CMP462: Natural Language Processing

Lecture 00: Introduction

Mohamed Alaa El-Dien Aly
Computer Engineering Department
Cairo University
Spring 2013

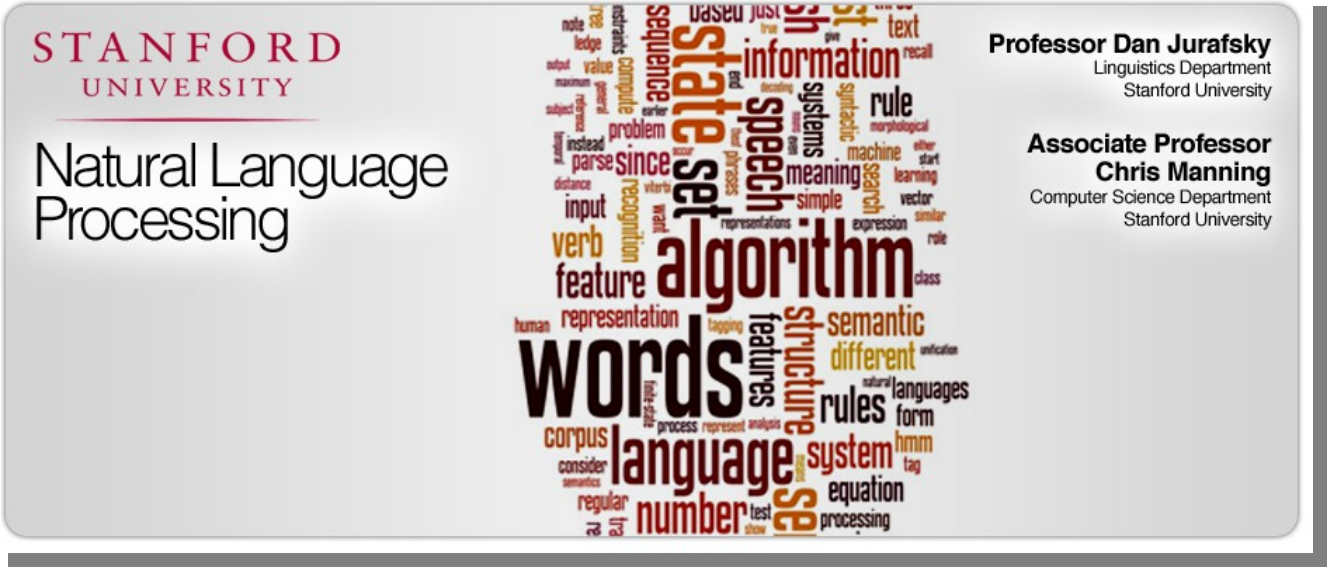
Agenda

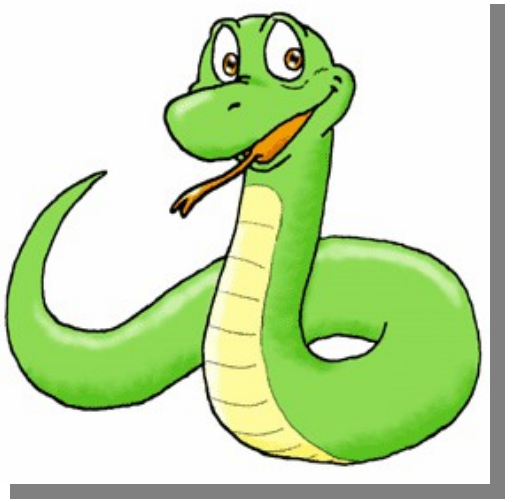
- Course Information
- Topics covered
- Logistics
- Introduction to NLP

Acknowledgment:
Most slides adapted from Chris Manning and Dan Jurafsky's NLP class on Coursera.

Course Information

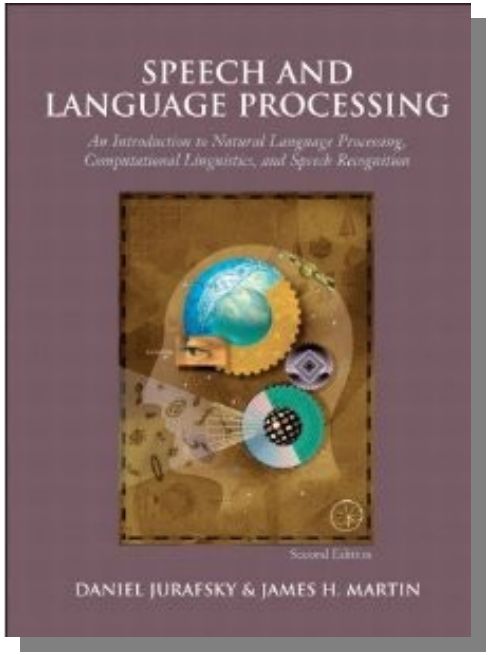
- Will cover an introduction to Natural Language Processing
- Follow closely the NLP class of Coursera
- Weekly programming assignments
- Work with Python this time

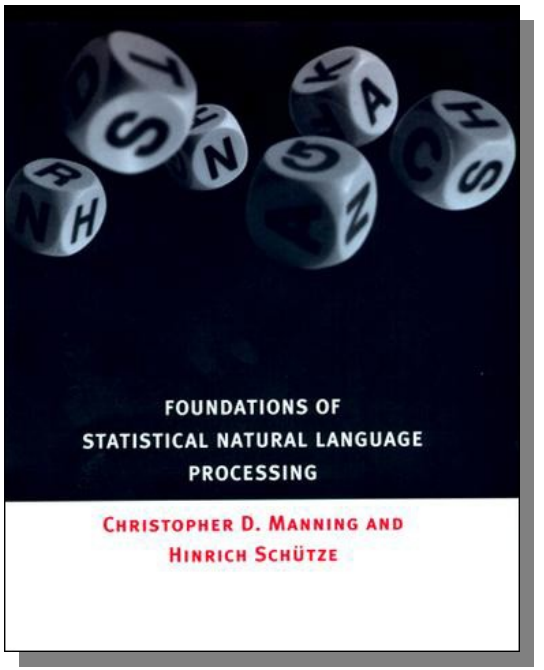




Course Information

- Textbooks:
  - *Speech and Language Processing*. Jurafsky and Martin.
  - *Foundations of Statistical Natural Language Processing*. Manning and Scheutze.





Class Topics

- Basic Text Processing
- Language Modeling
- Text Classification
- Sentiment Analysis
- Named Entity Recognition
- Relation Extraction
- Parts-of-Speech (POS) Tagging
- Parsing
- Information Retrieval

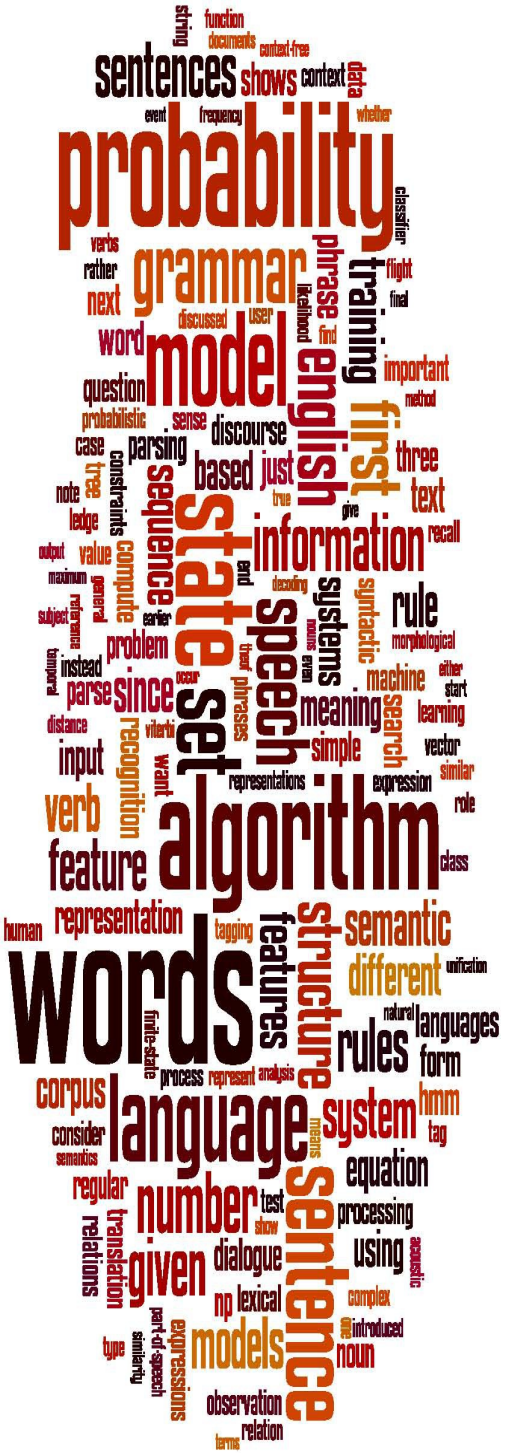
Logistics

- Grades
  - 30 points: weekly* programming assignments
  - 2 points: midterm
- Homeworks
  - Submitted to Moodle
  - NO late homeworks accepted
  - ZERO grade for copying
  - Collaboration OK, but solutions are written INDIVIDUALLY

Resources

- Class:
  - Textbooks
  - Slides
- Python
  - Python's homepage, tutorials, …: www.python.org
  - Spyder: a Python IDE
    - Linux, Windows: code.google.com/p/spyderlib
    - Windows: www.pythonxy.com
  - Natural Language Toolkit in Python with online book: www.nltk.org

Questions?



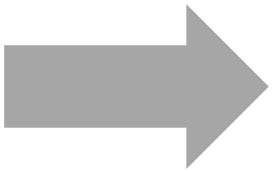
Introduction to NLP

What is Natural Language Processing?

Applications: 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

Dracula !!

Applications: Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Juraf

Hi Dan, we've now scheduled the curriculum m

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

-Chris

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Create new Calendar entry

Applications: 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 wor…nd those heavy, bulky professional cameras
- ✗ the camera feels flimsy, is plastic and ve…haveto be very delicate in the handling of this c…

Applications: Machine Translation

- Fully automatic

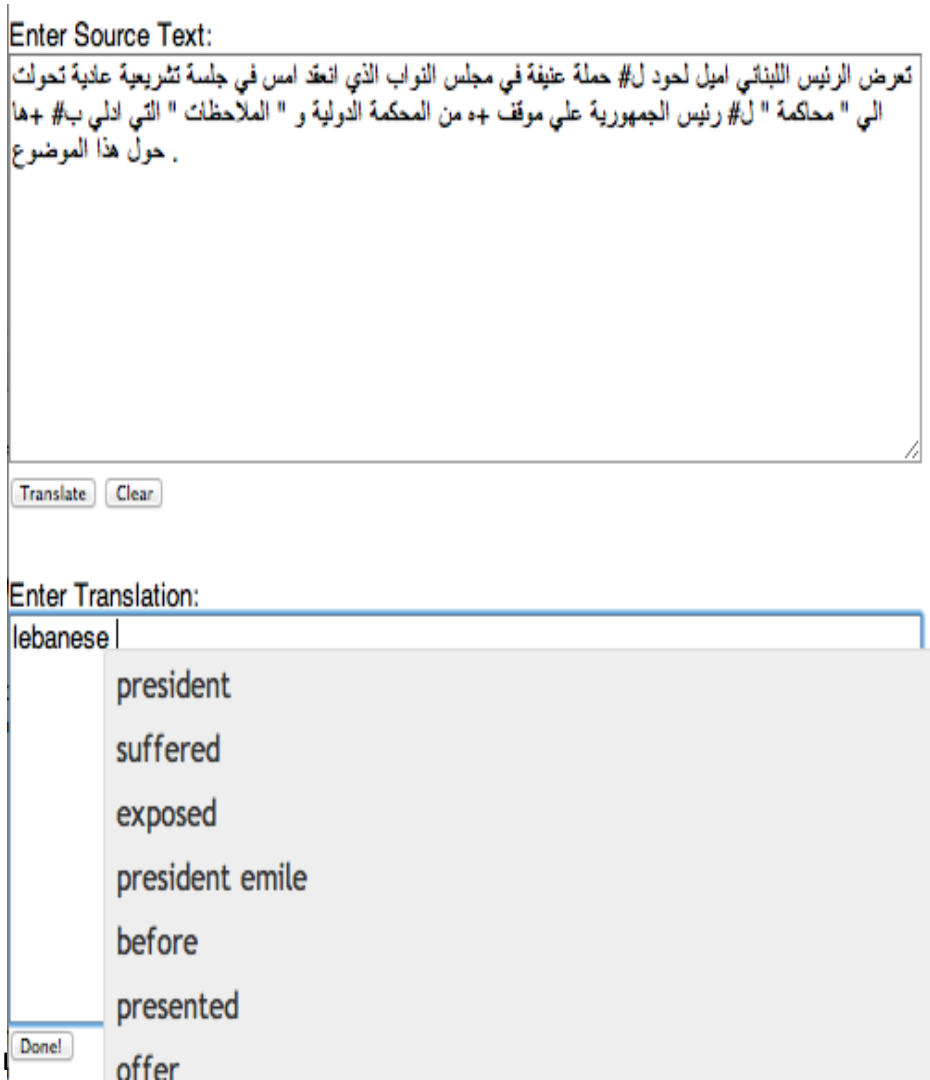
Enter Source Text:

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



Language Technology

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

making good progress

Sentiment analysis

Best roast chicken in San Francisco! ✓

The waiter ignored us for 20 minutes. ✗

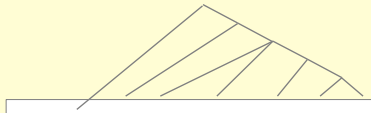
Coreference resolution

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 13th 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?

Ambiguity makes NLP hard: "Crash blossoms"

"Headline with two meanings"

Violinist Linked to JAL Crash Blossoms	Main verb *Linked* vs. *Blossoms*
Teacher Strikes Idle Kids	Main verb *Strikes* vs. *Idle*
Red Tape Holds Up New Bridges	WSD Holds up: *delay* vs *support*
Hospitals Are Sued by 7 Foot Doctors	… etc.
Juvenile Court to Try Shooting Defendant	
Local High School Dropouts Cut in Half	

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("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
  - Luckily, rough text features can often do half the job.

Recap

- Course Information
- Logistics
- Resources
- Introduction to NLP