

Lecture 1: Welcome!

CS447 Natural Language Processing

Julia Hockenmaier

juliahmr@illinois.edu

<https://courses.grainger.illinois.edu/cs447/>



CS447 Lecture 1: course Admin

Welcome to CS447!

Professor:

Julia Hockenmaier

juliahmr@illinois.edu

<http://juliahmr.cs.illinois.edu>

TAs:

Ching-Hua Yu

Liliang Ren

Keval Morabia

Going virtual: course structure

Structure

- **Lectures** will be prerecorded and uploaded to Mediaspace.
- **Office hours** will take place over Zoom
- **Additional synchronous Zoom activities:** Fridays 11:00am—12:15pm

Assessment (likely all done via Gradescope)

- 10 online, open-book **quizzes** throughout the semester
- 4 **programming assignments** (Python3, Gradescope)
- The 4th credit hour requires also a **literature review or research project.**

NB: We will add you to Gradescope.

Email us if you're not registered yet.

Going virtual: platforms

Website (for links to slides, videos, reading materials, policies, etc.)

<https://courses.grainger.illinois.edu/cs447/fa2020/>

Piazza (for discussion)

<https://piazza.com/class/ke375fpcbky5mb>

Please sign up!

Mediaspace Channel (where the videos will be hosted)

Subscribe here

<https://mediaspace.illinois.edu/channel/CS447+Natural+Language+Processing+Fall+2020/172894481/subscribe>

Going virtual: Lectures and Zoom activities

Lecture videos (“asynchronous delivery”) — REQUIRED MATERIALS —

- Lecture videos contain **all the necessary material** for quizzes and assignments
- I will upload class lecture **videos to our Mediaspace channel** and PDF **slides to our website** before the time our regular class was supposed to take place (Wednesdays/Fridays 11am CT).

Zoom activities (“synchronous delivery”) — OPTIONAL MATERIALS—

- **Fridays, 11am CT.** Zoom link on class website. Log in with your NetID.
- We will discuss materials in more depth, and perhaps do some exercises.
(Disclaimer: this is an experiment, let’s see how well this works)

We may record these sessions and upload them to a private Mediaspace channel that is only accessible to registered students.

How can you get the most out of CS447?

Be proactive!

- Watch the videos and read the required readings
- Start early with the assignments

Communicate:

- Participate in Piazza discussions
- Attend office hours

There are no stupid questions!

We'd *really* like to know if there is something you don't understand, because that means you've thought about the material, and we didn't explain it well.

Assessment

If you're taking this class for **3 credit hours**:

- $\frac{2}{3}$ of your credit will come from the 10 quizzes
- $\frac{1}{3}$ of your credit will come from the 4 programming assignments

If you're taking this class for **4 credit hours**:

- $\frac{1}{2}$ of your credit will come from the 10 quizzes
- $\frac{1}{4}$ of your credit will come from the 4 programming assignments
- $\frac{1}{4}$ of your credit will come from your literature review/research project.
- Each quiz will count as much as every other quiz, even if one quiz has more questions than the other.
- Each programming assignment will count as much as every other assignment, even if one has more parts than the other.

Quizzes

How?

- We are planning to use **Gradescope**
(but may migrate to Compass if Gradescope doesn't work well enough for this)
- **Open-book** quizzes, probably **with a time limit**
- Solutions will only be released after the deadline
- **One week/quiz** (on the weeks where no HW is released)

Why?

- We want to make sure you follow the material during the semester
- We want to evaluate that you understand the material

What?

- **Mostly short questions** (e.g. multiple choice)
- Probably also **some longer essay-type questions**.

Programming 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 (except for medical/religious exemptions)

Assignments Schedule

Current dates and materials not on Gradescope will be posted at
<https://courses.grainger.illinois.edu/cs447/fa2020/index.html>

Programming Assignments (3 weeks/assignment)

Assignments will be released by 11:59pm (or earlier)
and will be due by 11:59pm on the due date.

09/04—09/25 HW1

09/25—10/16 HW2

10/16—11/04 HW3

11/06—12/04 HW4

Quizzes (1 week/assignment)

4th Credit Hour: Additional Assessments

You can choose between a **Research Project** and a **Literature Review**.

We will provide more details at <https://courses.grainger.illinois.edu/cs447/fa2020/4credits.html>

We do not allow group projects

Deadlines:

Oct 1: Proposal (we will release a LaTeX template)

We won't give you a grade, but you need our approval on your topic.
(If you want to change your topic later, talk to us)

Nov 15: Status update report

(We won't grade you either;
this is just a checkpoint to make sure everything is on track)

Dec 9: Final report

[This is what you will be graded on]

4th Credit Hour: Research Project

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.

(We recommend resources such as Google Colab to run experiments)

Why?

To make sure you get a deeper knowledge of NLP by reading a few original papers in sufficient depth to build an actual 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 a number of original papers in sufficient depth to discuss and compare them, even if you don't build an actual system.

Grades

I don't grade "on a curve":

If everybody does really well in this class, everybody gets an A, not just the top X%.

I only assign letter grades at the end of the semester.

You should know what percent of the grade you have received so far, but I may not be able to tell you *precisely* what letter grade that may correspond to (although you should talk to me if you want to know whether you're doing well or not so well).

For assignments and quizzes, the undergrads' performance will determine the grading scale for everybody.

