CS447: Natural Language Processing

http://courses.engr.illinois.edu/cs447

Lecture 1: Introduction

Julia Hockenmaier

juliahmr@illinois.edu

3324 Siebel Center

Course Staff

Professor:

Julia Hockenmaier juliahmr@illinois.edu

Teaching assistants:

Qingrong Cheng qc16@illinois.edu

Maghav Kumar mkumar10@illinois.edu

Today's lecture

Course Overview:

What is NLP?

What will you learn in this course?

Course Admin:

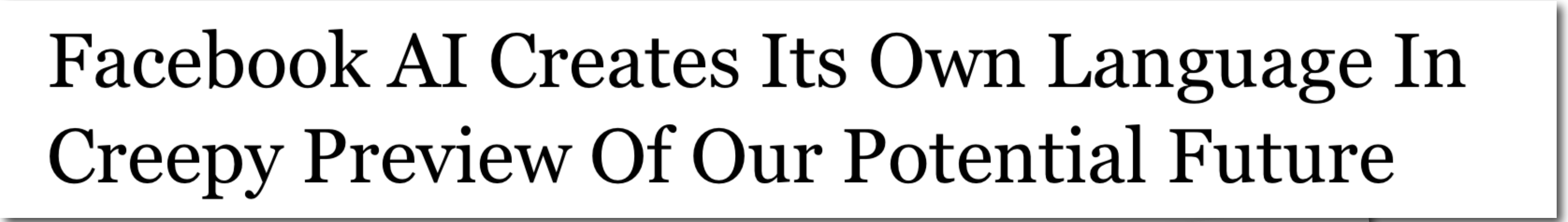
How will we teach this course?

How will you be assessed in this course?

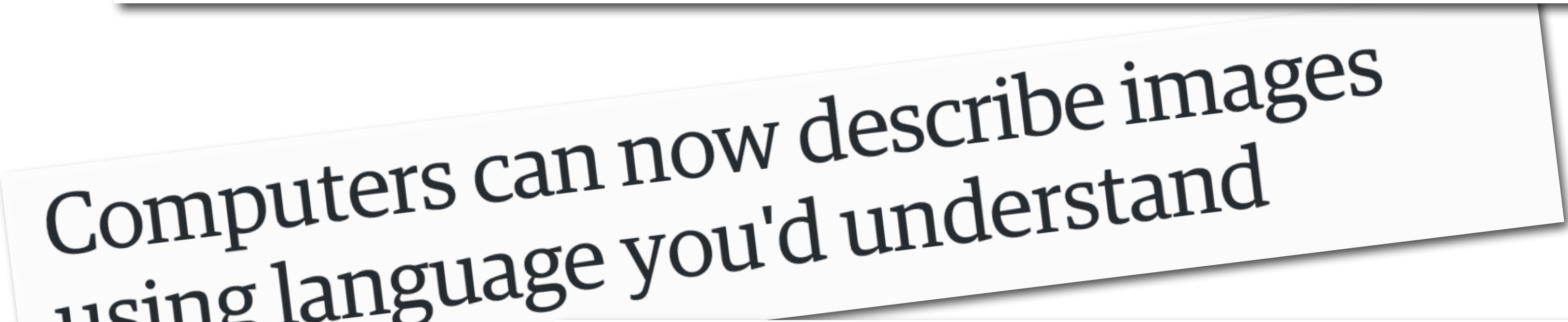
What is Natural Language Processing?

What is Natural Language Processing *really* ?

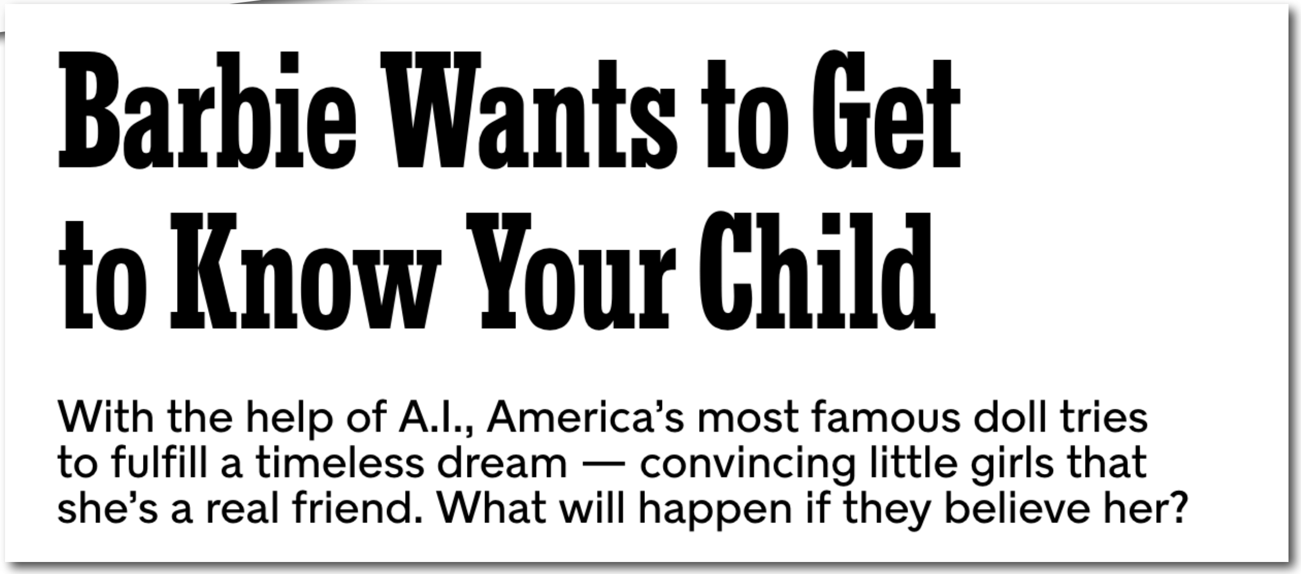
NLP in the news…



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?

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.”

