Introduction to NLP

Ruihong Huang
Texas A&M University

Some slides adapted from slides by
Dan Jurafsky, Luke Zettlemoyer, Ellen Riloff

- "An Aggie does not lie, cheat, or steal or tolerate those who do." For additional information, please visit: http://aggiehonor.tamu.edu.

- Upon accepting admission to Texas A&M University, a student immediately assumes a commitment to uphold the Honor Code, to accept responsibility for learning, and to follow the philosophy and rules of the Honor System. Students will be required to state their commitment on examinations, research papers, and other academic work. Ignorance of the rules does not exclude any member of the TAMU community from the requirements or the processes of the Honor System.

- The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services, currently located in the Disability Services building at the Student Services at White Creek complex on west campus or call 979-845-1637. For additional information, visit http://disability.tamu.edu.

- Piazza: CSCE 489, NLP

- https://piazza.com/class/jla0c036yum1ov

- course page:

- http://faculty.cse.tamu.edu/huangrh/Fall18-489/Fall18_489_nlp_foundation_technique.html

- Class participation: 5%

- Ten in-class quizzes: 20

- Two Programming Assignments: 20%

- Annotation assignment: 10%

- The Final Project: 30% (abstract: 5%, project progress report: 5%, presentation+report+code+data: 20%)

- Final exam: 15% (date: 12/12)

- Late Policy: 20% reduction per day for up to two days. Including programming assignments, annotation assignment, and the final project.

Programming Assignments

- Code: has to be runnable

- Report: how to run, results and analysis, remaining issues, known bugs.

The Final Project

- Due in a month (09/25, before the class starts): 1-page abstract

- Due in two months (11/06, before the class starts): 2-page progress report, already obtained results and a plan of remaining steps.

- By the end of the semester: submit code data and a final report, and a class presentation.

- Final report: 8 pages maximum, full description of the problem, approaches and evaluation results.

Basic Recipe of Forming a Project

- Choose a Topic and do a quick survey

- Prepare data

- Think about evaluation methods

- Start to work on it

The final Project

- Solving a mini core research problem you have identified by reading recent research papers from top NLP conferences.

- Developing a nice NLP application system.

Core research problems

- Semantics, word sense disambiguation

- Coreference resolution, discourse, pragmatics

- Consider to participate in a SemEval task (http://alt.qcri.org/semeval2019/index.php?id=tasks)

Applications

- Question-Answering

- Text Summarization

- Dialogue systems

- Sentiment Analysis

- Machine Translation

- Interdisciplinary applications……

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching
- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, sentiment analysis, question answering…

Question Answering: Jeopardy!

US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.

Information Extraction

Subject: curriculum meeti

Date: January 15, 2012

To: Dan Ju

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

15

Google Knowledge Graph

Text Summarization

- **Condensing documents**
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative

- **Very context-dependent!**

- **An example of analysis with**

WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.
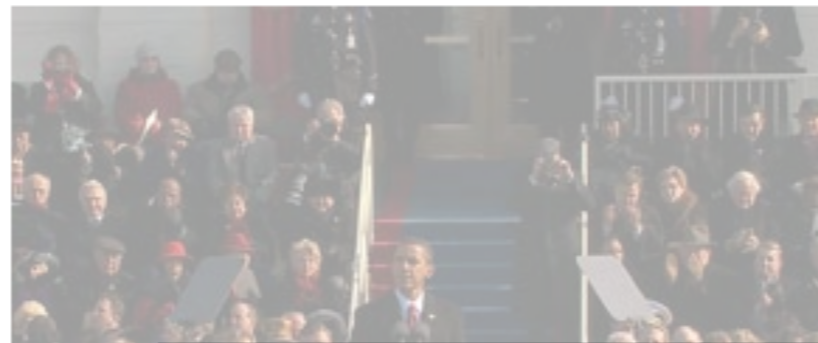
Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must begin

STORY HIGHLIGHTS
- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

said in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Obama, too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our

Human-machine Dialogs

Machine Translation

- Helping human translators

Enter Source Text:

这 不过 是 一 个 时 间 的 问 题 .