Academic Integrity

You can (and should) talk to each other about programming assignments, but you need to write all the code yourself.

We may use tools such as MOSS to detect plagiarism.

You should not talk to each other about the quizzes, although you can ask us clarification questions if (say) something is poorly worded.

If you're taking this class for four credits, your research project and literature survey also need to be your own work and you need to cite all sources.

You also can't get credit for the same work for two different classes.

If a project is related to your research, you need to let us know how it is related and who your advisor is if we have questions.

DRES accommodations

If you need any disability related accommodations, talk to DRES
(<http://disability.illinois.edu>, disability@illinois.edu, phone 333-4603)

If you are concerned you have a disability-related condition that is impacting your academic progress, there are academic screening appointments available on campus that can help diagnosis a previously undiagnosed disability by visiting the DRES website and selecting “Sign-Up for an Academic Screening” at the bottom of the page.”

Come and talk to me as well, especially once you have a letter of accommodation from DRES.

Do this early enough so that we can take your requirements into account!

Going virtual during a pandemic...

This semester is going to be different
(and possibly challenging) for all of us.

Please reach out to us if you're having any difficulties.
Communication is going to be essential.

Please be kind to each other.
Lots of people are really stressed right now.

Welcome to
CS447!

CS447 Lecture 01:
What will you learn
in this class?

What will you learn in this class?

– What is NLP?

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

– How does NLP work?

The fundamental **models, algorithms and representations** that have been developed for these tasks

– Why is NLP hard?

The relevant **linguistic concepts and phenomena** that have to be handled to do well at 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]

Our syllabus and textbook

You can find brief descriptions of our syllabus at

<https://courses.grainger.illinois.edu/cs447/fa2020/index.html>

[NB: if you hover with your mouse you will see links to reading materials]

Slides and links to videos/playlists will be uploaded here too.

Our Textbook:

Jurafsky and Martin, Speech and Language Processing 3rd ed.

<https://web.stanford.edu/~jurafsky/slp3/>

CS447 Lecture 1: What is NLP?

In Science Fiction

(Kubricks' 1968 movie 2001: A Space Odyssey)

A conversation onboard the Discovery One spacecraft between HAL 9000 (a sentient computer developed in Urbana, IL https://en.wikipedia.org/wiki/HAL_9000) and Dave, a human astronaut:

Dave: Open the pod bay doors, please, HAL. Open the pod bay doors, please, HAL. Hello, HAL, do you read me? Hello, HAL, do you read me? Do you read me, HAL? Do you read me, HAL? Hello, HAL, do you read me? Hello, HAL, do you read me? Do you read me, HAL?

HAL: Affirmative, Dave. I read you.

Dave: Open the pod bay doors, HAL.

HAL: I'm sorry, Dave. I'm afraid I can't do that.

Dave: What's the problem?

HAL: I think you know what the problem is just as well as I do.

Dave: What are you talking about, HAL?

HAL: This mission is too important for me to allow you to jeopardize it.

Dave: I don't know what you're talking about, HAL.

HAL: I know that you and Frank were planning to disconnect me. And I'm afraid that's something I cannot allow to happen.

Dave: Where the hell did you get *that* idea, HAL?

HAL: Dave, although you took very thorough precautions in the pod against my hearing you, I could see your lips move.

Dave: All right, HAL. I'll go in through the emergency airlock.

HAL: Without your space helmet, Dave, you're going to find that rather difficult.

Dave: *[sternly]* HAL, I won't argue with you anymore. Open the doors.

HAL: *[monotone voice]* Dave, this conversation can serve no purpose anymore. Good-bye.

[https://en.wikiquote.org/wiki/2001:_A_Space_Odyssey_\(film\)#Dialogue](https://en.wikiquote.org/wiki/2001:_A_Space_Odyssey_(film)#Dialogue)

In Science Fiction

(Kubricks' 1968 movie 2001: A Space Odyssey)

A conversation onboard the Discovery One spacecraft between HAL 9000 (a sentient computer developed in Urbana, IL https://en.wikipedia.org/wiki/HAL_9000) and Dave, a human astronaut:

Dave: Open the pod bay doors, please, HAL. Open the pod bay doors, please, HAL. Hello, HAL, do you read me? Hello, HAL, do you read me? Do you read me, HAL? Do you read me, HAL? Hello, HAL, do you read me? Hello, HAL, do you read me? Do you read me, HAL?

HAL: Affirmative, Dave. I read you.

Dave: Open the pod bay doors, HAL.

HAL: I'm sorry, Dave. I'm afraid I can't do that.

Dave: What's the problem?

HAL: I think you know what the problem is just as well as I do.

Dave: What are you talking about, HAL?

HAL: This mission is too important for me to allow you to jeopardize it.

Dave: I don't know what you're talking about, HAL.

HAL: I know that you and Frank were planning to disconnect me. And I'm afraid that's so.

Dave: Where the hell did you get *that* idea, HAL?

HAL: Dave, although you took very thorough precautions in the pod against my hearing you, I can hear you just as well as you can hear me.

Dave: All right, HAL. I'll go in through the emergency airlock.

HAL: Without your space helmet, Dave, you're going to find that rather difficult.

Dave: [sternly] HAL, I won't argue with you anymore. Open the doors.

HAL: [monotone voice] Dave, this conversation can serve no purpose anymore. Good-bye.