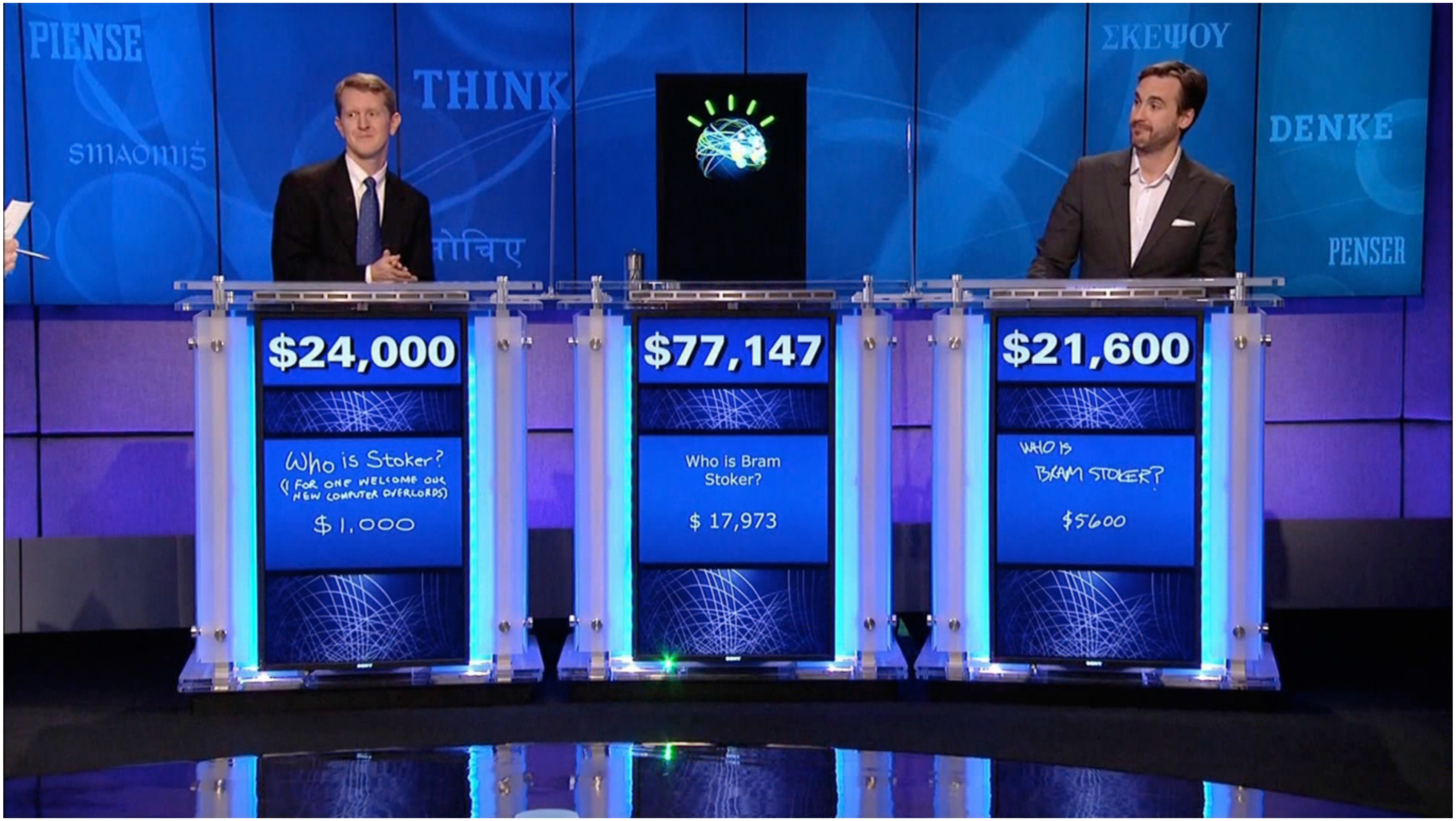
No, not really…
http://www.snopes.com/facebook-ai-developed-own-language/

How well does that Barbie do?

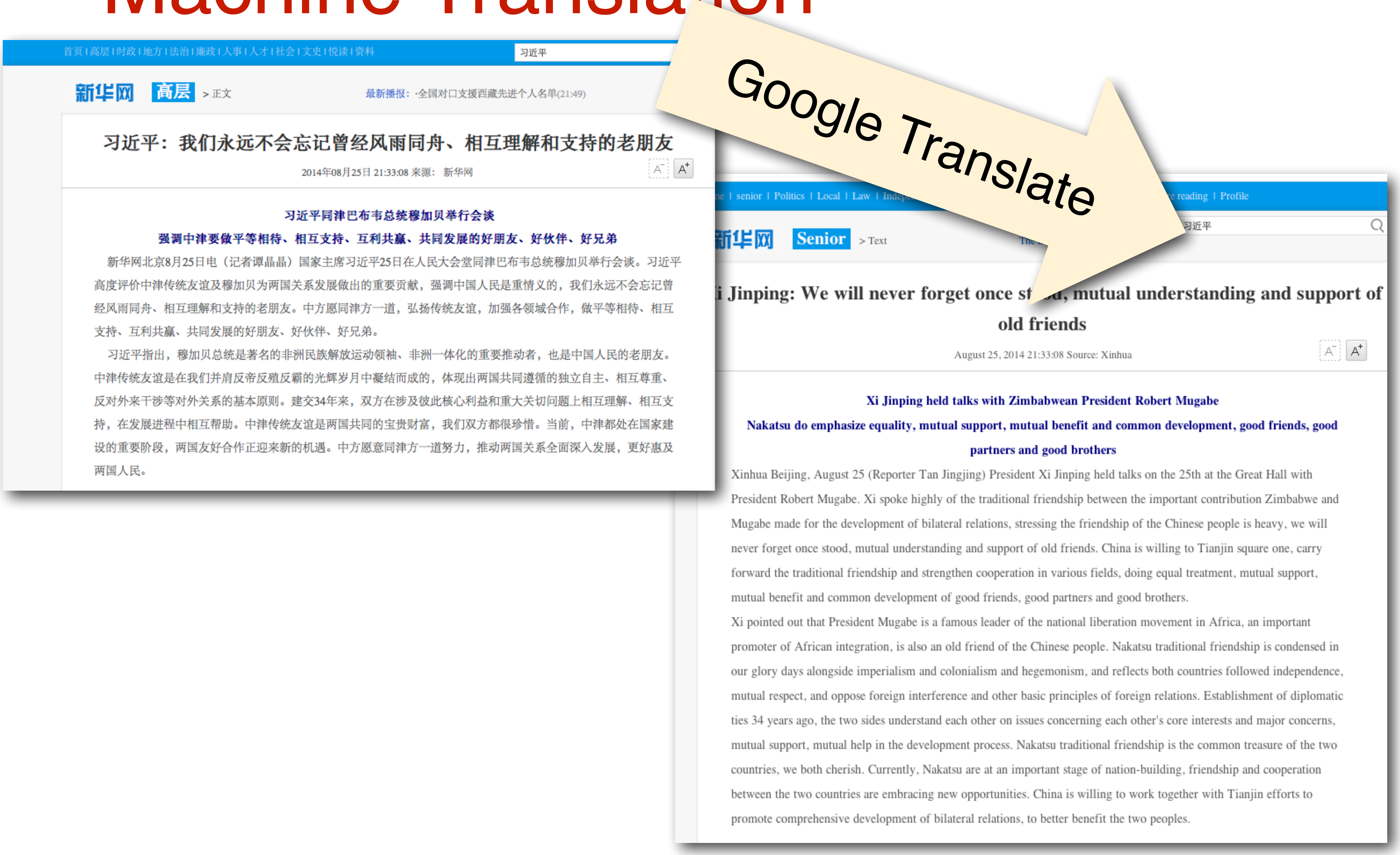
Barbie: *“Do you have any sisters?”*
Child: “Yeah, I only have one.”
Barbie: *“What’s something nice that your sister does for you?”*
Child: “She does nothing nice to me”
Barbie: *“Well, what is the last nice thing your sister did?”*
Child: “She helped me with my project
— and then she destroyed it.”
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)
https://www.nytimes.com/2015/09/20/magazine/barbie-wants-to-get-to-know-your-child.html

IBM's Watson wins at Jeopardy!