Fully automatic

Translation from Stanford's *Phrasal*:

This is only a matter of time.

19

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها حول هذا الموضوع .

Translate   Clear

Enter Translation:

lebanese

president

suffered

exposed

president emile

before

presented

offer

Done!

Inter-Disciplinary

Computer Science: artificial intelligence, machine learning

Linguistics: computational linguistics

Psychology: cognitive psychology, psycholinguistics

Statistics: probabilistic methods, information theory

Interactions with Linguists (History)

- 70s and 80s: more linguistic focus

-deeper models, toy domains, rule-based systems

- 90s: empirical revolution

-robust corpus-based methods, empirical evaluation

- 2000s: richer linguistic representations used in statistical approaches

Outline

- **Bag** of Words: Text classification

- **Sequence** of Words: language modeling, parts of speech tagging

- **Tree** of Words: syntactic parsing, dependency parsing

- **Semantics**: thesaurus, distributional, distributed

- **Discourse**, coreference, pragmatics

Language Technology

making good progress

still really hard

mostly solved

Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Part-of-speech (POS) tagging

ADJ   ADJ   NOUN VERB   ADV

Colorless green ideas sleep furiously.

Parsing

I can see Alcatraz from the window!

Named entity recognition (NER)

PERSON   ORG   LOC

Einstein met with UN officials in Princeton

Machine translation (MT)

第13届上海国际电影节开幕… ➡

The 13th Shanghai International Film Festival…

Information extraction (IE)

Party May 27 add

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

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 !!

Ambiguities inherent in Language

- Language is succinct and expressive.

- Human resolve ambiguities naturally.

Syntax: structural ambiguity

Time flies like an arrow.

Metaphor:
Time/NOUN flies/VERB like/PREP an/ART arrow/NOUN

New Fly Species:
Time/NOUN flies/NOUN like/VERB an/ART arrow/NOUN

Stopwatch Imperative:
Time/VERB flies/NOUN like/PREP an/ART arrow/NOUN

Syntax: structural ambiguity (attachment)

