Natural Language Processing:

Background and Overview

Regina Barzilay and Michael Collins

EECS/CSAIL

September 8, 2005

Course Logistics

Instructor Regina Barzilay, Michael Collins

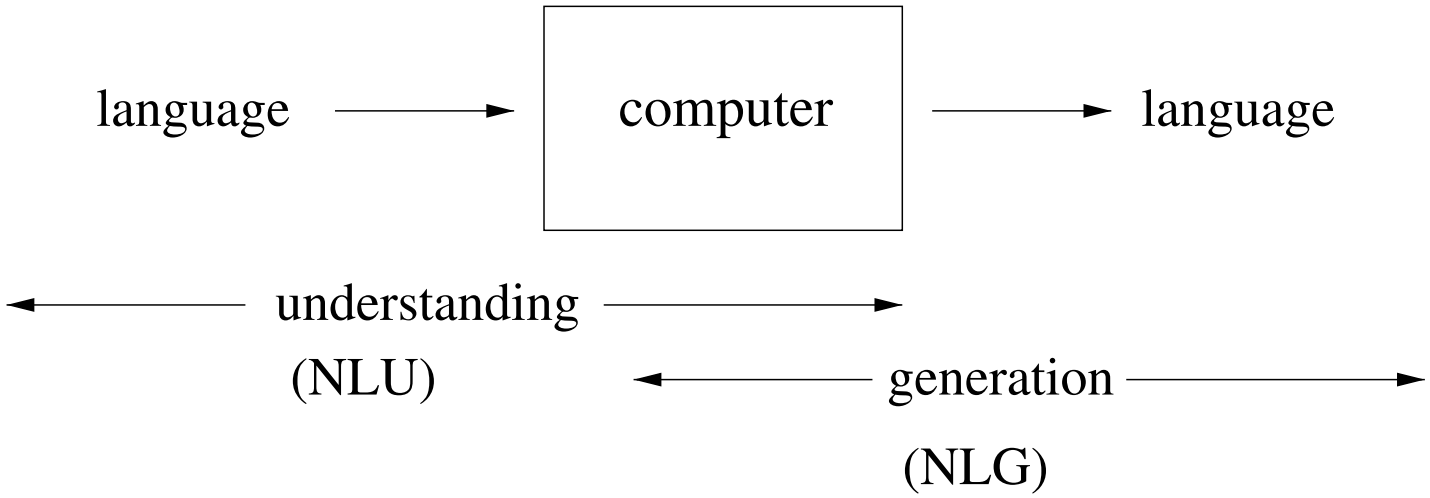
Classes Tues&Thurs 13:00–14:30

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



Alternative Views on NLP

- Computational models of human language processing
  - Programs that operate internally the way humans do
- Computational models of human communication
  - Programs that interact like humans
- Computational systems that efficiently process text and speech

An Online Translation Tool: Input

Tout ce que vous produisez pour credit dans ce cours doit etre votre propre travail. Vous pouvez parler avec les autres etudiants (et les professeurs) de votre approche du probleme, mais ensuite vous devez rsoudre le probleme par vous-meme. Ce n'est pas seulement la facon la plus ethique d'apprendre le contenu de cette classe, mais aussi la plus efficace.

An Online Translation Tool: Output

All that you produce for credit in this course must be your own work. You can speak with the other students (and the professors) about your approach about the problem, but then you must solve the same problem by you. It is not only the most ethical way to learn the contents from this class, but also most effective.

Original

Everything you do for credit in this subject is supposed to be your own work. You can talk to other students (and instructors) about approaches to problems, but then you should sit down and do the problem yourself. This is not only the ethical way but also the only effective way of learning the material.

MIT Translation System

Fifa Will Severely Punish Football Pitches of Deceptive Acts

Text removed for copyright reasons.

Information Extraction

Firm XYZ is a full service advertising agency specializing in direct and interactive marketing. Located in Bigtown CA, Firm XYZ 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: Firm XYZ

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Bigtown, CA.
COMPANY	Firm XYZ
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 ("Does the number of jobs in accounting increases over the years?")

NLP Applications: 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)

ATIS example

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

Other NLP Applications

- Grammar Checking
- Sentiment Classification
- Report Generation
- . . .