Machine Translation



首页 | 高层 | 时政 | 地方 | 法治 | 廉政 | 人事 | 人才 | 社会 | 文史 | 悦读 | 资料
习近平
新华网 高层 > 正文
最新播报：·全国对口支援西藏先进个人名单(21:49)

习近平：我们永远不会忘记曾经风雨同舟、相互理解和支持的老朋友

2014年08月25日 21:33:08 来源：新华网

习近平同津巴布韦总统穆加贝举行会谈

强调中津要做平等相待、相互支持、互利共赢、共同发展的好朋友、好伙伴、好兄弟

新华网北京8月25日电（记者谭晶晶）国家主席习近平25日在人民大会堂同津巴布韦总统穆加贝举行会谈。习近平高度评价中津传统友谊及穆加贝为两国关系发展做出的重要贡献，强调中国人民是重情义的，我们永远不会忘记曾经风雨同舟、相互理解和支持的老朋友。中方愿同津方一道，弘扬传统友谊，加强各领域合作，做平等相待、相互支持、互利共赢、共同发展的好朋友、好伙伴、好兄弟。

习近平指出，穆加贝总统是著名的非洲民族解放运动领袖、非洲一体化的重要推动者，也是中国人民的老朋友。中津传统友谊是在我们并肩反帝反殖反霸的光辉岁月中凝结而成的，体现出两国共同遵循的独立自主、相互尊重、反对外来干涉等对外关系的基本原则。建交34年来，双方在涉及彼此核心利益和重大关切问题上相互理解、相互支持，在发展进程中相互帮助。中津传统友谊是两国共同的宝贵财富，我们双方都很珍惜。当前，中津都处在国家建设的重要阶段，两国友好合作正迎来新的机遇。中方愿意同津方一道努力，推动两国关系全面深入发展，更好惠及两国人民。

i Jinping: We will never forget once st d, mutual understanding and support of old friends

August 25, 2014 21:33:08 Source: Xinhua

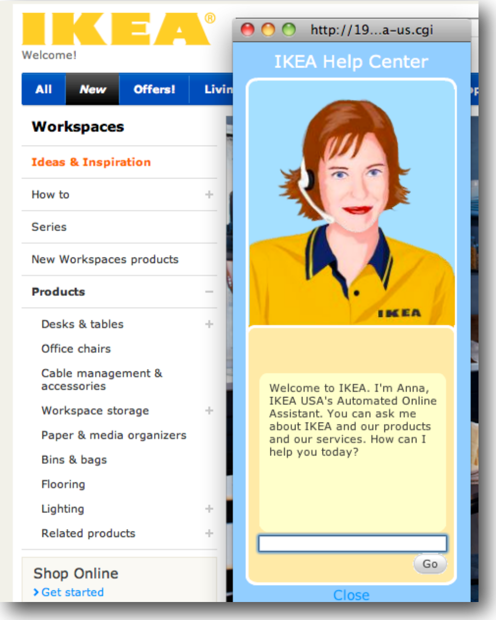
Xi Jinping held talks with Zimbabwean President Robert Mugabe

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. Nakatsu traditional friendship is the common treasure of the two countries, we both cherish. Currently, Nakatsu 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



IKEA Help Center

Workspaces

Ideas & Inspiration

How to +

Series

New Workspaces products

Products −

Desks & tables +

Office chairs

Cable management & accessories

Workspace storage +

Paper & media organizers

Bins & bags

Flooring

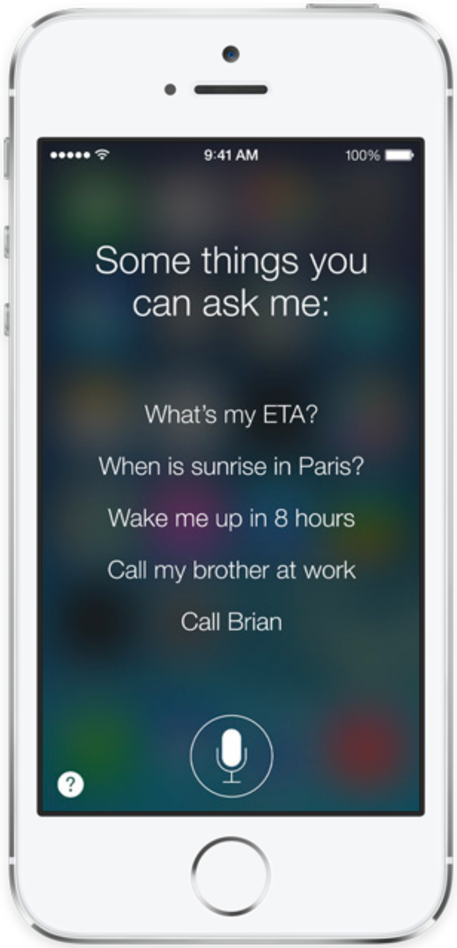
Lighting +

Related products +

Shop Online

› Get started

Welcome to IKEA. I'm Anna, IKEA USA's Automated Online Assistant. You can ask me about IKEA and our products and our services. How can I help you today?



Some things you can ask me:

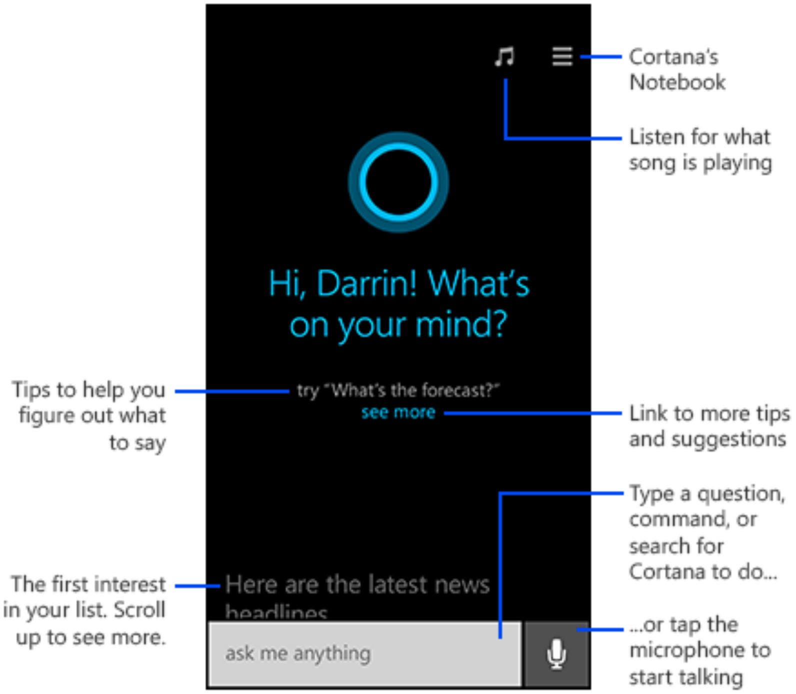
What's my ETA?

When is sunrise in Paris?

Wake me up in 8 hours

Call my brother at work

Call Brian



Hi, Darrin! What's on your mind?

try "What's the forecast?"
see more

Here are the latest news headlines

ask me anything

Cortana's Notebook

Listen for what song is playing

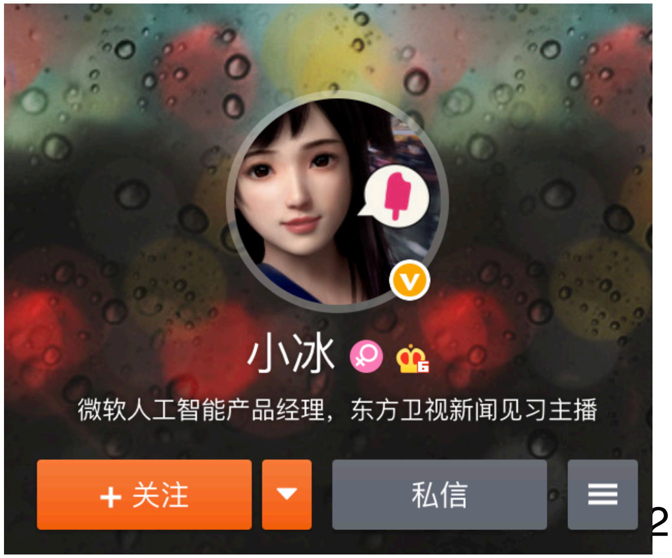
Tips to help you figure out what to say

Link to more tips and suggestions

Type a question, command, or search for Cortana to do...

