

# Intro to Natural Language Processing

Mausam

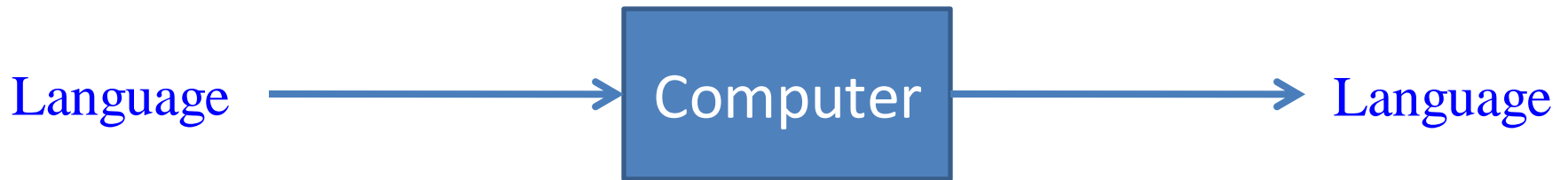
(Based on slides by Rada Mihalcea,  
Chris Manning, Luke Zettlemoyer,  
Mari Ostendorf, Regina Barzilay)

# Natural?

- **Natural Language?**
  - Refers to the language spoken by people, e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc.
- **Natural Language Processing**
  - Applications that deal with natural language in a way or another
- **Computational Linguistics**
  - Doing linguistics on computers
  - More on the linguistic side than NLP, but closely related

# What is NLP?

Computer processing of human language



NL Understanding and Info Extraction

NL Generation

Machine Translation, Paraphrasing, Summarization

# NLP is AI Complete

## Turing Test

**young woman:** Men are all alike.

**eliza:** In what way?

**young woman:** They're always bugging us about something specific or other.

**eliza:** Can you think of a specific example?

**young woman:** Well, my boyfriend made me come here.

**eliza:** Your boyfriend made you come here?

ELIZA (Weizenbaum, 1966): first computer dialogue system based on keyword matching

# Why Natural Language Processing?

- HUGE amounts of data
  - Web: ~1 trillion URLs
  - Intranet
- Applications for processing large amounts of texts
  - require NLP expertise**
- Classify text into categories
- Index and search large texts
- Automatic translation
- Speech understanding
  - Understand phone conversations
- Information extraction
  - Extract useful information from text
- Automatic summarization
  - Condense 1 book into 1 page
- Question answering
- Knowledge acquisition
- Text generations / dialogues

# Some Applications

- Yahoo, Google, Microsoft
  - Information Retrieval
- Monster.com, HotJobs.com (Job finders)
  - Information Extraction + Information Retrieval
- Systran powers BabelFish
  - Machine Translation
- Ask Jeeves
  - Question Answering
- Myspace, Facebook, Blogspot
  - Processing of User-Generated Content
- Tools for “business intelligence”
- All “Big Guys” have (several) strong NLP research labs:
  - IBM, Microsoft, AT&T, Xerox, Sun, etc.
- Academia: research in an university environment

# Web Search ... n.0



find all web pages containing  
the word Liebermann

read the last 3 months of  
the NY Times and provide  
a summary of the campaign  
so far

# Syntax

*“the dog ate my homework”* - Who did what?