Why is NLP Hard?

"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?

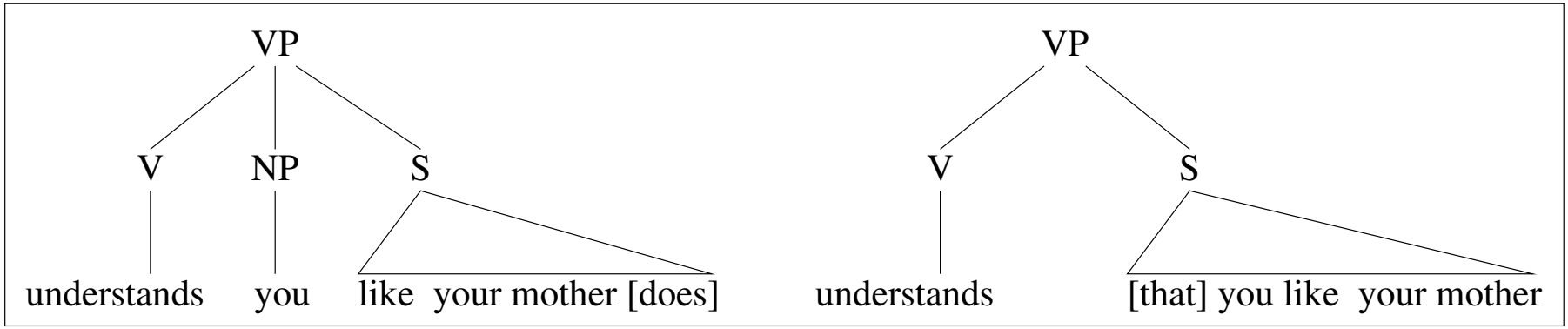
Ambiguity at Many Levels

At the acoustic level (speech recognition):

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

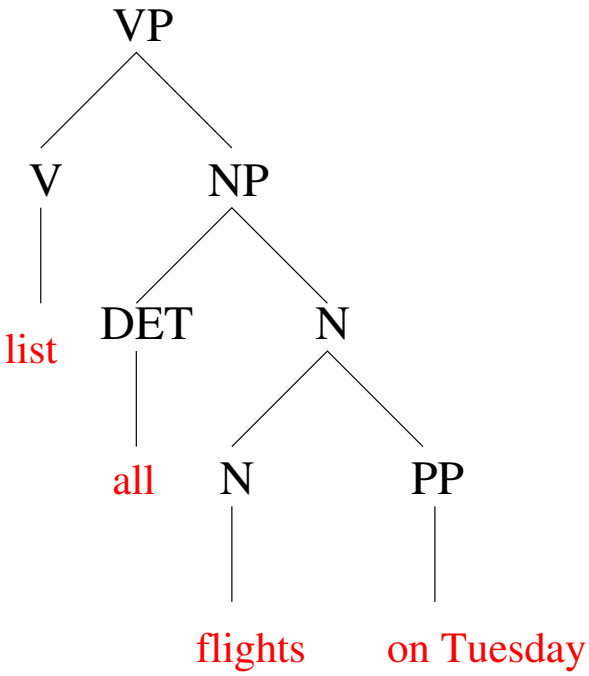
Ambiguity at Many Levels

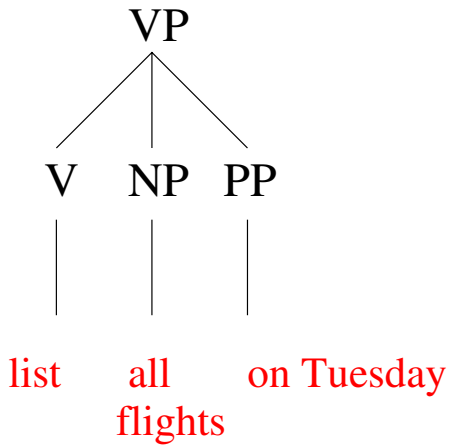
At the syntactic level:



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

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

Text removed for copyright reasons.