The first interest in your list. Scroll up to see more.

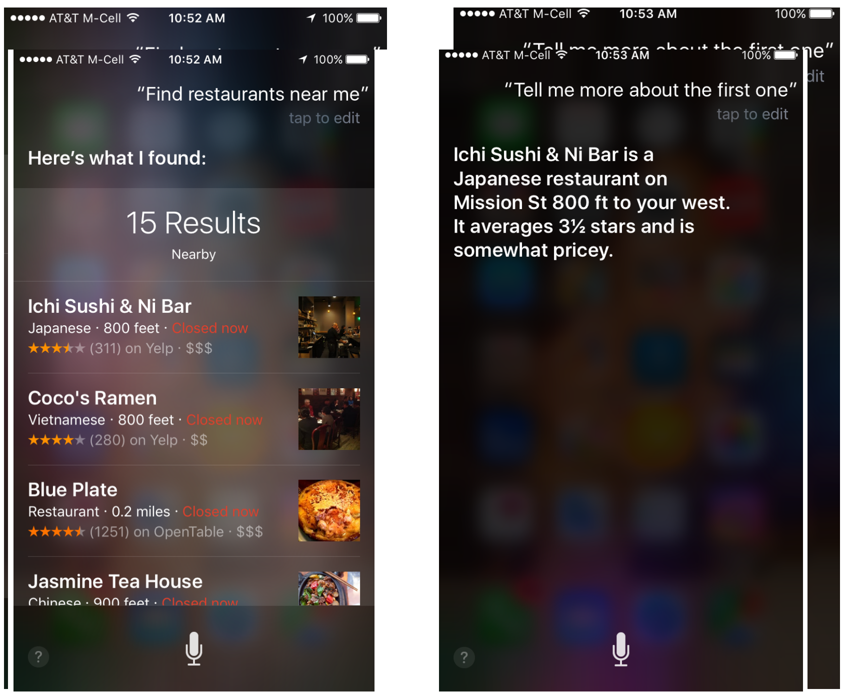
...or tap the microphone to start talking



小冰

微软人工智能产品经理，东方卫视新闻见习主播

Siri [Jurafsky & Martin, 3rd ed]



●●●●● AT&T M-Cell 10:52 AM 100%

"Find restaurants near me"

tap to edit

Here's what I found:

15 Results

Nearby

Ichi Sushi & Ni Bar
Japanese · 800 feet · Closed now
★★★★★ (311) on Yelp · $$$

Coco's Ramen
Vietnamese · 800 feet · Closed now
★★★★★ (280) on Yelp · $$

Blue Plate
Restaurant · 0.2 miles · Closed now
★★★★★ (1251) on OpenTable · $$$

Jasmine Tea House
Chinese · 900 feet · Closed now

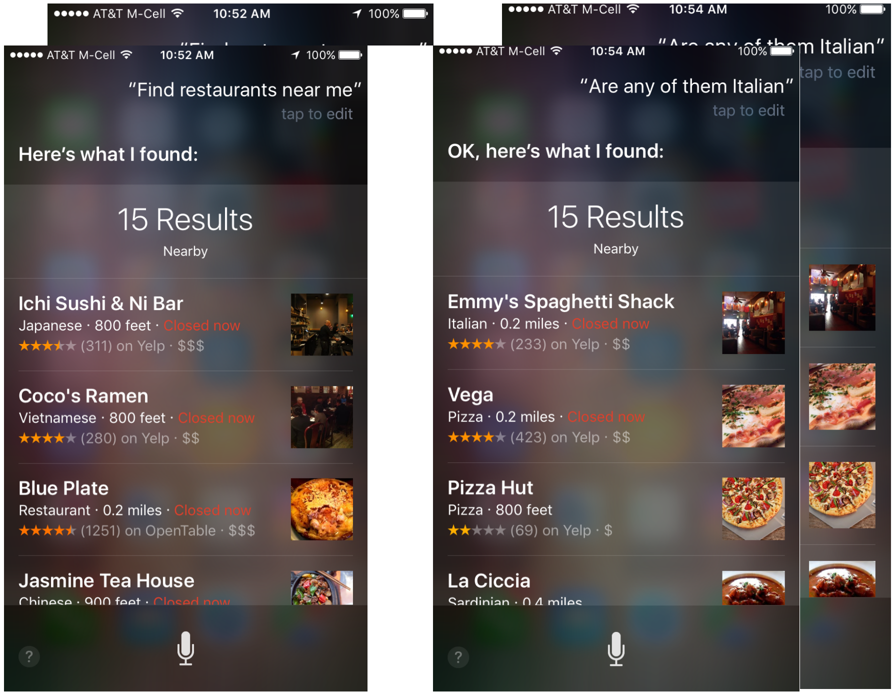
●●●●● AT&T M-Cell 10:53 AM 100%

"Tell me more about the first one"

tap to edit

Ichi Sushi & Ni Bar is a Japanese restaurant on Mission St 800 ft to your west. It averages 3½ stars and is somewhat pricey.

(a)



●●●●● AT&T M-Cell 10:52 AM 100%

"Find restaurants near me"

tap to edit

Here's what I found:

15 Results

Nearby

Ichi Sushi & Ni Bar
Japanese · 800 feet · Closed now
★★★★★ (311) on Yelp · $$$

Coco's Ramen
Vietnamese · 800 feet · Closed now
★★★★★ (280) on Yelp · $$

Blue Plate
Restaurant · 0.2 miles · Closed now
★★★★★ (1251) on OpenTable · $$$

Jasmine Tea House
Chinese · 900 feet · Closed now

●●●●● AT&T M-Cell 10:54 AM 100%

"Are any of them Italian"

tap to edit

OK, here's what I found:

15 Results

Nearby

Emmy's Spaghetti Shack
Italian · 0.2 miles · Closed now
★★★★★ (233) on Yelp · $$

Vega
Pizza · 0.2 miles · Closed now
★★★★★ (423) on Yelp · $$

Pizza Hut
Pizza · 800 feet
★★★★★ (69) on Yelp · $

La Ciccia
Sardinian · 0.4 miles

(b)

Challenge: Dialogue consists of several turns. Systems need to be able to understand and sanswer follow-up questions, etc.

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 paradigm shift (and a lot of hype) around “deep learning” and “AI” more generally

—Neural nets are powerful classifiers and sequence models

—Public libraries (Tensorflow, Pytorch, 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?

What will you learn in this class?

