Introduction

Thanks to Dan Jurafsky for reuse of (some) slides!
What is NLP?

• We study what it takes to make computers perform useful and interesting tasks involving human language
• We are also interested in the insights that we can gain about human language from the study of computational models
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
Hi Dan, we’ve now scheduled the curriculum meeting. It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris
Size and weight

✓ • nice and compact to carry!
✓ • since the camera is small and light, I won’t need to carry around those heavy, bulky professional cameras either!
✗ • 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
- Helping human translators

Enter Source Text:

这不过是一个时间的问题.

Translation from Google Translate:

This is just a question of time.
## Language Technology

### mostly solved

**Spam detection**
- Let’s go to Agra! ✓
- Buy VIAGRA ... ✗

**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

### 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?

---

**UPPSALA UNIVERSITY**
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
Why else is natural language understanding difficult?

**non-standard English**
Great job @justinbieber! Were SOO PROUD of what you've 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
    - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$ low
  - Luckily, rough text features can often do half the job.
This course

- We focus on key theory and methods for NLP:
  - Regular expressions and finite automata
  - Probability and statistical inference
  - Statistical models of morphology, syntax and semantics
- You should also get an orientation about NLP applications
- Main textbook:
  - Jurafsky and Martin, *Speech and Language Processing*
<table>
<thead>
<tr>
<th>Week</th>
<th>Theory/Models</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>Finite automata</td>
<td>Basic text processing</td>
</tr>
<tr>
<td>W2</td>
<td>Probability theory</td>
<td>Language modeling</td>
</tr>
<tr>
<td>W3</td>
<td>Sequence models</td>
<td>Part-of-speech tagging</td>
</tr>
<tr>
<td>W4</td>
<td>Tree models</td>
<td>Syntactic parsing</td>
</tr>
<tr>
<td>W5</td>
<td>Models of meaning</td>
<td>Computational semantics</td>
</tr>
</tbody>
</table>

*W6 is reserved for oral presentations of application areas*
Flipped-classroom teaching

• Traditional (university) teaching:
  • Students are passive in the classroom (lectures)
  • Students are active on their own (homework, assignments, exam prep)

• The flipped classroom:
  • Students are active in the classroom (problem solving)
  • Students are passive on their own (lectures) but also active (quizzes)

• This course will not be completely flipped
  • Classroom activity is key – and requires preparation
Examination and assignments

• First five weeks:
  • Two smaller in-class assignments (obligatory attendance)
  • One hand-in assignment based on the in-class assignments

• Final week:
  • Oral presentation of application area
  • Written literature review on same area

• Don’t leave the application survey until the last week!
Application areas

• Some suggestions:
  • Information extraction
  • Question answering and summarisation
  • Dialogue and conversational agents
  • Machine translation
  • Text generation and summarisation
  • Opinion mining and sentiment analysis
  • Authoring tools (spell checking, grammar checking)

• Think about it!

Ch. 22–25 in textbook
For in-class assignments we will often use Python
  • You don’t need to know Python to take the course
  • But you should make sure you know how to run a Python program
  • And you will probably learn some Python in the process

For hand-in assignments you can use any language you like
  • But since they are based on in-class assignments, it is often most convenient to use Python here as well