- *I saw the Grand Canyon flying to New York.*

- *I watered the plant with yellow leaves.*

- *I saw the man on the hill with the telescope.*

```
                        S
             ┌──────────┴──────────┐
            NP                     VP
          ┌──┴──┐          ┌───────┴───────┐
         Det    N          V               NP
          │     │          │        ┌───────┼────────┐
          │     │          │       Det      N        PP
          │     │          │        │       │    ┌───┴────┐
         the   dog        saw       a      man   P        NP
                                                 │     ┌───┴───┐
                                                 in   Det      N
                                                      │        │
                                                     the      park
```
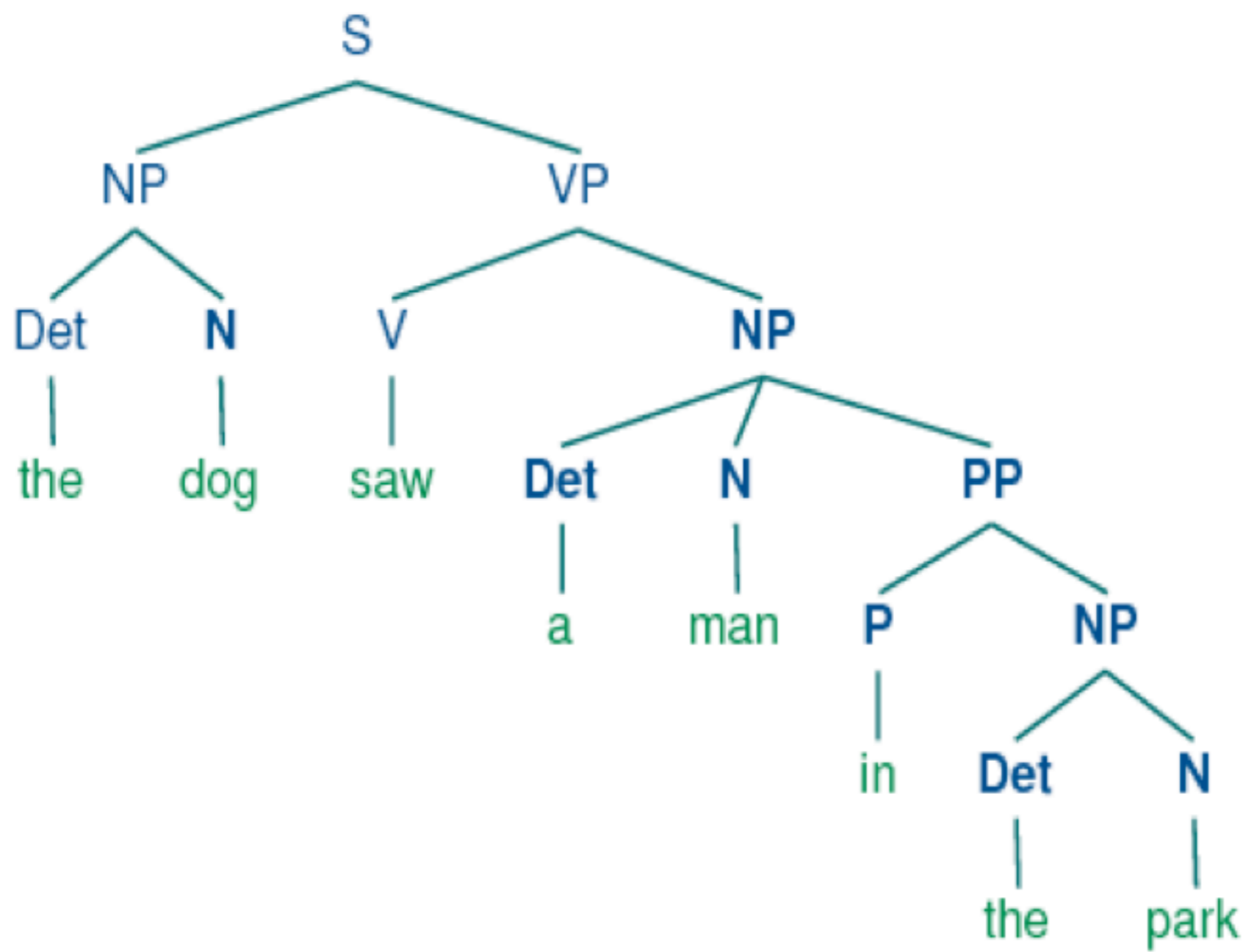
But syntax doesn't tell us much about meaning…

- *Colorless green ideas sleep furiously. [Chomsky]*

- *plastic cat food can cover*

Semantics: Lexical Ambiguity

- *I walked to the bank ...*

  of the river.

  to get money.

- *The bug in the room ...*

  was planted by spies.

  flew out the window.

- *I work for John Hancock ...*

  and he is a good boss.

  which is a good company.

- Discourse, Pragmatics

Discourse: coreference

A Short Story

President John F. Kennedy was assassinated.
The **president** was shot yesterday.
Relatives said that **John** was **a good father**.
JFK was **the youngest president** in history.
His family will bury **him** tomorrow.
Friends of **the Massachusetts native** will hold a candlelight service in **Mr. Kennedy**'s home town.

Pragmatics

Rules of Conversation

- Can you tell me what time it is?
- Could I please have the salt?

Speech Acts

- I bet you $50 that the Jazz will win tonight.
- Will you marry me?

NLP: a branch of AI

- Lack of world knowledge
- inferences

World Knowledge, Inferences

John went to the diner.

He ordered a steak.

He left a tip and went home.

John wanted to commit suicide.
He got a rope.

- Sparsity!!!

Zipf＇s Law

- the frequency of any word is [inversely proportional](#) to its rank： f = K / r

- fat-tail, most words occur only a couple of times

- high lexical diversity -> data sparseness

Goals of the class

- Key tasks, algorithms

- Essentially skills to build your system

- (Hopefully) see problems, holes, gaps, start research