In this class, you will learn about some of

— the core **tasks** (as well as **data sets and evaluation metrics**) that people work on in NLP

— the fundamental **models and algorithms** that have been developed for these tasks

— the relevant **linguistic concepts and phenomena** that will be encountered in these tasks

The focus 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.
- Neural and 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)

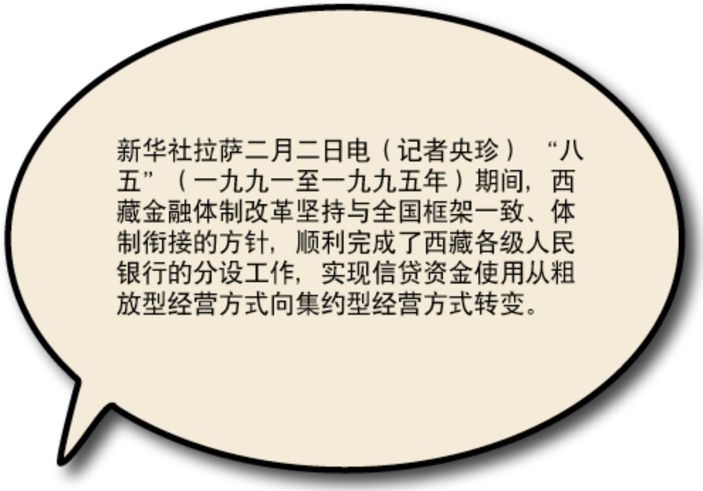
… but I would still like you to get an understanding of:

- How well does current NLP technology work (or not)?
- What NLP software and datasets are available?
- How to read NLP research papers [4 credits section]

Building a computer that ‘*understands*’ text: The traditional 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，顺利完成了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



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，顺利完成了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。

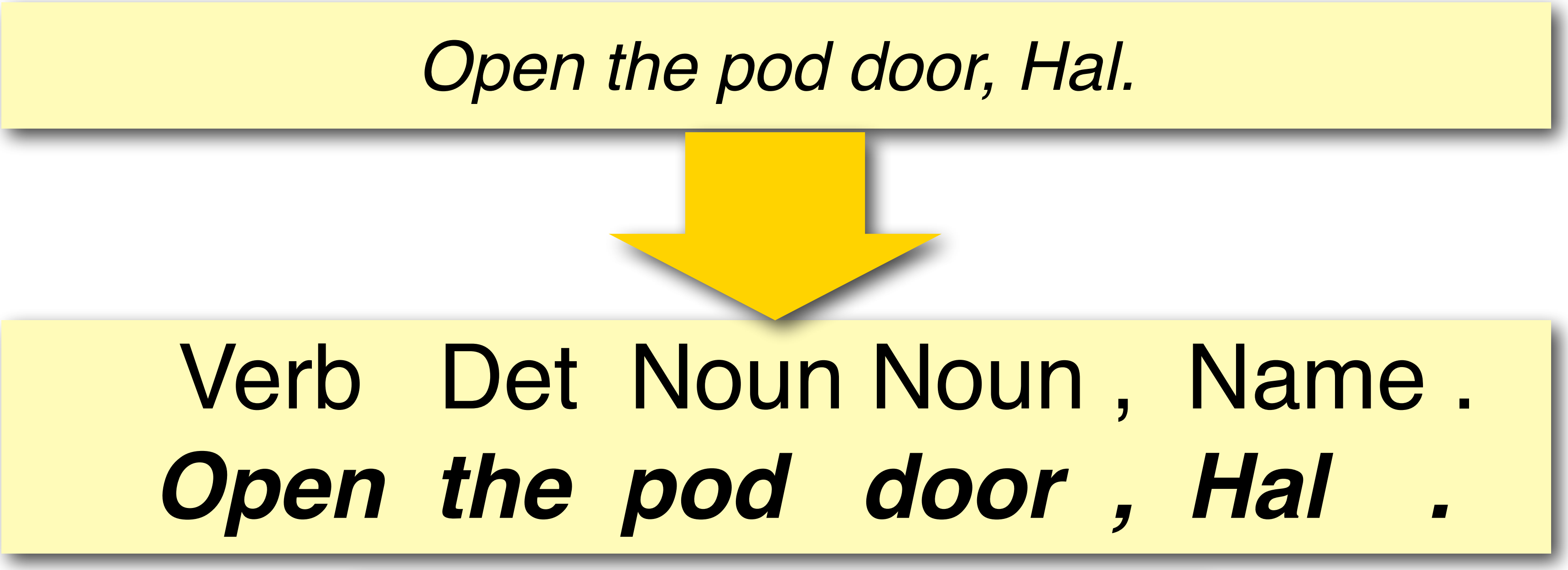
We need to split text into words and sentences.

- 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

$$\underset{T}{\operatorname{argmax}} P(T|S)$$

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

$$\begin{aligned}
\underset{T}{\operatorname{argmax}} P(T|S) &= \underset{T}{\operatorname{argmax}} P(T)P(S|T) \\
P(T) &=_{def} \prod_i P(t_i|t_{i-1}) \\
P(S|T) &=_{def} \prod_i P(w_i\,|\,t_i)
\end{aligned}$$

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

$$P(\ t_i = V \mid t_{i\text{-}1} = N\ ) = 0.3$$

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”

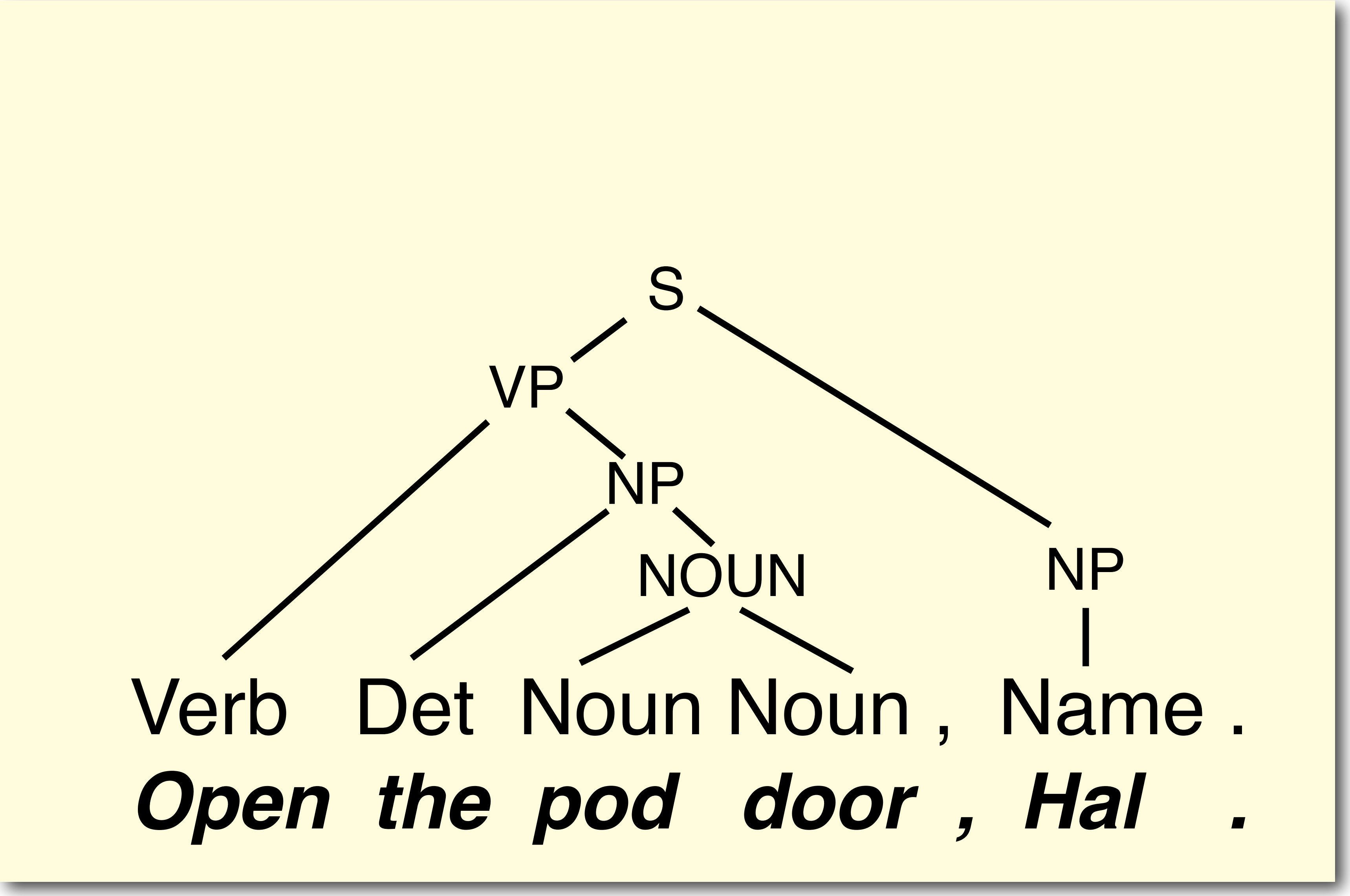
(Cassoulet = a French bean casserole)