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, ...
  - Context-dependent knowledge: co-reference, ...
- 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
  $p(determiner|noun) = \frac{freq(noun,determiner)}{freq(noun)}$
- 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)

Noun	**plural?**	**first appearance**	**determiner**
defendant	no	yes	the
cars	yes	no	null
FBI	no	no	the
concert	no	yes	a

Goal: Learn classification function that can predict unseen examples

Classification Approach

- Learn a function from $X \rightarrow Y$ (in the previous example, $\{\text{“}the''\text{,''}\ a'', null\}$)
- 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 $D(X|Y)$

Beyond Classification

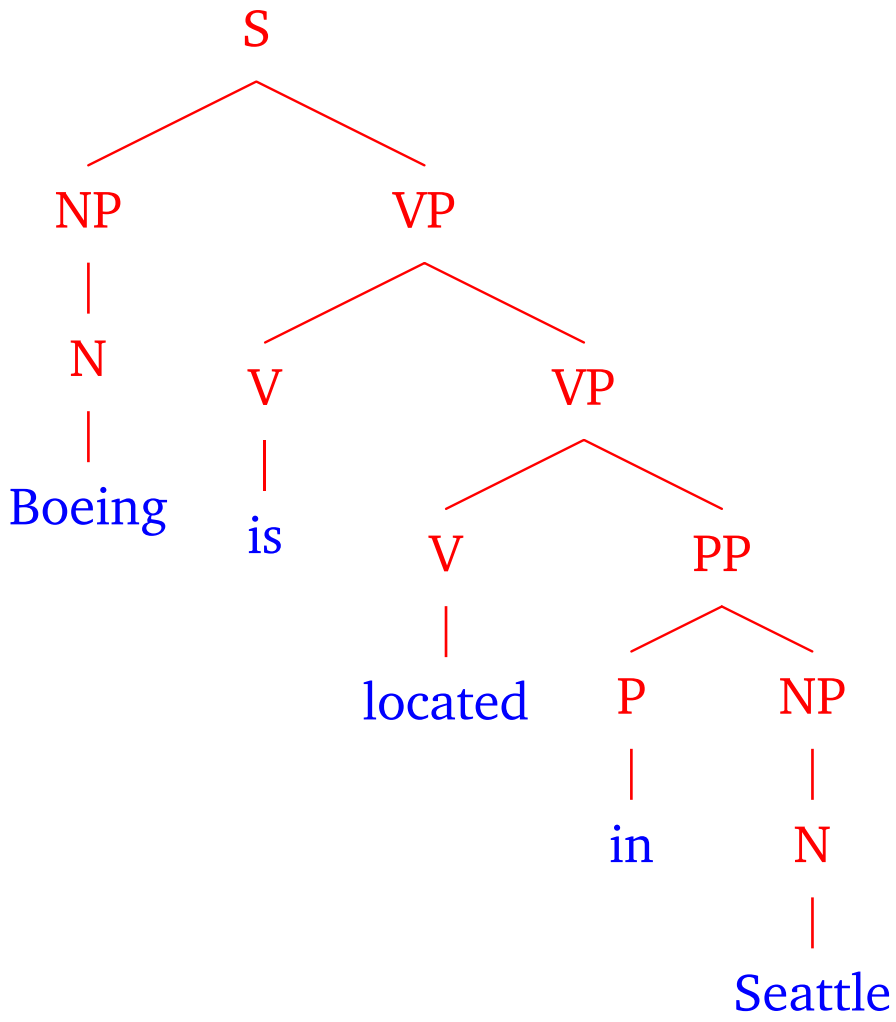
Many NLP applications can be viewed as a mapping from one complex set to another:

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

Classification framework is not suitable in these cases!