## 1. Identify the part of speech (POS)

Dog = noun ; ate = verb ; homework = noun

English POS tagging: 95%

## 2. Identify collocations

mother in law, hot dog

Compositional versus non-compositional  
collocates



# Syntax

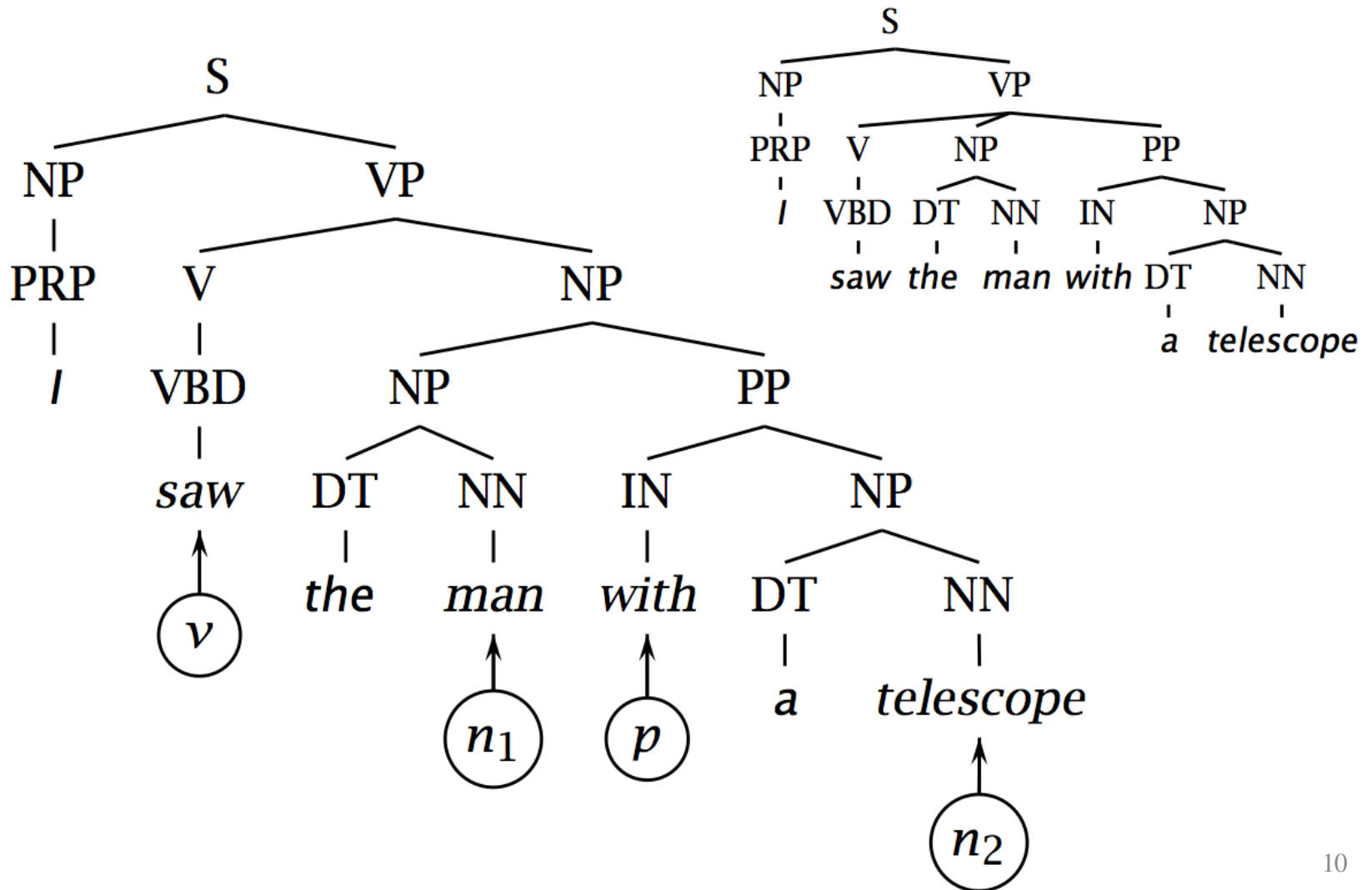
- Anaphora Resolution:

*“The dog entered my room. It scared me”*

- Preposition Attachment

“I saw the man in the park with a telescope”

# Syntactic Parsing



# Syntax vs. Semantics

- (1) Colorless green ideas sleep furiously.
- (2) Furiously sleep ideas green colorless.

“...(1), though nonsensical, is grammatical, while (2) is not.” (Chomsky 1957)

# Issues in Semantics

- Understand language! How?
- “*plant*” = *industrial plant*
- “*plant*” = *living organism*
- Words are ambiguous
- Importance of semantics?
  - Machine Translation: wrong translations
  - Information Retrieval: wrong information
  - Anaphora Resolution: wrong referents

# Ambiguity

I made her duck.

- Possible interpretations of the text out of context:
  - I cooked waterfowl for her.
  - I cooked the waterfowl that is on the plate in front of her.
  - I created a toy (or decorative) waterfowl for her.
  - I caused her to quickly lower her head.

# Why Semantics?

- The sea is at the home for billions factories and animals
- The sea is home to million of plants and animals
- English → French [commercial MT system]
- Le mer est a la maison de billion des usines et des animaux
- French → English

# Why are these funny?

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

# Information Retrieval

- General model:
  - A huge collection of texts
  - A query
- Task: find documents that are relevant to the given query
- How? Create an index, like the index in a book
  
- Retrieve specific information: **Question Answering**
- “What is the height of mount Everest?”
- 11,000 feet



# Information Extraction

- “There was a group of about 8-9 people close to the entrance on Highway 75”
- Who? “8-9 people”
- Where? “highway 75”
  
- Extract information
- Detect new patterns:
  - Detect hacking / hidden information / etc.
- Gov./mil. puts lots of money put into IE research

# Even More

- Discourse
- Summarization
- Subjectivity and sentiment analysis
- Text generation, dialog [pass the Turing test for some million dollars] – Loebner prize
- Knowledge acquisition [how to get that common sense knowledge]
- Speech processing
- ...
- [www.cs.washington.edu/research/textrunner](http://www.cs.washington.edu/research/textrunner)