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

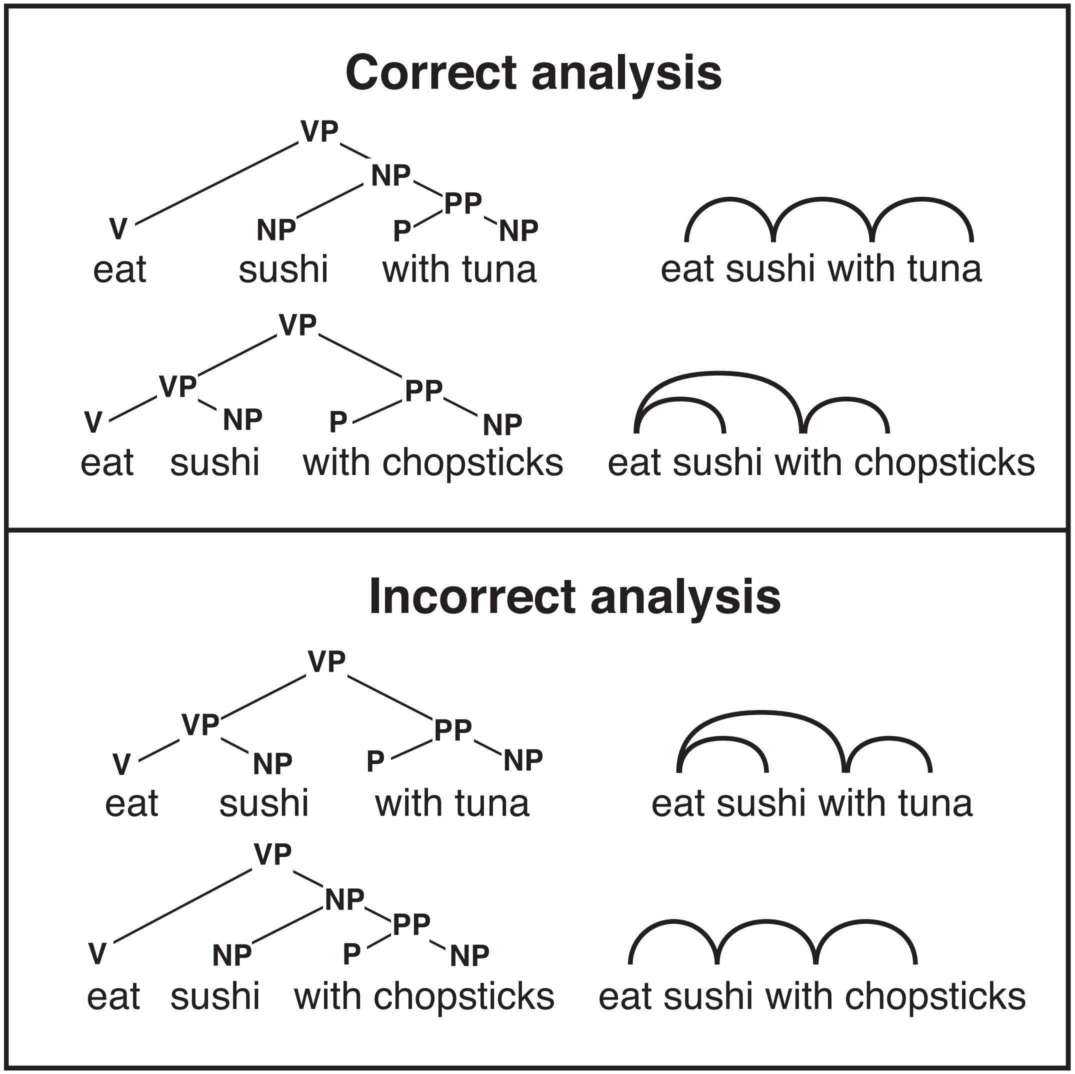
Our models need to be able to deal with this.

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



Observation: Structure corresponds to meaning



Question: what is grammar?

Grammar formalisms

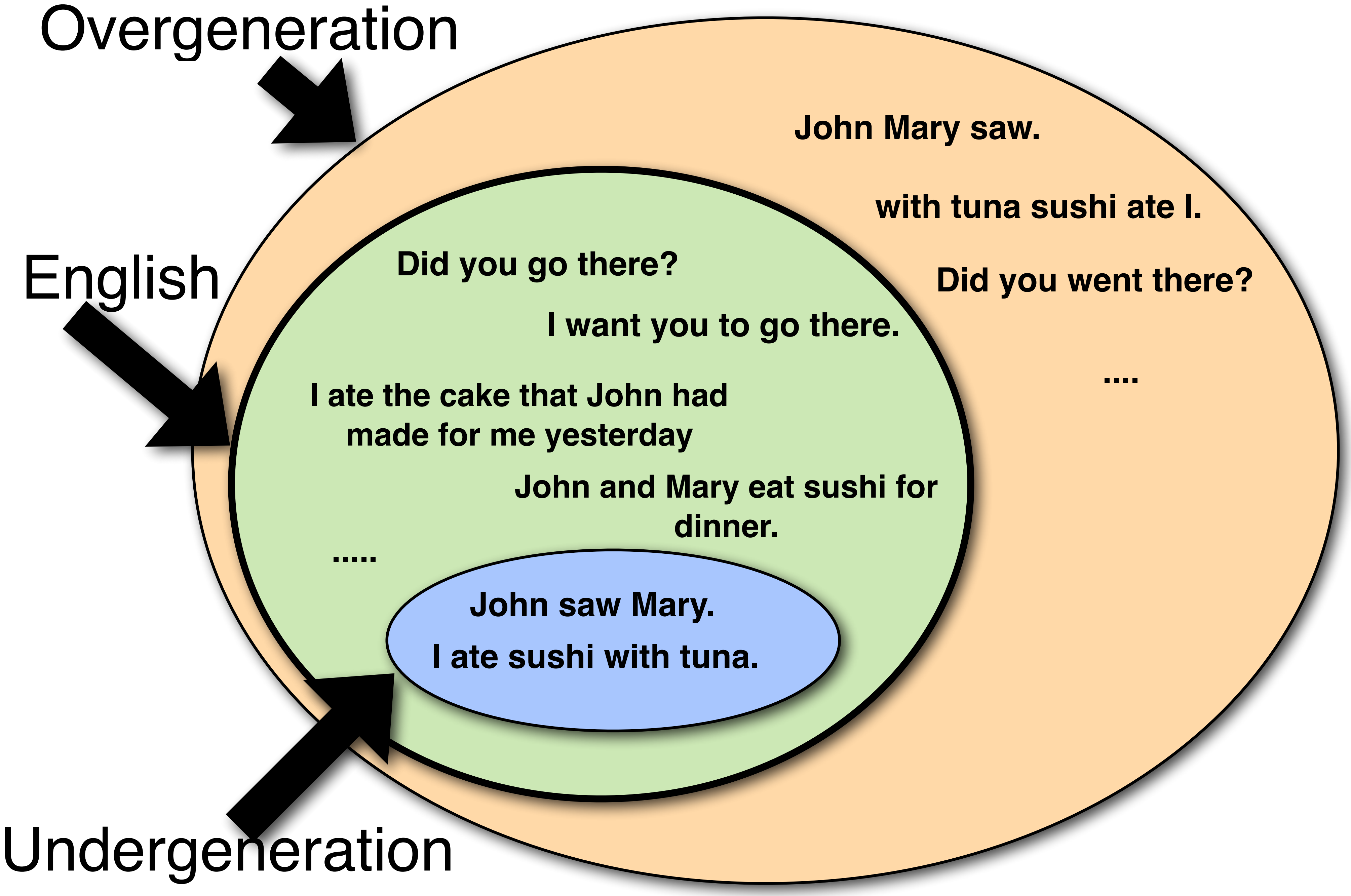
(= linguists' programming languages)

A precise way to define and describe
the structure of sentences.

Specific grammars

(= linguists' programs)

Implementations (in a particular formalism)
for a particular language (English, Chinese,....)



Overgeneration

English

Undergeneration

John Mary saw.

with tuna sushi ate I.

Did you went there?

....

Did you go there?

I want you to go there.

I ate the cake that John had made for me yesterday

John and Mary eat sushi for dinner.

.....

John saw Mary.

I ate sushi with tuna.

NLP and automata theory

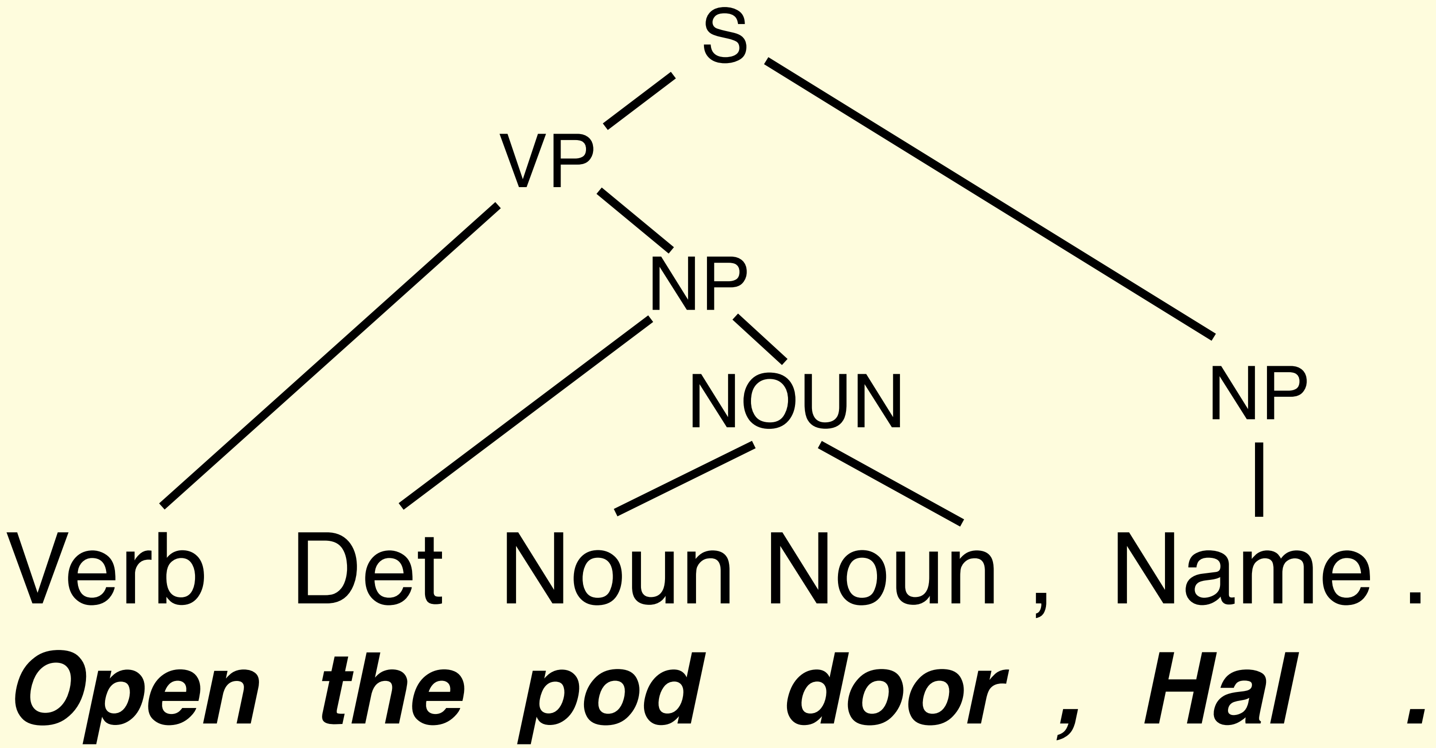
What kind of grammar/automaton
is required to analyze natural language?

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)))



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)