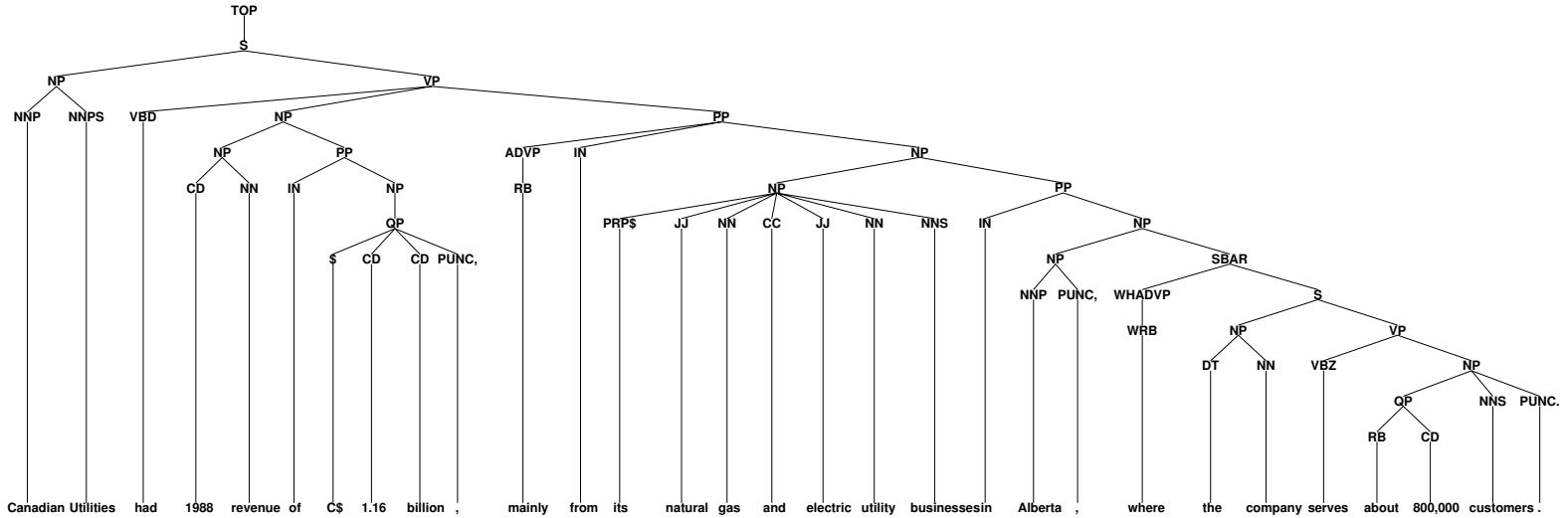
Parsing (Syntactic Structure)

Boeing is located in Seattle.



Data for Parsing Experiments

- 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 .

Example: Machine Translation

Он благополучно избегнул встречи с своею хозяйкой на лестнице. He had successfully avoided meeting his landlady on the staircase.
Каморка его приходилась под самою кровлей высокого пятиэтажного дома и походила более на шкаф, чем на квартиру. His garret was under the roof of a high, five-storied house and was more like a cupboard than a room.
Квартирная же хозяйка его, у которой он нанимал эту каморку с обедом и прислугой, помещалась одною лестницей ниже, в отдельной квартире. The landlady who provided him with garret, dinners, and attendance, lived on the floor below.

Mapping in Machine Translation

"... one naturally wonders if the problem of translation could conceivably be treated as a problem of cryptography. When I look at an article in Russian, I say: 'this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' " (Weaver 1955)

Source: Weaver, W. "Translation" (1955). In *Readings in Machine Translation*.
Edited by S. Nirenburg, H. L. Somers and Y. A. Wilks. Cambridge MA: MIT Press, 2003.
ISBN: 0262140748

Learning for MT

- Parallel corpora are available in several language pairs
- Basic idea: use a parallel corpus as a training set of translation examples
- Goal: learn a function that maps a string in a source language to a string in a target language

Unsupervised Methods

Find patterns in unannotated data

- Word Segmentation

- Grouping of words based on their syntactic/semantic properties
- Grammar Induction

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

Jurafsky, David, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Upper Saddle River, NJ: Prentice-Hall, 2000. ISBN: 0130950696.

Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press, 1999. ISBN: 0262133601.

Prerequisites

- Interest in language and basic knowledge of English
- Some basic linear algebra, probability, algorithms at the level at the level of 6.0465F
- Some programming skills

Assessment

- Midterm (20%)
- Final (30%)
- 6-7 homeworks (50%)

Todo

- Check the class webpage
- Reading: Lillian Lee *"I'm sorry Dave, I'm afraid I can't do that: Linguistics, Statistics, and Natural Language Processing circa 2001"*

Next lecture (9/16th): Syntax and Parsing