Language Processing: Background and Overview 24