Lecture 1:
Introduction
Course Staff

Professor:  
Julia Hockenmaier  juliahmr@illinois.edu

Teaching assistants:  
Dhruv Agarwal  dhruva2@illinois.edu  
Sai Krishna Bollam  sbollam2@illinois.edu  
Zubin Pahuja  zpahuja2@illinois.edu
Today’s lecture

Course Overview:
   What is NLP?
   What will you learn in this course?

Course Admin:
   How will we teach this course?
   How will you be assessed in this course?
What is Natural Language Processing?
What is Natural Language Processing really?
NLP in the news...

Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future

Computers can now describe images using language you'd understand

Barbie Wants to Get to Know Your Child

With the help of A.I., America’s most famous doll tries to fulfill a timeless dream — convincing little girls that she’s a real friend. What will happen if they believe her?
IBM’s Watson wins at Jeopardy!
Machine Translation

Google Translate

Xi Jinping held talks with Zimbabwean President Robert Mugabe

Nakatsu do emphasize equality, mutual support, mutual benefit and common development, good friends, good partners and good brothers

Xinhua Beijing, August 25 (Reporter Tan Jingjing) President Xi Jinping held talks on the 25th at the Great Hall with President Robert Mugabe. Xi spoke highly of the traditional friendship between the important contribution Zimbabwe and Mugabe made for the development of bilateral relations, stressing the friendship of the Chinese people is heavy, we will never forget once stood, mutual understanding and support of old friends. China is willing to Tianjin square one, carry forward the traditional friendship and strengthen cooperation in various fields, doing equal treatment, mutual support, mutual benefit and common development of good friends, good partners and good brothers.

Xi pointed out that President Mugabe is a famous leader of the national liberation movement in Africa, an important promoter of African integration, is also an old friend of the Chinese people. Nakatsu traditional friendship is condensed in our glory days alongside imperialism and colonialism and hegemonism, and reflects both countries followed independence, mutual respect, and oppose foreign interference and other basic principles of foreign relations. Establishment of diplomatic ties 34 years ago, the two sides understand each other on issues concerning each other's core interests and major concerns, mutual support, mutual help in the development process. Nakata traditional friendship is the common treasure of the two countries, we both cherish. Currently, Nakata are at an important stage of nation-building, friendship and cooperation between the two countries are embracing new opportunities. China is willing to work together with Tianjin efforts to promote comprehensive development of bilateral relations, to better benefit the two peoples.
Dialog systems, chatbots, digital assistants
The same follow-up questions that Siri couldn’t answer in 2014 receive appropriate responses when posed to Siri in 2017. For example, when asked if Cleverbot’s favorite show was Doctor Who, it responded with a simple statement: “Yes, so funny.” In contrast, Forbes-Riley and Litman’s (2011) student was able to follow the conversation and ask more detailed questions, such as “Why did you choose to work for Google?”

We begin in the next section with chatbots, introducing the three major chatbot architectures: rule-based systems, information retrieval systems, and transduction architectures (the architecture) that underlies most modern task-based dialog systems, as shown in Fig. 29.2. Figure 29.3 shows the tutoring domain: An excerpt from a tutorial interaction between a physics student and the ITSPOKE system of the VoiceXML system that is a simple industrial instantiation of the architecture, as described in Carpenter (2017).

As an illustration of the tutoring domain, consider the following discussion between a physics student and the Tutor system:

**Student:** What will the velocity of the object be a second after that (where the initial velocity is 9.8 m/s and the acceleration is 9.8 m/s)?

**Tutor:** Right. Now back to the man and his keys. We’ve shown that they both have an acceleration of 9.8 m/s². So what’ll its velocity be, a second after that?