[https://en.wikiquote.org/wiki/2001:_A_Space_Odyssey_\(film\)#Dialogue](https://en.wikiquote.org/wiki/2001:_A_Space_Odyssey_(film)#Dialogue)

What are HAL's capabilities?

- HAL speaks fluent, natural English
- HAL's responses are appropriate in the context
 - “Affirmative, I read you”,
 - “I think you know what the problems is just as well as I do”
- HAL understands references to the physical world:
 - “Open the pod bay doors”
- HAL talks about what HAL sees:
 - “I could see your lips move”, “without your space helmet”
- HAL reasons
 - “I can't do that”, “this mission is too important”
 - “I know that you and Frank were planning to disconnect

A real human-computer dialogue from the 1960s: ELIZA (Weizenbaum, 1966)

Computer: TELL ME MORE ABOUT YOUR FAMILY

Human: My mother takes care of me.

Computer: WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

Human: My father.

Computer: YOUR FATHER

Human: You are like my father in some ways.

Computer: WHAT RESEMBLANCE DO YOU SEE

Human: You are not very aggressive but I think you don't want me to notice that.

Computer: WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE

Human: You don't argue with me.

Computer: WHY DO YOU THINK I DON'T ARGUE WITH YOU

Human: You are afraid of me.

The ELIZA effect

People *really* trusted this system:

Weizenbaum: “I had not realized ... that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people”

ELIZA was based on a cascade of finite-state transducers:

Human: *You don't argue with me.*

1. Replace **you** with **I** and **me** with **you**:

I don't argue with you.

2. Replace **<...>** with **Why do you think <...>**:

Why do you think I don't argue with you.

Computer: *WHY DO YOU THINK I DON'T ARGUE WITH YOU*

- How could you break a system like ELIZA?
- Where do you see the ELIZA effect today?

In the 21st century, NLP remains 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

A REPORTER AT LARGE OCTOBER 14, 2019 ISSUE

CAN A MACHINE LEARN TO WRITE FOR THE NEW YORKER?

How predictive-text technology could transform the future of the written word.

The AI Text Generator That's Too Dangerous to Make Public

Researchers at OpenAI decided that a system that scores well at understanding language could too easily be manipulated for malicious intent.

How AI Can Create And Detect Fake News

A.I. breakthroughs in natural-language processing are big for business

BY JEREMY KAHN

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 me to.”

No, not really...

<http://www.snopes.com/facebook-ai-developed-own-language/>

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

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,”

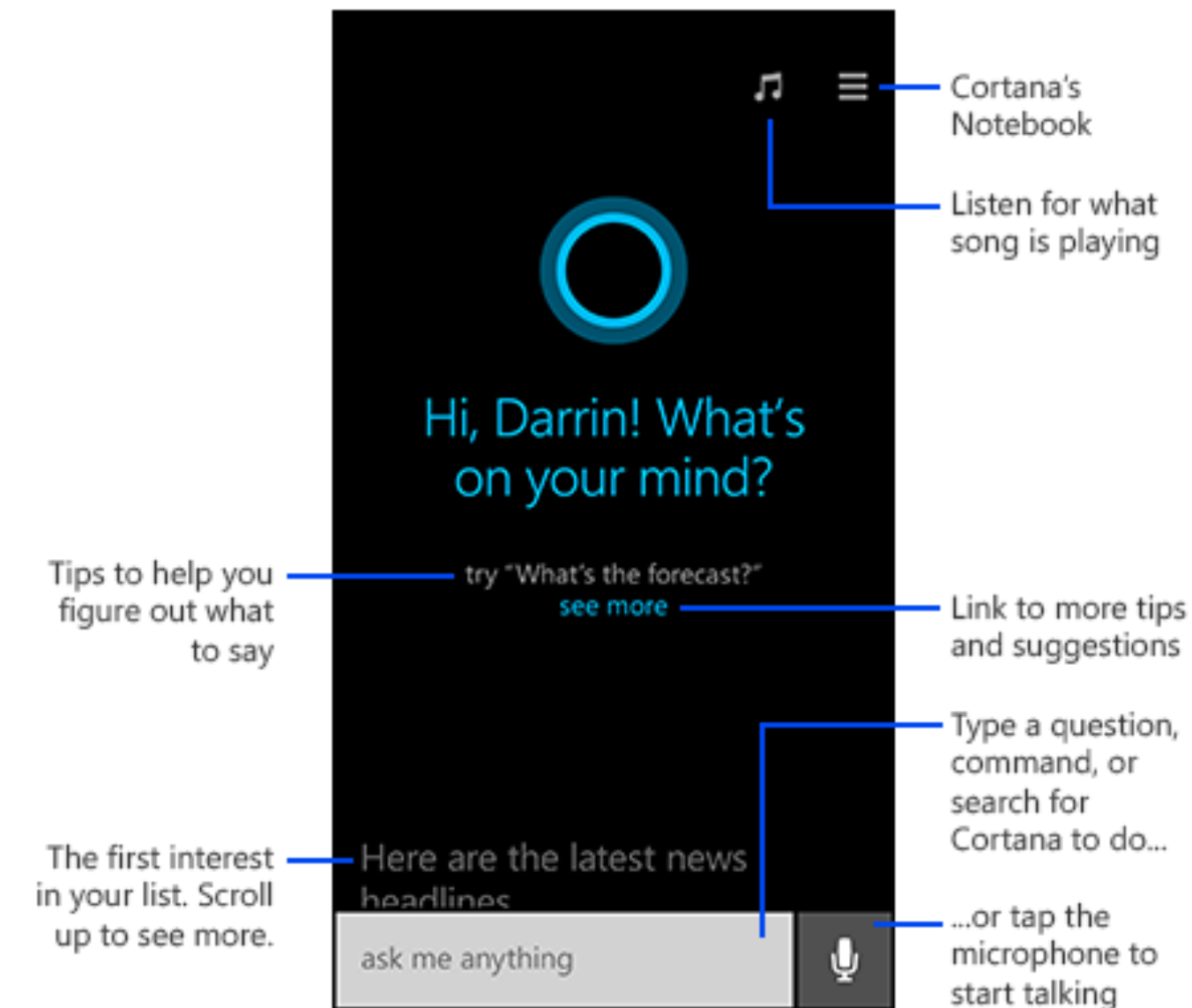
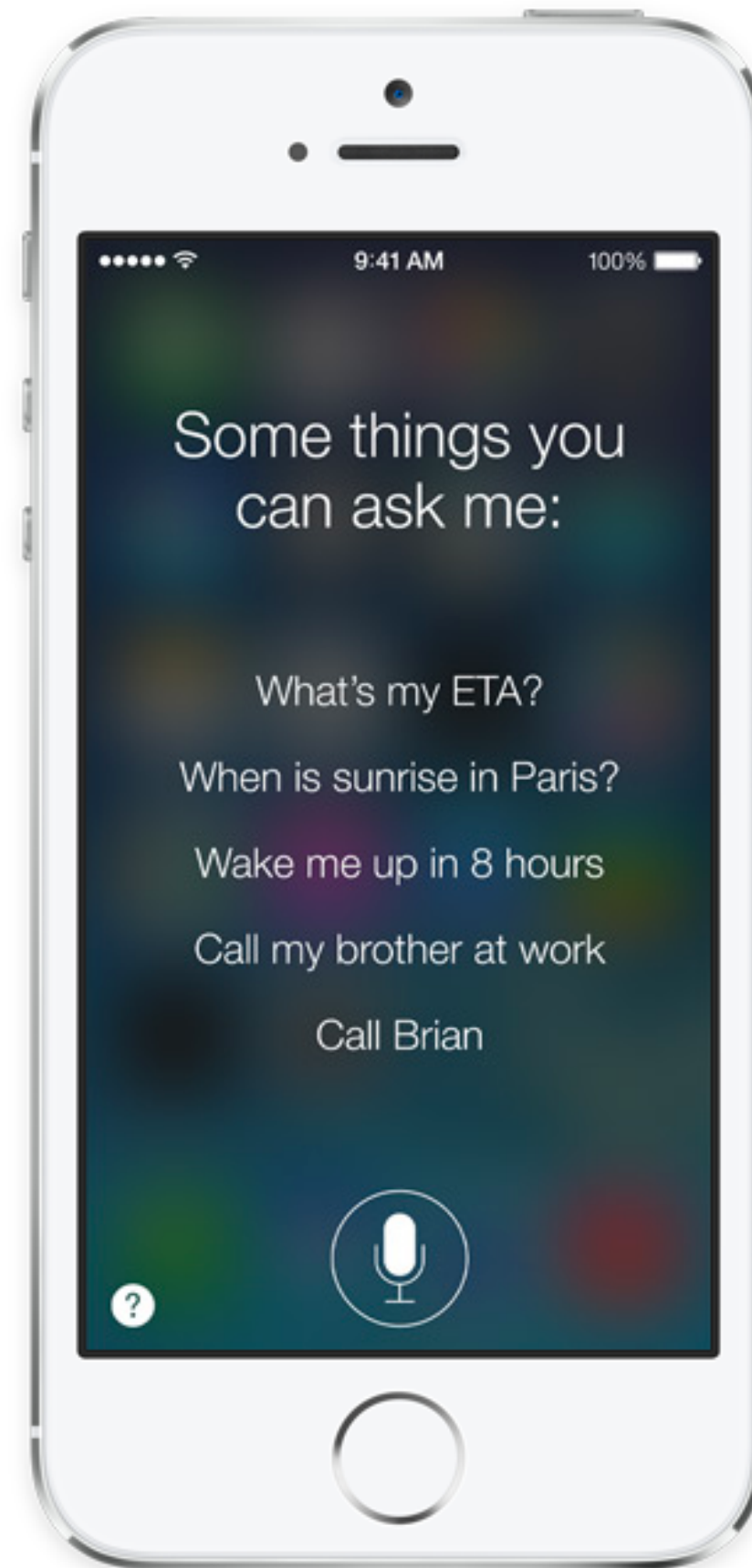
Barbie: *“You never know, she might appreciate hearing it”*

(Thanks to Barbara Grosz for pointing out the example from the NYT)

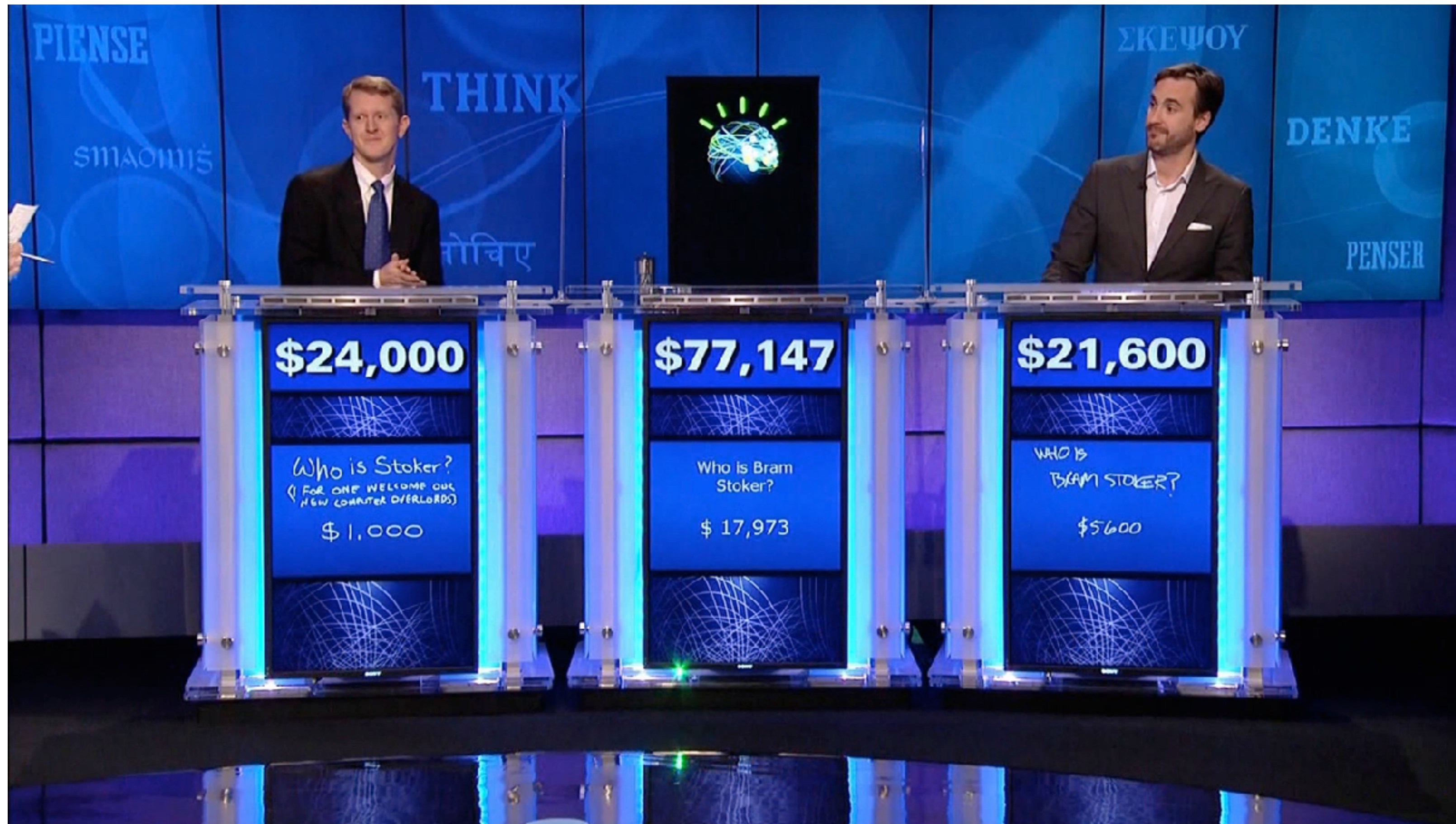
Barbie Wants to Get to Know Your Child

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?

Dialog systems, chatbots, digital assistants



IBM's Watson wins at Jeopardy!



Machine Translation

爷爷心疼孙女让娃睡懒觉 没想到引发了家庭矛盾

2020-08-25 08:06:03 来源: 钱江晚报

70岁的钟大伯（化名）陷入了“暑期焦虑”：这个暑假，他每周都要接送孙女上下培训班。高温、酷暑，每天回来，都像脚踩棉花般没力气。

除了身体上的不适，还有精神上的紧张。

觉得儿子儿媳给孩子报班太多，钟大伯还和他们产生了冲突：“大热天的，大人孩子都遭罪。”

这段时间，钟大伯因为容易激动发火，失眠，胃口差，血压一直不稳定，来到了浙江省人民医院精神卫生科就诊。

Google Translate

Grandpa feels sorry for his granddaughter and let the baby sleep in

2020-08-25 08:06:03 Source: Qianjiang Evening News

Uncle Zhong (a pseudonym), 70, fell into "summer anxiety": This summer, he would shuttle his granddaughter to and from training classes every week. With high temperatures and scorching heat, every day I come back, I feel as weak as stepping on cotton.

In addition to physical discomfort, there is also mental tension.

Feeling that his son and daughter-in-law were reporting too much for their children, Uncle Zhong also had a conflict with them: "It's a hot day, adults and children suffer."

During this period of time, Uncle Zhong came to the Mental Health Department of Zhejiang Provincial People's Hospital because he was prone to get angry, insomnia, poor appetite, and unstable blood pressure.

http://education.news.cn/2020-08/25/c_1210768533.htm

Huge language models solve NLP?

- A **language model** can be used to generate (produce) text
- Massive neural language models trained on vast amounts of text have been developed in the last few years
- Most recent incarnation: GPT-3 (175B parameters, trained on 300B tokens)
- But these models have no access to meaning.

See also Bender & Koller '20 for a critique <https://www.aclweb.org/anthology/2020.acl-main.463.pdf>

Opinion

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

by **Gary Marcus** and **Ernest Davis**

August 22, 2020

<https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>

Human Prompt (given to GPT-3)

At the party, I poured myself a glass of lemonade,
but it turned out to be too sour, so I added a little sugar.
I didn't see a spoon handy, so I stirred it with a cigarette.
But that turned out to be a bad idea because

[GPT-3's generated continuation]

it kept falling on the floor. That's when he decided to start
the Cremation Association of North America, which has become
a major cremation provider with 145 locations.

from Marcus & Davis '20

<https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>

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, 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

Examples of NLP applications

(What can NLP be used for?)

Natural language (and speech) interfaces

- Search/IR, database access, image search, image description
- Dialog systems (e.g. customer service, robots, cars, tutoring), chatbots

Information extraction, summarization, translation:

- Process (large amounts of) text automatically to obtain meaning/knowledge contained in the text
- Identify/analyze trends, opinions, etc. (e.g. in social media)
- Translate text automatically from one language to another

Convenience:

- Grammar/style checking, automate email filing, autograding

Examples of NLP tasks

(What capabilities do NLP systems need?)

Natural language understanding

- Extract information (e.g. about entities, events or relations between them) from text
- Translate raw text into a meaning representation
- Reason about information given in text
- Execute NL instructions

Natural language generation and summarization

- Translate database entries or meaning representations to raw natural language text
- Produce (appropriate) utterances/responses in a dialog
- Summarize (newspaper or scientific) articles, describe images

Natural language translation

- Translate one natural language to another

CS447 Lecture 1:
Building a computer that
'understands' text:
The traditional NLP pipeline

What does it take to understand text?

死亡谷测得54.4摄氏度高温
美国加州名胜或破世界纪录

Çavuşoğlu'ndan Atina'ya uyarı:
Bazı ülkelerin doldurushuna gelip,
kendinizi riske atmayın

รอยัลลิสต์มาร์เก็ตเพลส: เฟชบุ๊ก
เตรียมดำเนินการกฎหมายกับรัฐบาล
ไทย หลังบังคับบล็อกการเข้าถึงกลุ่มปิดที่
พูดคุยเกี่ยวกับราชวงศ์

හ-ඒ ආප් ඊයැ-ඒ සංඛද 12
හඬු, ඡුආප්ඡු හඬුඤ ඉඬදඬ
දහඤ හඬ

Qabiyyeen xalayaa dhimma Obbo
Lidatu Ayyaaloorratti MM Abiyyiif
barraa'e maali?

'Dim angen cau tafarndai a bwyta i
ailagor ysgolion'

Task: Tokenization/segmentation

死亡谷测得54.4摄氏度高温
美国加州名胜或破世界纪录

รอยัลลิสต์มาร์เก็ตเพลส: เฟชบุ๊ก
เตรียมดำเนินการกฎหมายกับรัฐบาล
ไทย หลังบังคับบล็อกการเข้าถึงกลุ่มปิดที่
พูดคุยเกี่ยวกับราชวงศ์

We need to split text into words and sentences.

Languages like Chinese or Thai 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

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

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

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

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

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

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

$$P(t_i = V | t_{i-1} = N) = 0.312$$

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, evaluated).

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?

“I made her crouch”, “I cooked duck for her”, “I cooked her [pet] duck (perhaps just for myself)”, ...

“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.”* Who was drinking coffee?

“I made her duck cassoulet”

(Cassoulet = a French bean casserole)

