



# Introduction to NLP

What is Natural Language Processing?



## Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S  
"AN ACCOUNT OF THE PRINCIPALITIES OF  
WALLACHIA AND MOLDOVIA"  
INSPIRED THIS AUTHOR'S  
MOST FAMOUS NOVEL



Bram Stoker

Dan Jurafsky



## Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

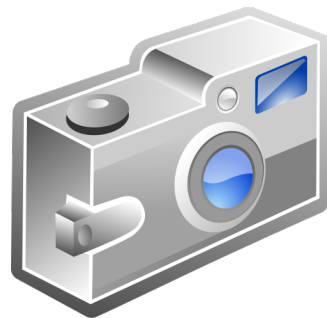
It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry



# Information Extraction & Sentiment Analysis



Attributes:

zoom  
affordability  
size and weight  
flash  
ease of use

Size and weight

- ✓ • nice and compact to carry!
- ✓ • since the camera is small and light, I around those heavy, bulky professio
- ✗ • the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera





# Machine Translation

- Fully automatic

Enter Source Text:

这不过是一个时间的问题。

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators

Enter Source Text:

تعرض الرئيس اللبناني اميل لحود لـ حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت الي " محاكمة " لـ رئيس الجمهورية علي موقفه من المحكمة الدولية و " الملاحظات " التي ادلي بها حول هذا الموضوع .

Translate Clear

Enter Translation:

lebanese

- president
- suffered
- exposed
- president emile
- before
- presented
- offer

Done!



# Language Technology

making good progress

mostly solved

## Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

## Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

## Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

## Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

## Coreference resolution

Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

## Parsing

I can see Alcatraz from the window!

## Machine translation (MT)

第13届上海国际电影节开幕...

The 13<sup>th</sup> Shanghai International Film Festival...

## Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party  
May 27  
add

still really hard

## Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

## Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

## Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

## Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Dan Jurafsky



# Ambiguity makes NLP hard: “Crash blossoms”



Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Dan Jurafsky



# Ambiguity is pervasive

*New York Times* headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%





## In-video quizzes!

- Most lectures will include a little quiz
  - Just to check basic understanding
  - Simple, multiple-choice.
  - You can retake them if you get them wrong



# Why else is natural language understanding difficult?

## non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

## segmentation issues

the New York-New Haven Railroad  
the New York-New Haven Railroad

## idioms

dark horse  
get cold feet  
lose face  
throw in the towel

## neologisms

unfriend  
Retweet  
bromance

## world knowledge

Mary and Sue are sisters.  
Mary and Sue are mothers.

## tricky entity names

Where is *A Bug's Life* playing ...  
*Let It Be* was recorded ...  
... a mutation on the *for* gene ...

But that's what makes it fun!



## Making progress on this problem...

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - $P(\text{"maison"} \rightarrow \text{"house"})$  **high**
    - $P(\text{"L'avocat g n ral"} \rightarrow \text{"the general avocado"})$  **low**
  - Luckily, rough text features can often do half the job.

Dan Jurafsky



# This class

- Teaches key theory and methods for statistical NLP:
  - Viterbi
  - Naïve Bayes, Maxent classifiers
  - N-gram language modeling
  - Statistical Parsing
  - Inverted index, tf-idf, vector models of meaning
- For practical, robust real-world applications
  - Information extraction
  - Spelling correction
  - Information retrieval
  - Sentiment analysis

Dan Jurafsky



## Skills you'll need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Java or Python programming
  - Weekly programming assignments



# Introduction to NLP

# What is Natural Language Processing?