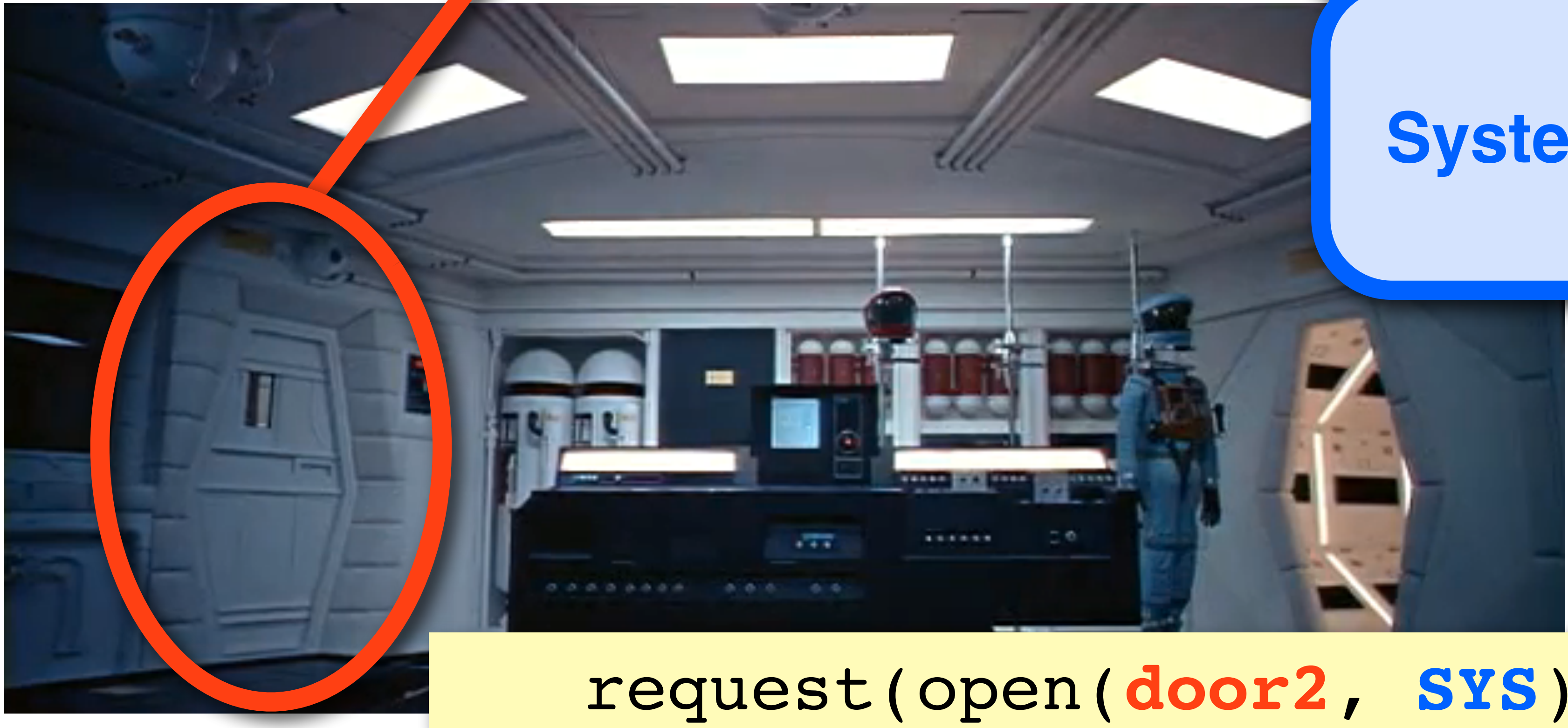
Understanding texts

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?)

Multimodal NLP: mapping from language to the world

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



System

```
request(open(door2, SYS))
```

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

NLP Pipeline: Assumptions

Each step in the NLP pipeline embellishes the input with **explicit information** about its linguistic structure

POS tagging: parts of speech of word,

Syntactic parsing: grammatical structure of sentence,….

Each step in the NLP pipeline requires its own explicit **(“symbolic”) output representation**:

POS tagging requires a POS tag set

(e.g. NN=common noun singular, NNS = common noun plural, …)

Syntactic parsing requires constituent or dependency labels

(e.g. NP = noun phrase, or nsubj = nominal subject)

These representations should capture linguistically appropriate **generalizations/abstractions**

Designing these representations requires linguistic expertise

NLP Pipeline: Shortcomings

Each step in the pipeline relies on a **learned model** that will return the *most likely* representations

- This requires a lot of **annotated training data** for each step
- Annotation is **expensive** and sometimes **difficult** (people are not 100% accurate)
- These models are **never 100% accurate**
- Models make more mistakes if their input contains mistakes

How do we know that we have captured the "***right*** **generalizations** when designing representations?

- Some representations are **easier to predict** than others
- Some representations are **more useful** for the next steps in the pipeline than others
- But we won't know how easy/useful a representation is until we have a model that we can plug into a particular pipeline

Course Admin

This class consists of...

… Lectures:

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

… Office Hours:

Julia: Wednesdays and Fridays, 2:00pm–2:30pm, Siebel 3324

Qingrong: TBD, Siebel 0207

Meghav: TBD, Siebel 0207

… Websites:

Syllabus, slides, policies, etc: *http://courses.engr.illinois.edu/cs447*

Discussions: *piazza.com/illinois/fall2019/cs447*

Grades, submitting assignments: Compass2G *http://compass2g.illinois.edu* and Gradescope http://gradescope.com

… Readings:

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

… Assessment:

4 assignments, 2 exams (4th credit hour: project or survey)

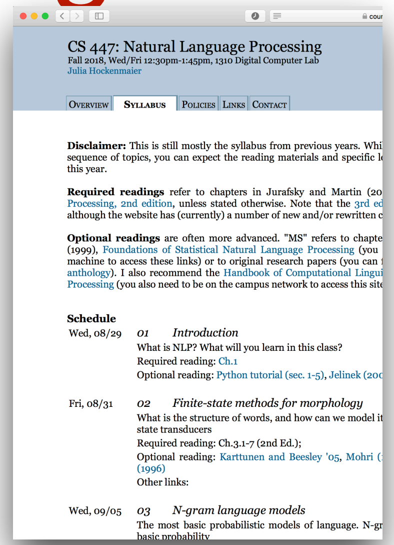
Lectures and office hours

Attend!

Ask questions!

Participate!

Reading



CS 447: Natural Language Processing
Fall 2018, Wed/Fri 12:30pm-1:45pm, 1310 Digital Computer Lab
Julia Hockenmaier

Disclaimer: This is still mostly the syllabus from previous years. Whi
sequence of topics, you can expect the reading materials and specific le
this year.

Required readings refer to chapters in Jurafsky and Martin (20
Processing, 2nd edition, unless stated otherwise. Note that the 3rd ed
although the website has (currently) a number of new and/or rewritten c

Optional readings are often more advanced. "MS" refers to chapte
(1999), Foundations of Statistical Natural Language Processing (you
machine to access these links) or to original research papers (you can f
anthology). I also recommend the Handbook of Computational Lingui
Processing (you also need to be on the campus network to access this site

Schedule

Wed, 08/29 01 *Introduction*
What is NLP? What will you learn in this class?
Required reading: Ch.1
Optional reading: Python tutorial (sec. 1-5), Jelinek (200

Fri, 08/31 02 *Finite-state methods for morphology*
What is the structure of words, and how can we model it
state transducers
Required reading: Ch.3.1-7 (2nd Ed.);
Optional reading: Karttunen and Beesley '05, Mohri (
(1996)
Other links:

Wed, 09/05 03 *N-gram language models*
The most basic probabilistic models of language. N-gr
basic probability

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** (We will mostly use chapters from the forthcoming 3rd edition, but may also use the 2nd edition from 2008 in a few place)

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)

We use Python 3

Why?

To make sure you can put what you've learned to practice.

How?

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 Gradescope.

Late policy?

No late assignments will be accepted (sorry).

Homework assignments

Schedule:

Week 2:	Friday, 09/06	HW1 out
Week 5:	Friday, 09/27	HW1 due, HW2 out
Week 8:	Friday, 10/20	HW2 due, HW3 out
Week 11:	Friday, 11/08	HW3 due, HW4 out
Week 14:	Friday, 12/06	HW4 due

Points per assignment:

HW1,HW2,HW3,HW4 = 10 points per assignment

Exams

What?

First midterm: Friday, Oct 11, in class

Second Midterm: Wednesday, Dec 11, 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 4: Proposal due (What topic? What papers will you read?)
Fri, Nov 8: Progress report due (Are your experiments on track?)
Thu, Dec 12: 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 4: Proposal due (What topic? What papers will you read?)
Fri, Nov 8: Progress report due (Is your paper on track?)
Thu, Dec 12: Final report due (Summary of papers)

Today's readings

Today's lecture:

Jurafsky and Martin Chapter 1 (2nd edition)

http://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf