Natural Language Processing

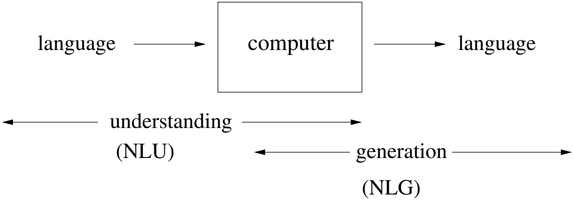
Michael Collins, Columbia University

Overview

- 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



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.

Information Extraction

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 automative account. Experience in online marketing, automative 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

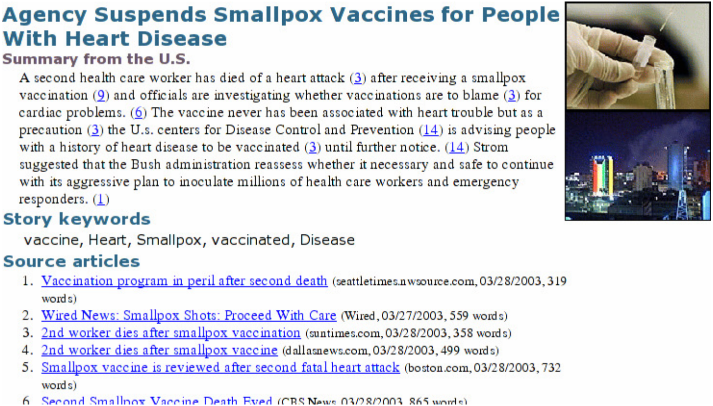
⇓

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Irvine, CA
COMPANY	10TH DEGREE
SALARY	$50,000-$80,000

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 (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.nwsource.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 (suntimes.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)
6. Second Smallpox Vaccine Death Eyed (CBS News 03/28/2003, 865 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 ⇒ 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

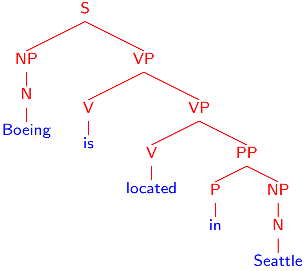
Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ./.

Basic NLP Problems: Parsing

INPUT:

Boeing is located in Seattle.

OUTPUT:



Overview

- 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”

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?

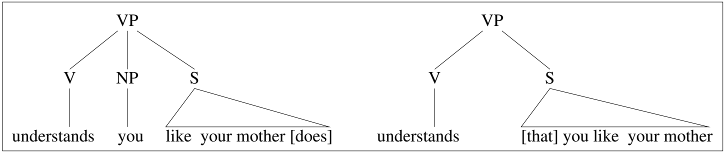
Ambiguity at Many Levels

At the acoustic level (speech recognition):

1. " ... a computer that understands you like your mother"
2. " ... a computer that understands you lie cured mother"

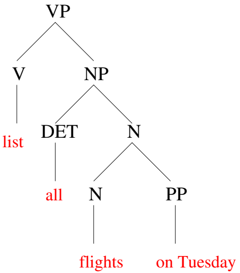
Ambiguity at Many Levels

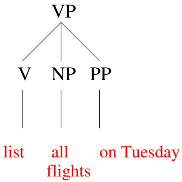
At the syntactic level:



Different structures lead to different interpretations.

More Syntactic Ambiguity





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

Overview

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

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
- Tagging, hidden Markov models
- Statistical parsing
- Machine translation
- Log-linear models, discriminative methods
- Semi-supervised and unsupervised learning for NLP

Prerequisites

- Basic linear algebra, probability, algorithms
- Programming skills

Assessment

- Questions during the lectures
- 3 homeworks

Books

Comprehensive notes for the course will be provided at
`http://www.cs.columbia.edu/~mcollins`

Additional useful background:
Jurafsky and Martin:
Speech and Language Processing (2nd Edition)