Lecture 1: Introduction

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

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?

IBM’s Watson wins at Jeopardy!

Machine Translation
Dialog systems

Sentiment/opinion analysis

The topics of this class

What will you learn in 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, not 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

Source: www.amazon.com
Building a computer that ‘understands’ text: The NLP pipeline

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
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)

How do we decide?

We want to know the most likely tags \( T \) for the sentence \( S \)

\[
\text{argmax}_T P(T | S)
\]

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

\[
\text{argmax}_T P(T) = \text{argmax}_T P(T | S) P(S | T)
\]

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

\[
P(S | T) = \text{def} \prod_i P(w_i | i)
\]

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

\[
P(t_i = V | t_{i-1} = N) = 0.3
\]

“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:
- S
  - VP
    - NP
      - NOUN
      - Noun
    - NP
      - Name.

Incorrect analysis:
- S
  - VP
    - NP
      - Noun
      - Verb
    - NP
      - Det
      - Nouns

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)))

Verb Det Noun Noun, Name.
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
∃x∃y(pod_door(x) & Hal(y) & request(open(x, y)))

We also distinguish between
- Lexical semantics (the meaning of words) and
- Compositional semantics (the meaning of sentences)

Multimodal NLP: mapping from language to the world

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

Understanding texts

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

This class consists of...

… Lectures:
Tuesdays and Thursdays, 2:00pm–3:15 pm, SC0216

… Office hours:
Julia: Tuesdays and Thursdays, 3:15pm–4:00pm, SC3324
Chris: Wednesdays, 5–6pm, SC1312; Thursdays, 10–11am, SC 1312
Ryan: Mondays, 1–2pm, SC 1312; Wednesdays, 3–4pm, SC 1312

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

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

… Assessment:
4 Assignments, 2 exams (4th credit hour: project or survey)
Reading

The textbook:
Jurafsky and Martin, *Speech and Language Processing* (2nd edition, 2008)

Other material (Slides, additional reading):
Posted on course website

For the homework:

Homework assignments

What?
4 assignments (mostly programming), plus homework 0
We use Python 2.7 and the Natural Language Toolkit (NLTK)

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

How?
You will have two 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).

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 reviews

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

Schedule:
Week 1: Tue, 08/25 HW0 out (today!)
Week 3: Tue, 09/08 HW0 due
Week 3: Thu, 09/10 HW1 out
Week 6: Thu, 10/01 HW1 due, HW2 out
Week 9: Thu, 10/22 HW2 due, HW3 out
Week 12: Thu, 11/12 HW3 due, HW4 out
Week 15: Tue, 12/08 HW4 due (last lecture)

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: Thursday, Oct 8 (Week 7), in class
Final exam: TBC; tentatively: Wednesday, Dec 16 (morning)
(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?
Oct 6 (Wk 6): Proposal due (What topic? What papers will you read?)
Nov 12 (Wk 12): Progress report due (Are your experiments on track?)
Dec 10 (Reading Day): 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?
Oct 6 (Wk 6): Proposal due (What topic? What papers will you read?)
Nov 12 (Wk 12): Progress report due (Is your paper on track?)
Dec 10 (Reading Day): Final report due (Summary of papers)

Course Outline (tentative)

Lectures 2–5: Words and strings (morphology, language models)
Lectures 7–10: Sequence labeling (POS tagging etc.)
Lectures 11–12: Lexical similarities, word clustering
Lecture 13: Review for midterm
Lecture 14: Midterm exam
Lecture 15–21: Syntax and Parsing
Lecture 22–24: Machine Translation
Lecture 25–28: Semantics, Discourse
Lecture 29: Review for Final Exam

http://courses.engr.illinois.edu/cs447/syllabus.html
Today’s readings

Today’s lecture:
Jurafsky and Martin Chapter 1
http://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf

Homework 0
Due two weeks from today.
https://courses.engr.illinois.edu/CS447/HW/cs447_HW0.zip

Read Python tutorial sections 1—5
https://docs.python.org/2/tutorial/