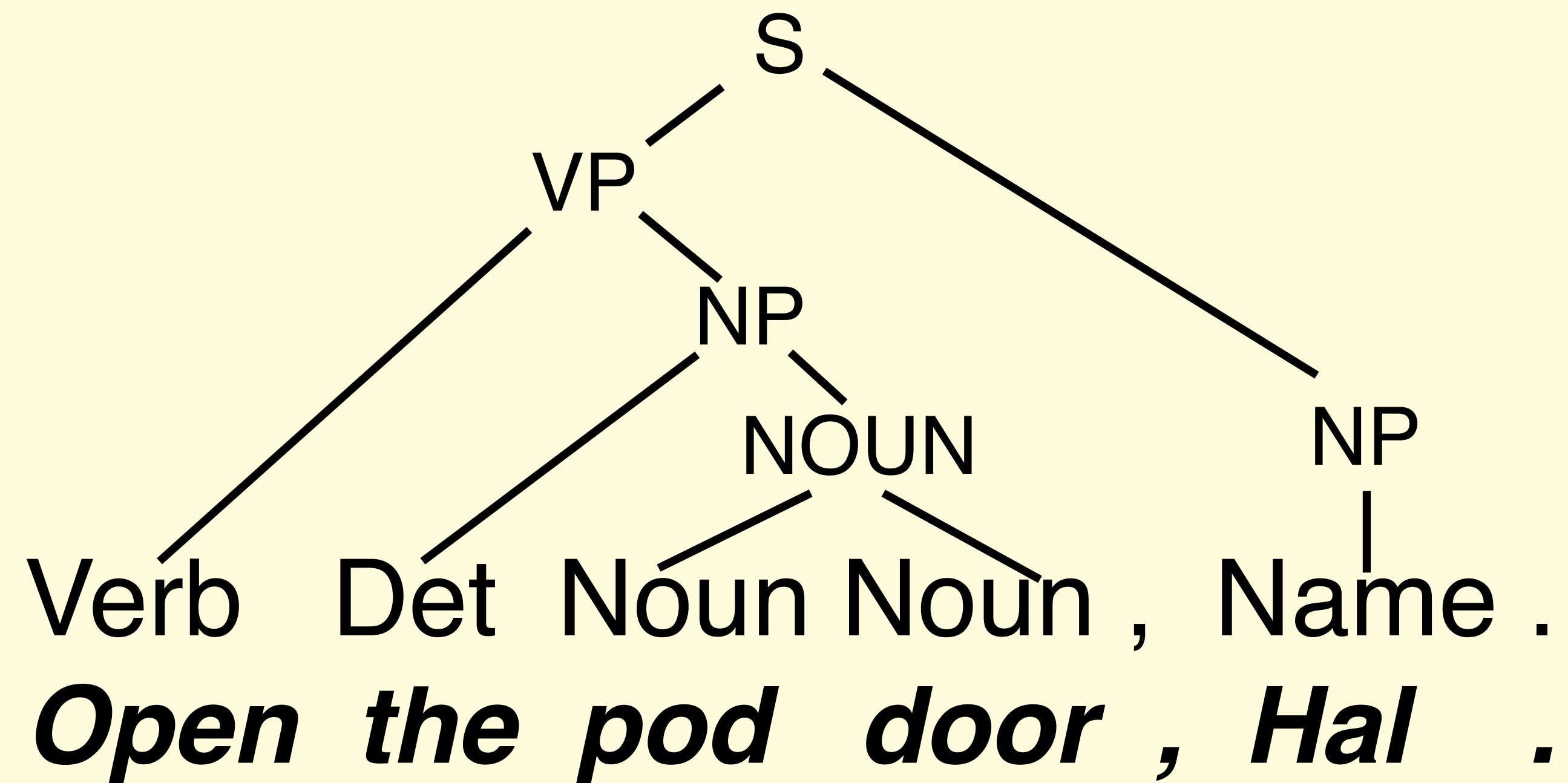
The second major problem in NLP is **coverage**:

We will always encounter **unfamiliar words and constructions**.

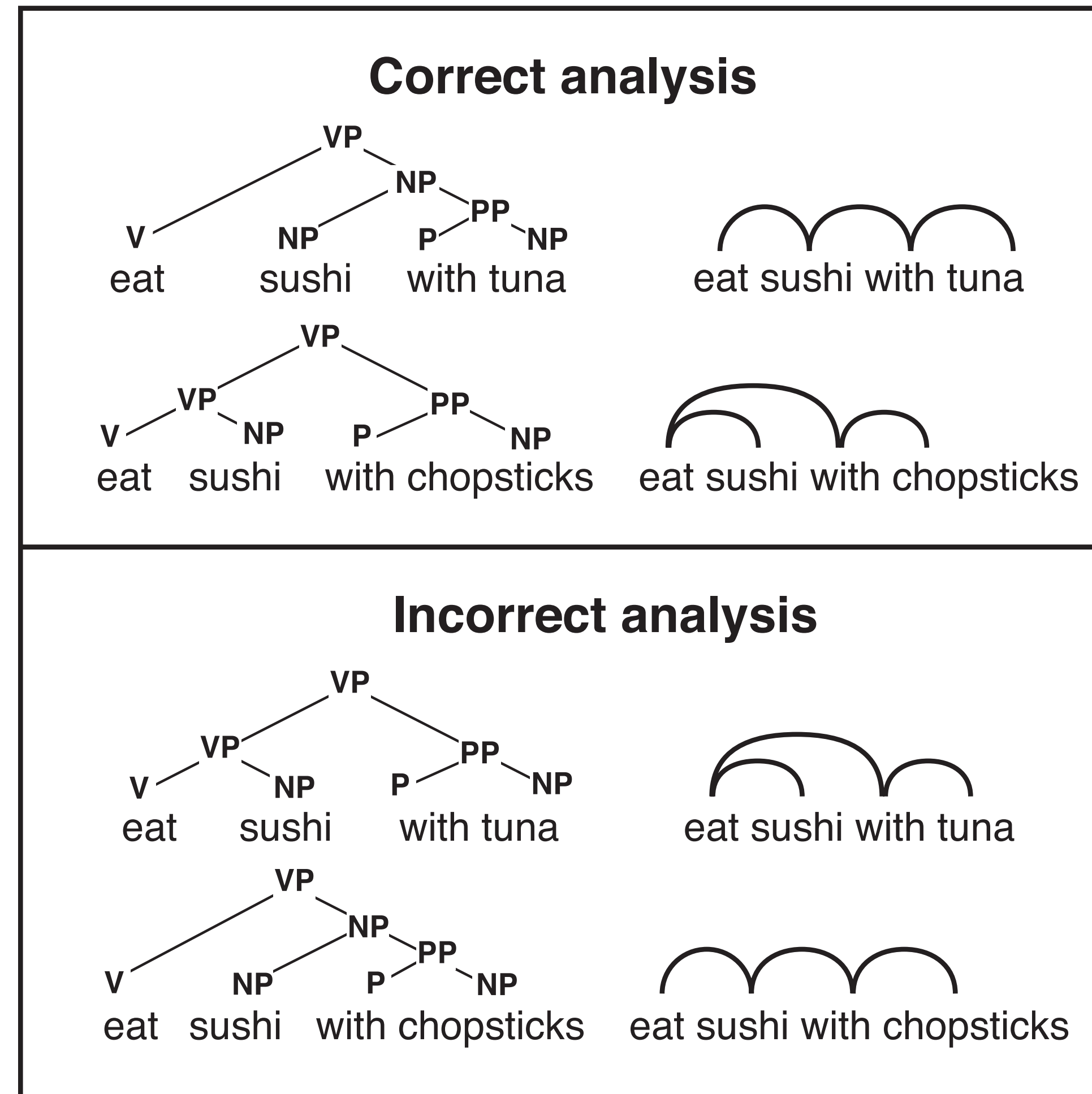
Our models need to be able to deal with this.

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



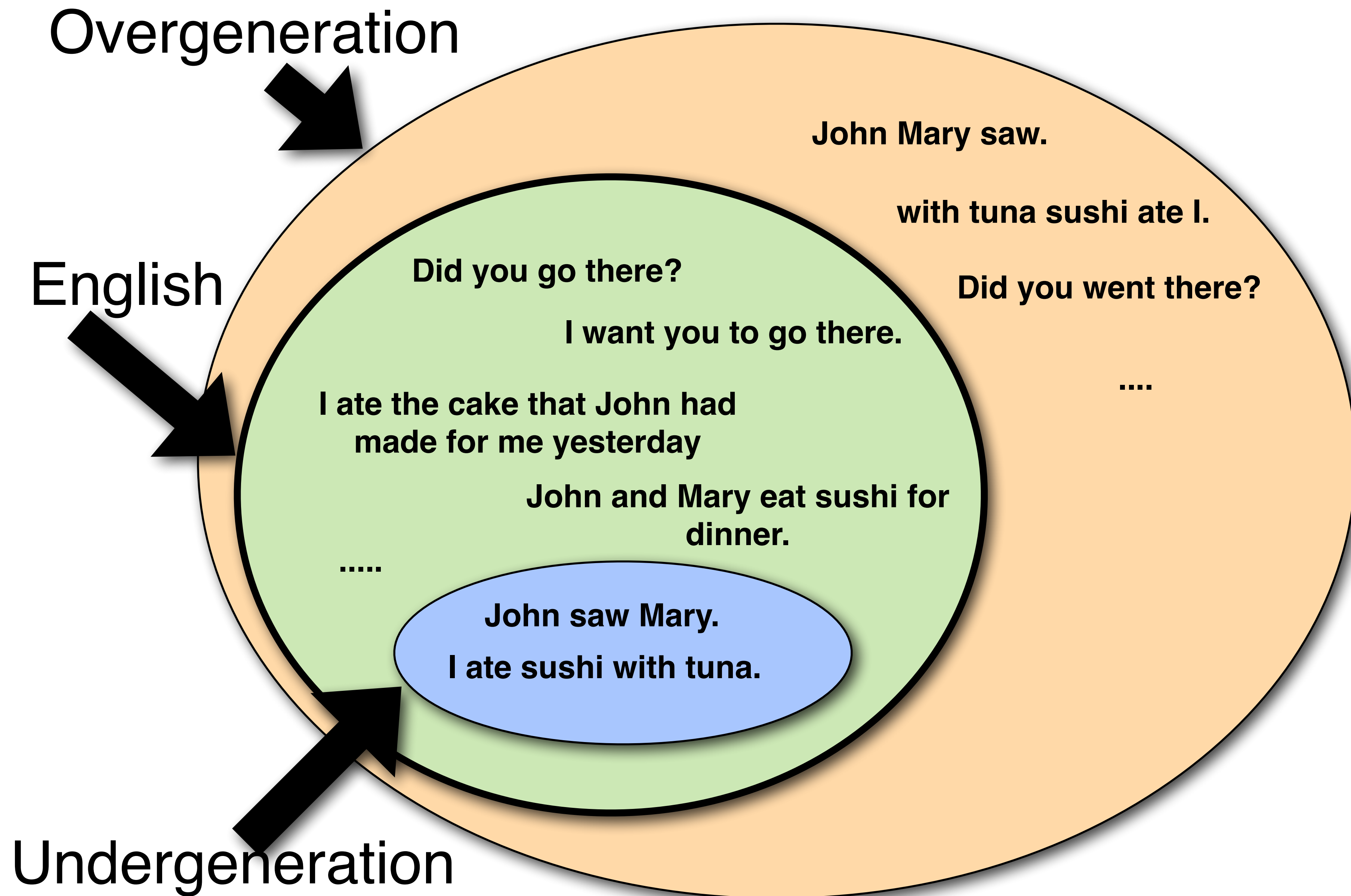
Question: what is grammar?

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,....)



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.

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**

$\exists x \exists y (\text{pod_door}(x) \ \& \ \text{Hal}(y) \ \& \ \text{request}(\text{open}(x, y)))$

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?)

Is Carl Lewis wealthy? What about Jesse Owens?

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

Sidestepping the NLU pipeline

Many current neural approaches for natural language understanding and generation go directly from the raw input to the desired final output.

With large amounts of training data, this often works better than the traditional approach.

- We will soon discuss why this may be the case.

But these models don't solve everything:

- How do we incorporate knowledge, reasoning, etc. into these models?
- What do we do when don't have much training data?
(e.g. when we work with a low-resource language)