**Student:** Is it 19.6 m/s?

**Tutor:** Good. A second later, its velocity will be 9.8 m/s + 9.8 m/s = 19.6 m/s. This is because its acceleration tells us that every second, its velocity increases by 9.8 m/s. So, how do their velocities compare with each other, at every second during the fall?
Sentiment/opinion analysis

Customer Reviews
Haier HLT71 7-Inch Handheld LCD TV by Haier

Average Customer Rating
⭐⭐⭐⭐⭐ (688 customer reviews)

- Image quality: ⭐⭐⭐⭐⭐ (359)
- Portability: ⭐⭐⭐⭐⭐ (359)
- Ease of use: ⭐⭐⭐⭐⭐ (356)
- Features: ⭐⭐⭐⭐⭐ (353)

The Most Helpful Reviews

The most helpful favorable review:
⭐⭐⭐⭐⭐ FANTASTIC 7 inch portable LCD TV!
I am really thrilled with my purchase of this Haier HLT71 7-inch, ATSC 2009 portable LCD TV! I have never heard of the name brand Haier, and I had heard some nightmare stories about some of the name brands having poor pictures on little screens, so I thought I'd give this Haier a try.

I ordered this tv directly from Amazon and was delivered fast and in great condition.

Source: www.amazon.com
Have chatbots developed their own language?

https://code.facebook.com/posts/1686672014972296/deal-or-no-deal-training-ai-bots-to-negotiate/

Bob: “I can can I I everything else.”
Alice: “Balls have zero to me to me to me to me to me to me to me to me to me to.”

No, not really…
http://www.snopes.com/facebook-ai-developed-own-language/
How well does that Barbie do?

Barbie: “Do you have any sisters?”
Child: “Yeah, I only have one.”
Barbie: “What’s something nice that your sister does for you?”
Child: “She does nothing nice to me”
Barbie: “Well, what is the last nice thing your sister did?”
Child: “She helped me with my project — and then she destroyed it.”
Barbie: “Oh, yeah, tell me more!”
Child: “That’s it, Barbie,”
Barbie: “Have you told your sister lately how cool she is?”
Child: “No. She is not cool,”
Barbie: “You never know, she might appreciate hearing it”

(Thanks to Barbara Grosz for pointing out the example from the NYT)
What is the current state of NLP?

Lots of commercial applications and interest.
   Some applications are working pretty well already, others not so much.

A lot of hype around “deep learning” and “AI”
   — Neural nets are powerful classifiers and sequence models
   — Public libraries (Tensorflow, Torch, Caffe, etc.) and datasets make it easy for anybody to get a model up and running
   — “End-to-end” models put into question whether we still need the traditional NLP pipeline that this class is built around
   — We’re still in the middle of this paradigm shift
   — But many of the fundamental problems haven’t gone away
What will you learn in this class?
The topics of this class

We want to identify the **structure and meaning** of **words, sentences, texts and conversations**

N.B.: we do not deal with speech (no signal processing)

We mainly deal with **language analysis/understanding**, and less with language generation/production

We focus on **fundamental concepts, methods, models, and algorithms**, not so much on current research:

- Data (natural language): linguistic concepts and phenomena
- Representations: grammars, automata, etc.
- Statistical models over these representations
- Learning & inference algorithms for these models
What you should learn

You should be able to answer the following questions:
- What makes natural language difficult for computers?
- What are the core NLP tasks?
- What are the main modeling techniques used in NLP?

We won’t be able to cover the latest research…
   (this requires more time, and a much stronger background in machine learning than I am able to assume for this class)

… but I would still like you to get an understanding of:
- How well does current NLP technology work (or not)?
- What NLP software is available?
- How to read NLP research papers [4 credits section]
Building a computer that ‘understands’ text: The NLP pipeline
新华社拉萨二月二日电（记者央珍）“八五”（一九九一至一九九五年）期间，西藏金融体制改革坚持与全国框架一致、体制衔接的方针，顺利完成西藏各级人民银行的分设工作，实现信贷资金使用从粗放型经营方式向集约型经营方式转变。
Task: Tokenization/segmentation

We need to split text into words and sentences.
- Languages like Chinese don’t have spaces between words.
- Even in English, this cannot be done deterministically:
  
  *There was an earthquake near D.C. You could even feel it in Philadelphia, New York, etc.*

NLP task:
What is the *most likely* segmentation/tokenization?
Task: Part-of-speech-tagging

Open the pod door, Hal.

Verb  Det  Noun Noun ,  Name .

Open the pod door , Hal .

open:
verb, adjective, or noun?
Verb: open the door
Adjective: the open door
Noun: in the open
How do we decide?

We want to know the most likely tags $T$ for the sentence $S$

$$\arg\max_T P(T|S)$$

We need to define a statistical model of $P(T | S)$, e.g.:

$$\arg\max_T P(T|S) = \arg\max_T P(T)P(S|T)$$

$$P(T) = \text{def} \prod_i P(t_i|t_{i-1})$$

$$P(S|T) = \text{def} \prod_i P(w_i|t_i)$$

We need to estimate the parameters of $P(T | S)$, e.g.:

$$P( t_i =V \mid t_{i-1} =N ) = 0.3$$
Disambiguation requires statistical models

Ambiguity is a core problem for any NLP task

Statistical models* are one of the main tools to deal with ambiguity.

*more generally: a lot of the models (classifiers, structured prediction models) you learn about in CS446 (Machine Learning) can be used for this purpose. You can learn more about the connection to machine learning in CS546 (Machine learning in Natural Language).

These models need to be trained (estimated, learned) before they can be used (tested).

We will see lots of examples in this class (CS446 is NOT a prerequisite for CS447)
“I made her duck”

What does this sentence mean?

“duck”: noun or verb?
“make”: “cook X” or “cause X to do Y”?
“her”: “for her” or “belonging to her”?

Language has different kinds of ambiguity, e.g.:

**Structural ambiguity**

“I eat sushi with tuna” vs. “I eat sushi with chopsticks”
“I saw the man with the telescope on the hill”

**Lexical (word sense) ambiguity**

“I went to the bank”: financial institution or river bank?

**Referential ambiguity**

“John saw Jim. He was drinking coffee.”
“I made her duck cassoulet”

(Cassoulet = a French bean casserole)

The second major problem in NLP is coverage: We will always encounter unfamiliar words and constructions.

Our models need to be able to deal with this.

This means that our models need to be able to generalize from what they have been trained on to what they will be used on.
Task: Syntactic parsing

Open the pod door, Hal.
Observation: Structure corresponds to meaning

Correct analysis

Incorrect analysis
John saw Mary.
I ate sushi with tuna.

I ate the cake that John had made for me yesterday.

John and Mary eat sushi for dinner.

Did you go there?
I want you to go there.

Did you went there?

I ate tuna sushi with.

John Mary saw.
NLP and automata theory

What kind of grammar/automaton is required to analyze natural language?

What class of languages does natural language fall into?

Chomsky (1956)’s hierarchy of formal languages was originally developed to answer (some of) these questions.
Task: Semantic analysis

∃x∃y(pod_door(x) & Hal(y)
& request(open(x, y)))

Open the pod door, Hal.
Representing meaning

We need a meaning representation language.

“Shallow” semantic analysis: Template-filling
(Information Extraction)
   Named-Entity Extraction: Organizations, Locations, Dates,...
   Event Extraction

“Deep” semantic analysis: (Variants of) formal logic
   $\exists x \exists y (\text{pod\_door}(x) \& \text{Hal}(y) \& \text{request(open}(x,y)))$

We also distinguish between
Lexical semantics (the meaning of words) and
Compositional semantics (the meaning of sentences)
∃x∃y(pod\_door(x) & Hal(y) & request(open(x, y)))

Multimodal NLP: mapping from language to the world

request(open(door2, SYS))
More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field -- a multi-millionaire.

Who is Carl Lewis?
Did Carl Lewis break any world records? (and how do you know that?)
Summary: The NLP Pipeline

An NLP system may use some or all of the following steps:

**Tokenizer/Segmenter**
- to identify words and sentences

**Morphological analyzer/POS-tagger**
- to identify the part of speech and structure of words

**Word sense disambiguation**
- to identify the meaning of words

**Syntactic/semantic Parser**
- to obtain the structure and meaning of sentences

**Coreference resolution/discourse model**
- to keep track of the various entities and events mentioned
Course Admin
This class consists of...

... Lectures:
  Wednesdays and Fridays, 12:30pm–1:45 pm, DCL1310

... Office:
  Julia: Wednesdays and Fridays, 2pm–3pm, Siebel 3324
  Dhruv: TBD, Siebel 0207
  Sai Krishna: TBD, Siebel 0207
  Zubin: TBD, online

... Websites:
  Syllabus, slides, policies, etc: http://courses.engr.illinois.edu/cs447
  Discussions: piazza.com/illinois/fall2018/cs447
  Grades, submitting assignments: http://compass2g.illinois.edu

... Readings:
  Textbook + additional readings (http://courses.engr.illinois.edu/cs447)

... Assessment:
  4+1 assignments, 2 exams (4th credit hour: project or survey)
Lectures and office hours

Attend!
Ask questions!
Participate!
Reading

Course website: (slides, reading)
https://courses.engr.illinois.edu/cs447/fa2018/syllabus.html

The textbook: https://web.stanford.edu/~jurafsky/slp3/
Jurafsky and Martin, Speech and Language Processing
(3rd edition PDFs in prep.; 2nd edition, 2008 in print)

For some assignments:
Assessment

If you take this class for 3 hours credit:
   1/3 homework assignments
   1/3 midterm exam
   1/3 final exam

If you take this class for 4 hours credit:
   1/4 homework assignments
   1/4 midterm exam
   1/4 final exam
   1/4 literature review or project

We reserve the right to improve your grade by up to 5% depending on your class participation. If you’re in between grades, but attended class and participated frequently and actively in in-class discussions etc., we will give you the higher grade.
Homework assignments

What?
- 4 assignments (mostly programming), plus homework 0
- We use Python and the Natural Language Toolkit (NLTK)

Why?
- To make sure you can put what you’ve learned to practice.

How?
- You will have one weeks to complete HW0.
- You will have three weeks to complete HW1, HW2, HW3, HW4.
- Grades will be based on your write-up and your code.
- Submit your assignments on Compass.

Late policy?
- **No** late assignments will be accepted (sorry).
Homework assignments

Schedule:
Week 1:   Friday, 08/31   HW0 out
Week 2:   Friday, 09/07   HW0 due, HW1 out
Week 5:   Friday, 09/28   HW1 due, HW2 out
Week 8:   Friday, 10/19   HW2 due, HW3 out
Week 11:  Friday, 11/09   HW3 due, HW4 out
Week 14:  Friday, 12/07   HW4 due

Points per assignment:
HW0 = 2 points
(Did you submit [on time]? Was it in the right format?)
HW1,HW2,HW3,HW4 = 10 points per assignment
Exams

What?

Midterm exam: Friday, Oct 12, in class
Final exam: Wednesday, Dec 12, in class
(based on material after first midterm)

Why?

To make sure you understand what you learned well enough to explain and apply it.

How?

Essay questions and problem questions
Closed-book (no cheatsheets, no electronics, etc.)
Will be based on lectures and readings
4th credit hour: Research Projects

What?
You need to read and describe a few (2–3) NLP papers on a particular task, implement an NLP system for this task and describe it in a written report.

Why?
To make sure you get a deeper knowledge of NLP by reading original papers and by building an actual system.

When?
Fri, Oct 5: Proposal due (What topic? What papers will you read?)
Fri, Nov 9: Progress report due (Are your experiments on track?)
Thu, Dec 13: Final report due (Summary of papers, your system)
4th credit hour: Literature Survey

What?
You need to read and describe several (5-7) NLP papers on a particular task or topic, and produce a written report that compares and critiques these approaches.

Why?
To make sure you get a deeper knowledge of NLP by reading original papers, even if you don’t build an actual system.

When?
Fri, Oct 5: Proposal due (What topic? What papers will you read?)
Fri, Nov 9: Progress report due (Is your paper on track?)
Thu, Dec 13: Final report due (Summary of papers)
Course Outline (tentative)

Lectures 2–5: Morphology, language models
Lectures 7–10: Sequence labeling (POS tagging etc.)
Lectures 11–12: Syntax and Parsing
Lecture 13: Review for midterm

Midterm exam

Lectures 15–18: Semantics
Lectures 19–22: Machine Translation
Lectures 23–24: Discourse, Dialog
Lectures 25–27: Neural NLP
Lecture 28: Review for Final Exam

Final exam
Today’s readings

Today’s lecture:

Jurafsky and Martin Chapter 1 (2nd edition)
http://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf