COL 772
Natural Language Processing

Instructor: Mausam

(Slides adapted from Heng Ji, Dan Klein, Jurafsky & Martin, Noah Smith, Luke Zettlemoyer)

Personnel

- Instructor: Mausam, SIT 402, mausam@cs.washington.edu

- TAs:
    - Keshav Kolluru
    - Vipul Kumar Rathore

2

Logistics

- Course Website:
  www.cse.iitd.ac.in/~mausam/courses/col772/autumn2020

- Join class discussion group on Piazza (access code col772)
  https://piazza.com/iit_delhi/fall2020/col772/home

- Textbook:
  Yoav Goldberg. Neural Network Methods in Natural Language Processing. Morgan and Claypool (2017).
  Dan Jurafsky and James Martin. Speech and Language Processing,
      2nd Edition, Prentice-Hall (2008).
      3rd Edition (under preparation)
  Course notes by Michael Collins: www.cs.columbia.edu/~mcollins/

- Grading:
  - 30% assignments
  - 20% project
  - 20% midsem
  - 30% Major
  - Extra credit: constructive participation in class and discussion group

Assignments and Project

- **~3 programming assignments**
  - assignments done **individually**!
  - late policy: penalty of 10% maximum grade every day for a week

- **Project**
  - done in teams of two
    - Special permission needed for other team sizes
  - no late submissions allowed

- **Request HPC approval right away!**
  - Please don't complain if HPC is busy
  - TAs: do an extra class on best practices for HPC

Academic Integrity

- Cheating → negative penalty (and possibly more)
  - Exception: if one person/team is identified as cheater
    - Non-cheater gets a zero
- Collaboration is good!!! Cheating is bad!!! Who is a cheater?
  - No sharing of part-code
  - No written/soft copy notes
  - Right to information rule
  - Kyunki saas bhi kabhi bahu thi Rule
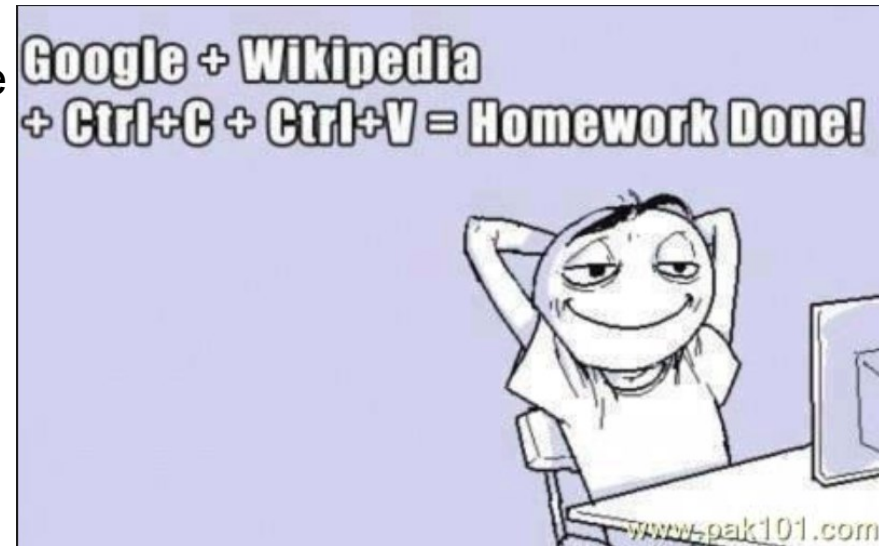
Class Requirements & Prereqs

- **Class requirements**
  - Uses a variety of skills / knowledge:
    - Probability and statistics
    - Deep learning
    - Probabilistic graphical models
    - Basic linguistics background
    - Excellent coding skills
  - Most people are probably missing one of the above
  - You will often have to work to fill the gaps

- **Official Prerequisites**
  - Data structures

- **Unofficial Prerequisites**

Unofficial Prerequisites

(1) program and year,
(2) number of background courses taken,
(3) grades in background courses,
(4) any related non-background courses taken,
(5) If your existing project benefits from the course,
(6) any NLP experience you may have,
(7) your overall GPA.

A willingness to learn whatever background you are missing

Form: https://forms.gle/CGVqQfQvDgK4uSUx8

Status: 210 students applied. 170+ filled the form. 50 seats!
I gotta make hard decisions. Sorry.

Timings

- Tue/Fri 2-3:30
  - Availability after class?

- Office hours
  - By appointment

- Audit criteria
  - no audit

Goals of this course

- Learn the issues and techniques of modern NLP
- Build realistic NLP tools
- Be able to read current research papers in the field
- See where the holes in the field still are!

- Computer Engineer
  - very relevant field in the modern age

- Computer Scientist
  - an excellent source of research problems

Theory vs. Modeling vs. Applications

- Lecture balance tilted towards modeling

- Assignment balance tilted towards applications

- ~No theorems or proofs

- Desired work – lots!

What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Probabilistic/Neural Modeling Techniques
    - Increasingly complex model structures/architectures
    - Learning and parameter estimation
    - Inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice…

# Three Way Crossing in ML
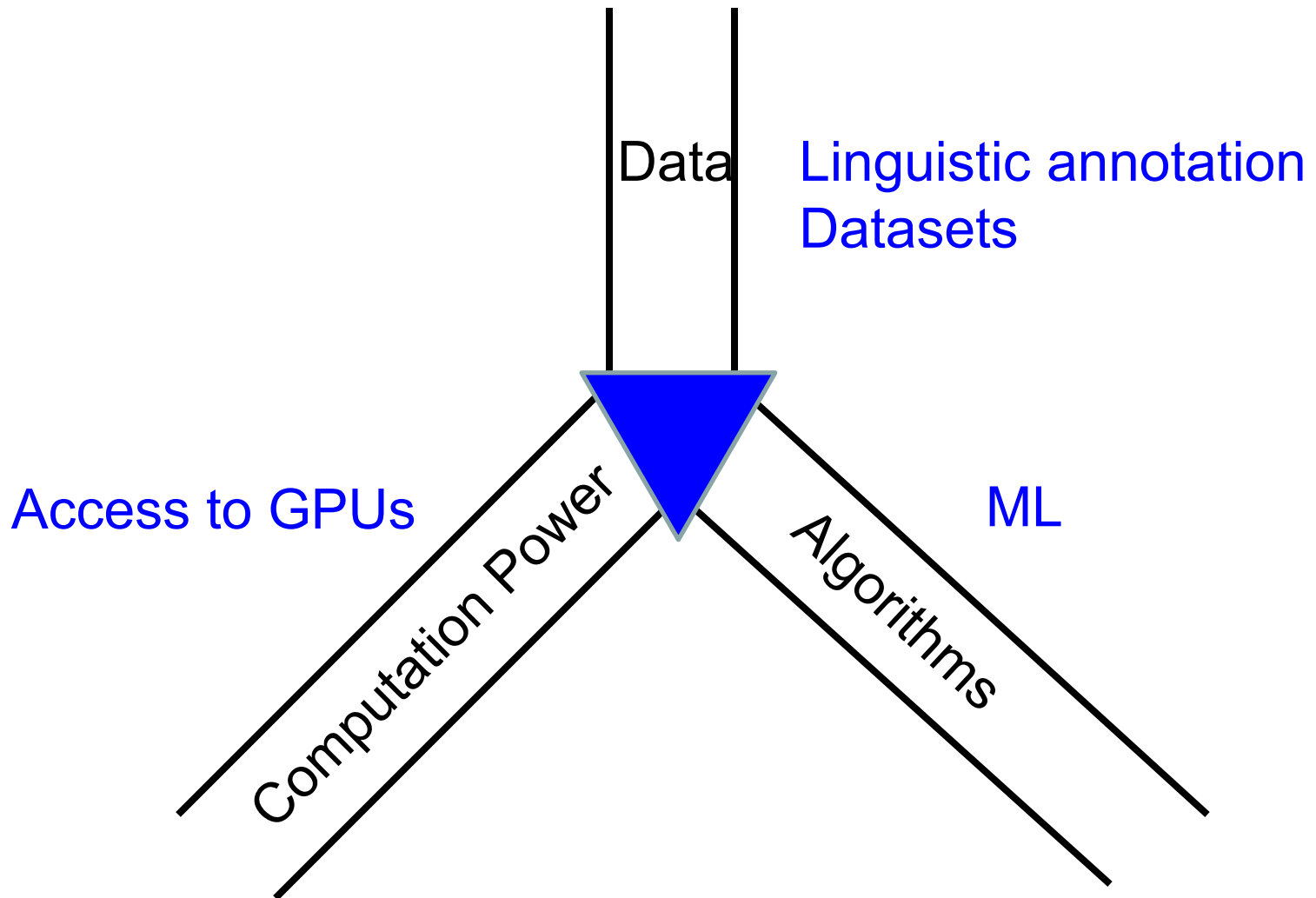
MOTIVATION

The Dream

- **It'd be great if machines could**
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- **But they can't:**
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge

The mystery

- What's now impossible for computers (and any other species) to do is effortless for humans
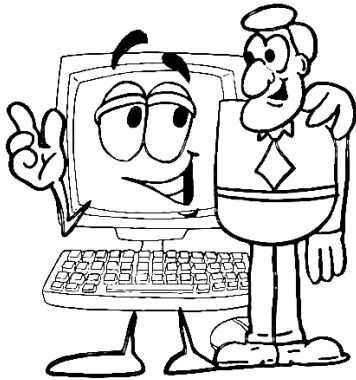
The mystery (continued)

- Patrick Suppes, eminent philosopher, in his 1978 autobiography:

  "…the challenge to psychological theory made by linguists to provide an adequate theory of language learning may well be regarded as the most significant intellectual challenge to theoretical psychology in this century."

- So far, this challenge is still unmet in the 21st century

- Natural language processing (NLP) is the discipline in which we study the tools that bring us closer to meeting this challenge

What is NLP?

- **Fundamental goal: *deep* understanding 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, dialog interfaces, question answering…
  - Unknown: human-level comprehension

NLP for Big Data

- ## Huge Size
  - Google processes 20 PB a day (2008)
  - Facebook has 300+ PB of user data + 4 PB/day (2020)

- ## 80% data is unstructured (IBM, 2010)
- ## More importantly, Heterogeneous
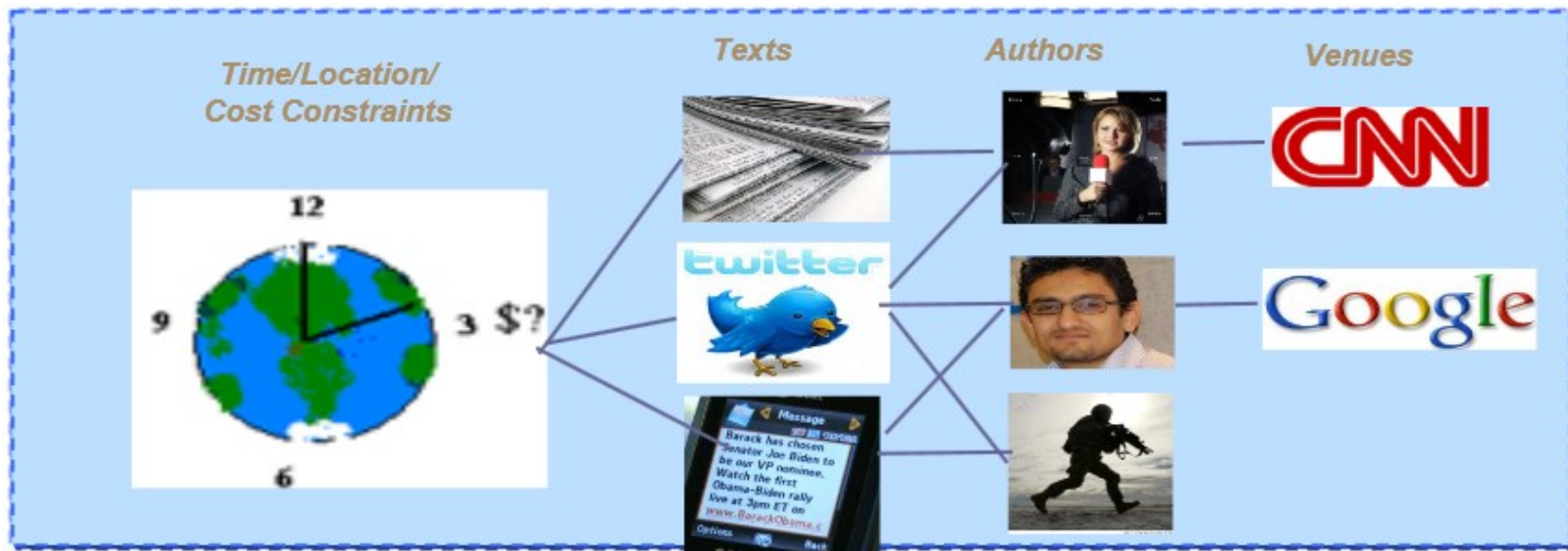
Some Early NLP History

- ## 1950s:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military (imagine that)
    - Toy models: MT using basically word-substitution
  - Optimism!

- ## 1960s and 1970s: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
    - "Little John was looking at his toy box. Finally, he found it. The box was in the pen. John was very happy."
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)

SHRDLU

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK.
Person: What does the box contain?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: What is the pyramid supported by?
Computer: THE BOX.
Person: How many blocks are not in the box?
Computer: FOUR OF THEM.
Person: Is at least one of them narrower than the one which i told you to pick up?
Computer: YES, THE RED CUBE.
Person: Is it supported?
Computer: YES, BY THE TABLE.
Person: Can the table pick up blocks?
Computer: NO.

Not yet solved: CLEVR (2016)

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.
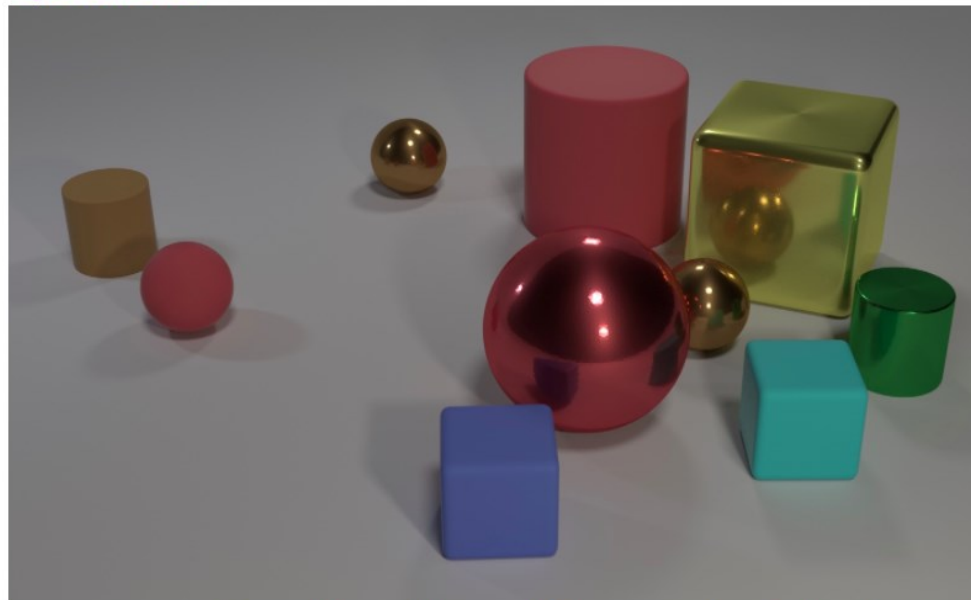
Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?

Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made o the same material as** the **small red sphere**?

Q: **How many** objects are **either** **small cylinders** or **red** things?

NLP History: pre-statistics

- (1) Colorless green ideas sleep furiously.
- (2) Furiously sleep ideas green colorless
    - It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not." (Chomsky 1957)
- 70s and 80s: more linguistic focus
    - Emphasis on deeper models, syntax and semantics
    - Toy domains / manually engineered systems
    - Weak empirical evaluation

NLP: machine learning and empiricism

"Whenever I fire a linguist our system performance improves." –Jelinek, 1988

- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: Probabilistic NLP gives way to Neural NLP

Three Generations of NLP

- Hand-crafted Systems – Knowledge Engineering [1950s–]
  - Rules written by hand; adjusted by error analysis
  - Require experts who understand both the systems and domain
  - Iterative guess-test-tweak-repeat cycle
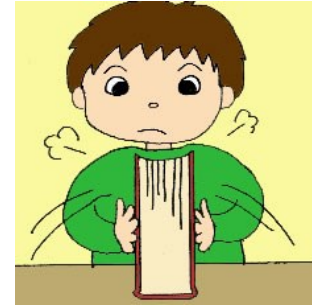
- Automatic, Trainable (Machine Learning) Systems with engineered features [1985s–2012]
  - The tasks use statistical models with hand-coded features
  - More robust techniques based on rich annotations
  - Perform better than rules (Parsing 90% vs. 75% accuracy)

- Automatic, Trainable Neural architectures with no/limited engineered features [2012--]
  - The tasks use deep learning
  - Perform better than hand-coded features but not by too much

What is Nearby NLP?

- ## Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it
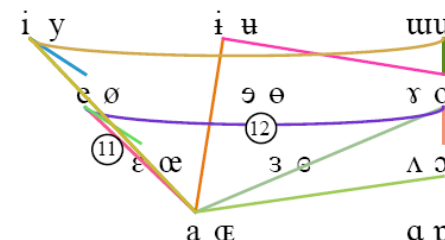
- ## Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!
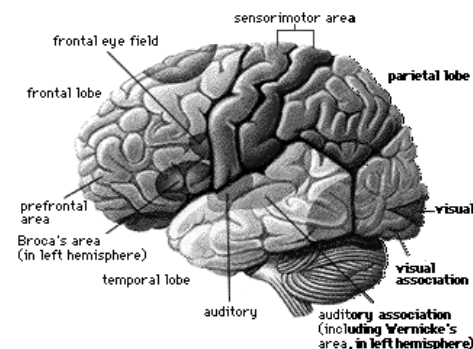
- ## Speech?
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP
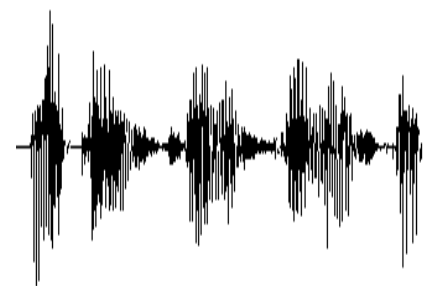
Web Search … n.0

find all web pages containing
the word Liebermann

read the last 3 months of
the NY Times and provide
a summary of Trump's
presidency so far

Caveat

NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with exact solutions/algorithms
- We can't let either of those facts get in the way of making progress

Speech and Language Processing
- Jurafsky and Martin

NLP is AI Complete

Turing Test

young woman: Men are all alike.
eliza: In what way?
young woman: They're always bugging us about something specific or other.
eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?

ELIZA (Weizenbaum, 1966): first computer dialogue system based on keyword matching

Levels of Linguistic Knowledge

speech

text

phonetics

orthography

phonology

morphology

lexemes

"shallower"

syntax

semantics

"deeper"

pragmatics

discourse
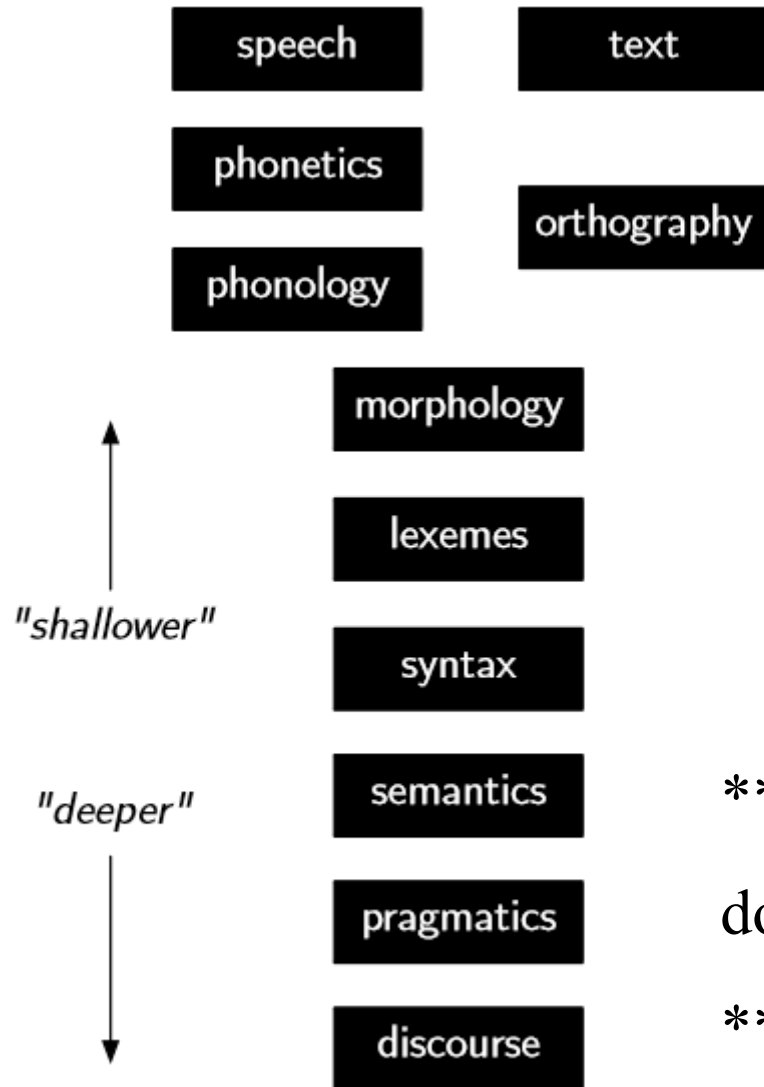
**we can go up, down and up and down and combine steps too!!

**every step is equally complex

Orthography

ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจาก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....

Morphology

uygarlaştıramadıklarımızdanmışsınızcasına
"(behaving) as if you are among those whom we could not civilize"

TIFGOSH ET HA-LELED BA-GAN
"you will meet the boy in the park"

unfriend, Obamacare, Manfuckinghattan

Suffix, prefix, infix (tmesis), circumfix

arundhati "leftist" roy

छुड़ाया chuRAyA → छुड़वाया chhurwAyA

The Challenge of Words

- Segmenting text into words
  - Thai example
- Morphological variation
  - Turkish and Hebrew examples
- Words with multiple meanings:
  - bank, mean
- Domain-specific meanings:
  - latex, sport, etc.
- Multiword expressions:
  - make a decision, take out, make up

POS Tagging

ikr   smh   he   asked   fir   yo   last   name

so   he   can   add   u   on   fb   lololol

Morphology+Syntax

A ship-shipping ship, shipping shipping-ships.

Syntactic Analysis
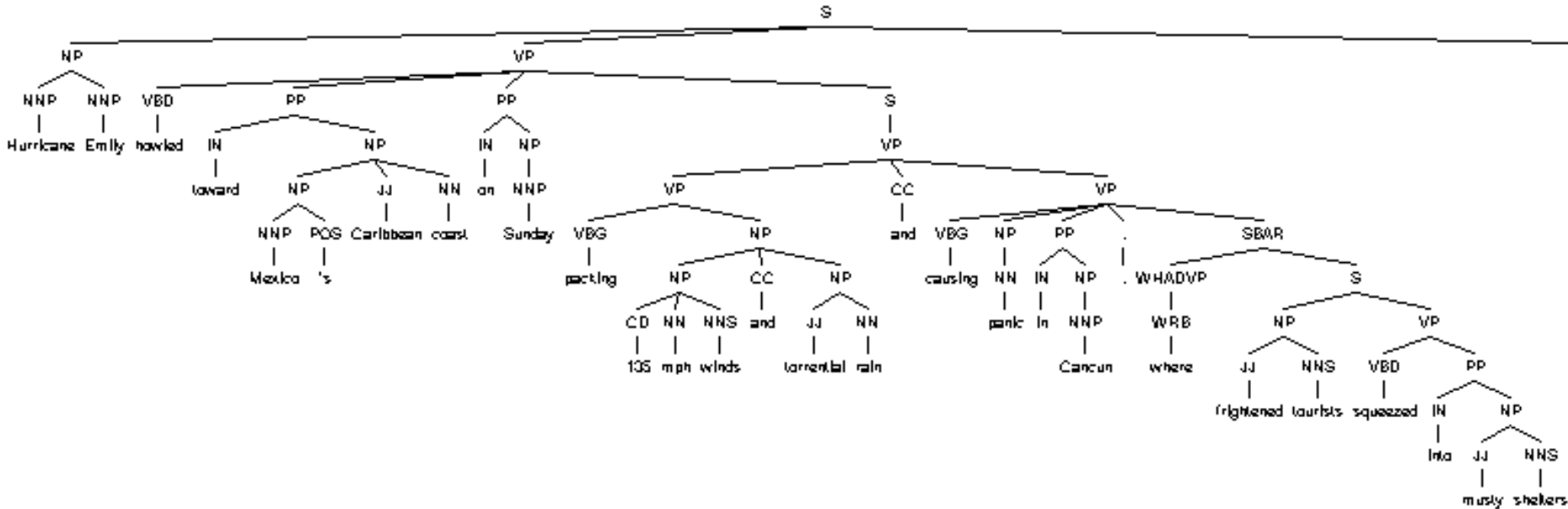
Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- SOTA: ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

Dependency Parsing
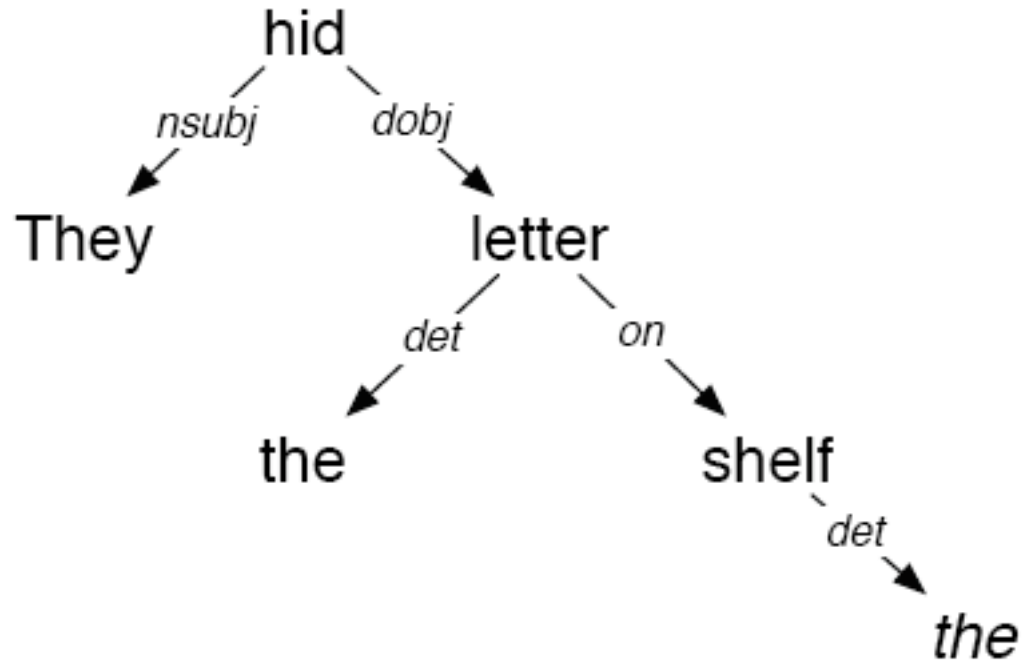
They hid the letter on the shelf

Dependency Relations

Argument Dependencies	Description
nsubj	nominal subject
csubj	clausal subject
dobj	direct object
iobj	indirect object
pobj	object of preposition
Modifier Dependencies	**Description**
tmod	temporal modifier
appos	appositional modifier
det	determiner
prep	prepositional modifier

Constituency Parsing
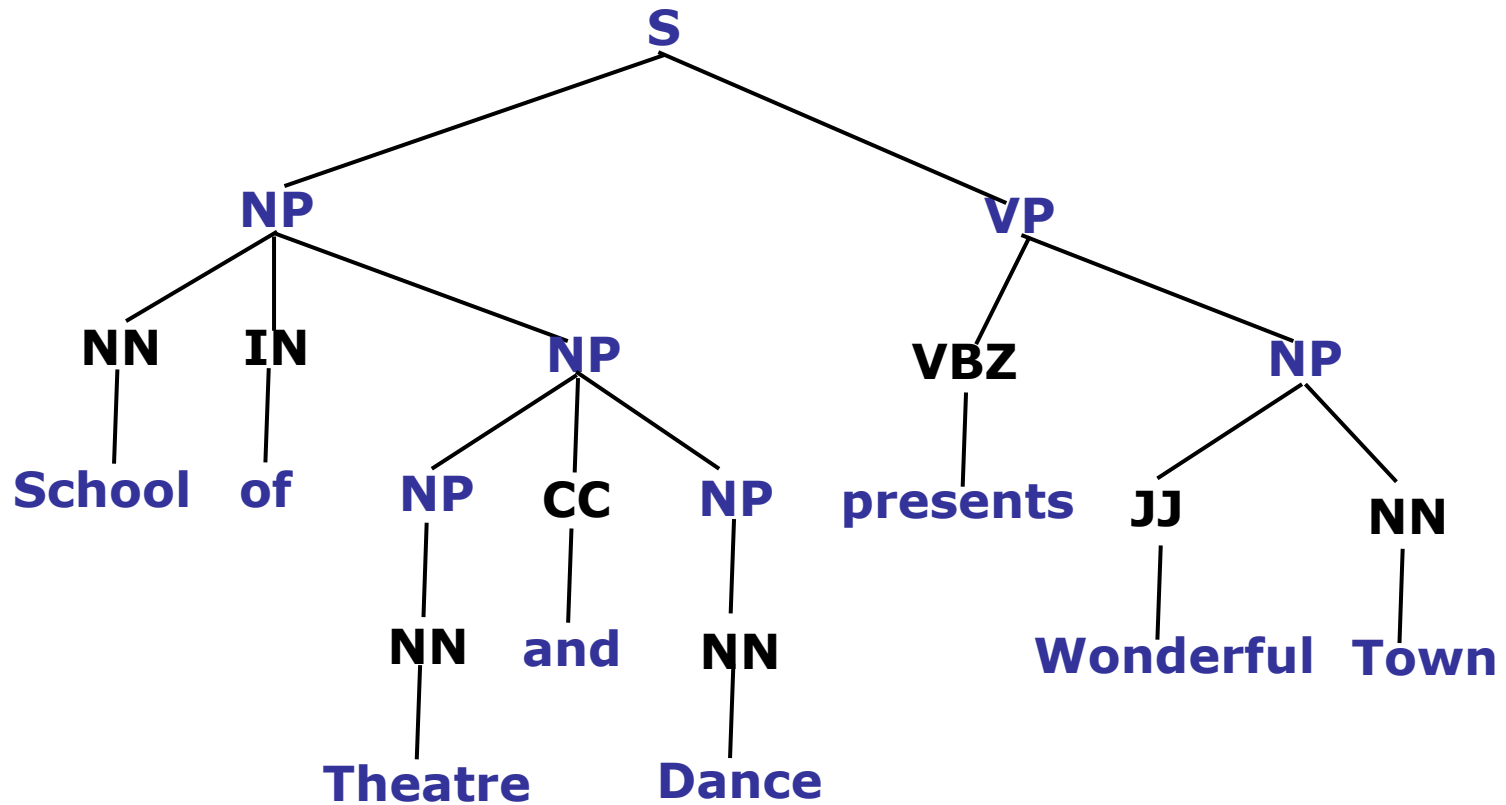
Semantics

- ## Lexical Semantics
  - Meanings of words

- ## Frame semantics
  - (shallow semantics)

- ## Sentential semantics
  - (full semantics)

Lexical Semantics

- ## Synonym
  - start-begin, bodoh-dungu, next to-near

- ## Antonym
  - clever x stupid, high x low, big x small, etc.

- ## Homonym
  - identical in spelling and pronounciation
  - bear, bank, …

- ## Homophones
  - sounds identical but are written differently
  - site-sight, piece-peace.

- ## Homograph
  - written identically but sound differently
  - lead, wind, minute

- ## Polysemy
  - a word which has two(or more) related meanings

Semantic Role Labeling: Adding Semantics into Trees
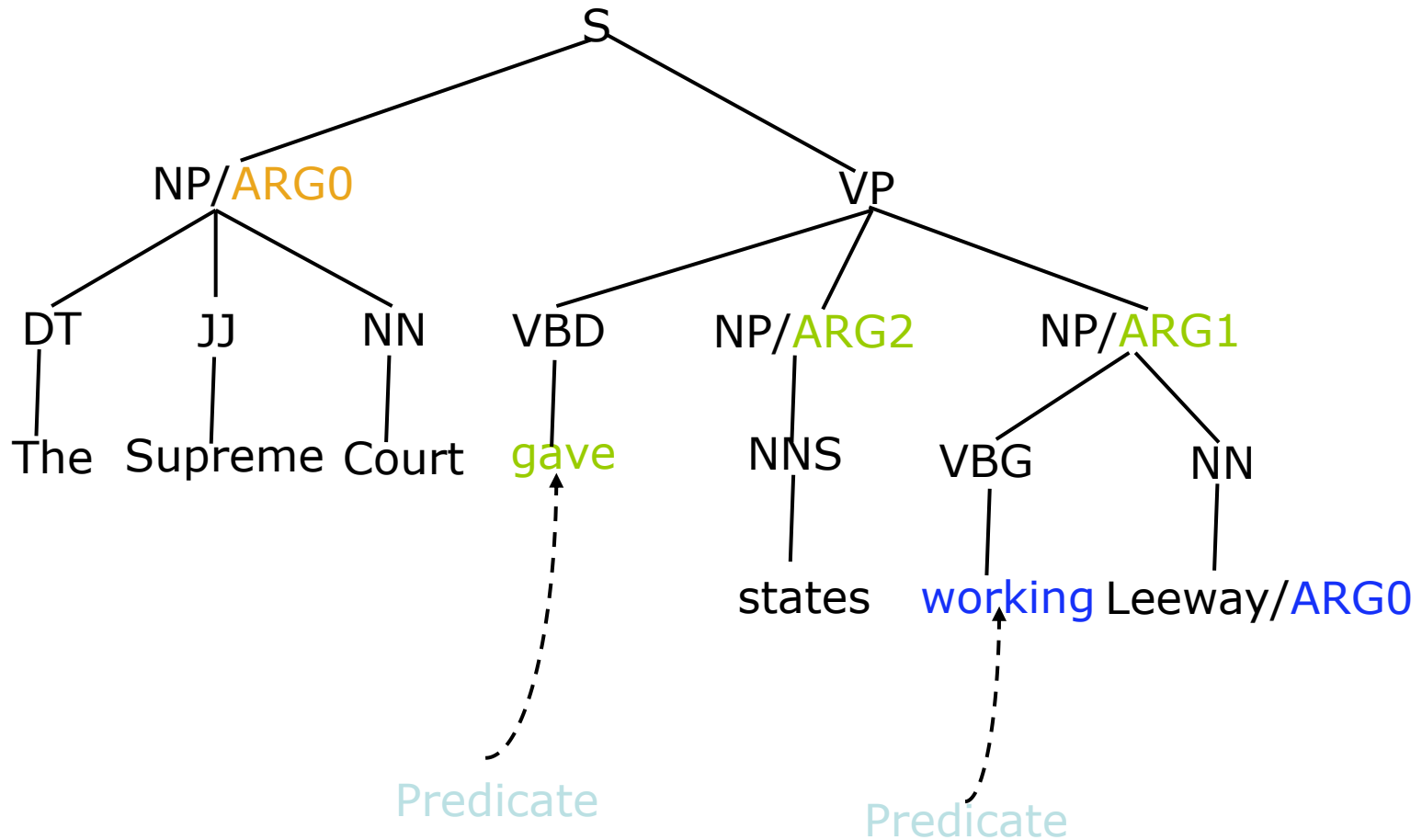
Core Arguments

- Arg0 = agent
- Arg1 = direct object / theme / patient
- Arg2 = indirect object / benefactive / instrument / attribute / end state
- Arg3 = start point / benefactive / instrument / attribute
- Arg4 = end point

Full Sentential Semantics

- **Formal/Logical Semantics**

  - All man are mortal

  - India borders a country that borders Afghanistan.

Coreference

- But the little prince could not restrain admiration:

- "Oh! How beautiful you are!"

- "Am I not?" the flower responded, sweetly. "And I was born at the same moment as the sun . . ."

- The little prince could guess easily enough that she was not any too modest--but how moving--and exciting--she was!

- "I think it is time for breakfast," she added an instant later. "If you would have the kindness to think of my needs--"

- And the little prince, completely abashed, went to look for a sprinkling-can of fresh water. So, he tended the flower.

Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 80% accuracy for pre-defined tables, 90%+ for single easy fields
- But remember: information is redundant!

Temporal Information Extraction

- *In **1975**, after being fired from Columbia amid allegations that he used company funds to pay for his son's bar mitzvah, **Davis** founded **Arista***
  - Is '1975' related to the employee_of relation between Davis and Arista?
  - If so, does it indicate START, END, HOLDS… ?

- Each classification instance represents a temporal expression in the context of the entity and slot value.

- We consider the following classes
  - START   *Rob joined Microsoft in 1999.*
  - END   *Rob left Microsoft in 1999.*
  - HOLDS   *In 1999 Rob was still working for Microsoft.*
  - RANGE   *Rob has worked for Microsoft for the last ten years.*
  - NONE   *Last Sunday Rob's friend joined Microsoft.*

NLP+Knowledge Base

pop recording artist, actress and fashion model. Born and raised in Houston, Texas, she enrolled in various performing
as first exposed to singing and dancing competitions as a child. Knowles rose to fame in the late 1990s as the lead
irl group Destiny's Child, one of the world's best-selling girl groups of all time.

Destiny's Child, Knowles released her debut solo album *Dangerously in Love* (2003), which spawned the number one
and "Baby Boy" and became one of the most successful albums of that year, earning her a then record-tying five
[3] Following the group's disbandment in 2005, Knowles released *B'Day* in 2006. It debuted at number one on the
d included the hits "Déjà Vu", "Irreplaceable" and "Beautiful Liar". Her third solo album *I Am... Sasha Fierce*, released in
luded the anthemic "Single Ladies (Put a Ring on It)". The album and its singles earned her six Grammy Awards,
for most Grammy Awards won by a female artist in one night.[4][5][6] Knowles is one of the most honored artists by the
l among female artists, with 16 awards—13 as a solo artist and three as a member of Destiny's Child.[7][8]

acting career in 2001, appearing in the musical film *Carmen: A Hip Hopera*. In 2006, she starred in the lead role in the
e 1981 Broadway musical *Dreamgirls*, for which she earned two Golden Globe nominations. Knowles launched her
, House of Deréon, in 2004, and has endorsed such brands as Pepsi, Tommy Hilfiger, Armani and L'Oréal. In 2010,
wles at number two on its list of the 100 Most Powerful and Influential Celebrities in the world;[9][10] she was also listed
l and influential musician in the world.[11] *Time* also included Knowles on its list of the "100 Most Influential People in the

- **Knowledge Base (KB)**
  - Attributes (a.k.a., "slots") derived from Wikipedia infoboxes are used to create the reference KB

- **Source Collection**
  - A large corpus of unstructured texts

Beyoncé Knowles

Knowles performing in 2007

Background information

Birth name	Beyoncé Giselle Knowles
Born	September 4, 1981 (age 29)[1]
Origin	Houston, Texas, United States
Genres	R&B, pop, soul
Occupations	Singer-songwriter, record producer, actress, dancer, choreographer, model, video director
Instruments	Vocals
Years active	1997–present

Knowledge Base Linking (Wikification)

Query = "James Parsons"

Knowledge Base Population (Slot Filling)


```
<query id="SF114">
    <name>Jim Parsons</name>
    <docid>eng-WL-11-174592-12943233</docid>
    <enttype>PER</enttype>
    <nodeid>E0300113</nodeid>
    <ignore>per:date_of_birth
            per:age per:country_of_birth
            per:city_of_birth</ignore>
</query>
```

Jim Parsons, a graduate ~~School Attended: University of Houston~~ er and Dance, won the Emmy on Sunday for Lead Actor in a Comedy Series for his work on The Big Bang Theory.

School Attended: University of Houston

Jim Parsons

Parsons in 2008

Born	James Joseph Parsons March 24, 1973 (age 37) Houston, Texas, U.S.
Occupation	Actor
Years active	2000–present

KB Slots

Person		Organization
per:alternate_names	per:title	org:alternate_names
per:date_of_birth	per:member_of	org:political/religious_affiliation
per:age	per:employee_of	org:top_members/employees
per:country_of_birth	per:religion	org:number_of_employees/members
per:stateorprovince_of_birth	per:spouse	org:members
per:city_of_birth	per:children	org:member_of
per:origin	per:parents	org:subsidiaries
per:date_of_death	per:siblings	org:parents
per:country_of_death	per:other_family	org:founded_by
per:stateorprovince_of_death	per:charges	org:founded
per:city_of_death		org:dissolved
per:cause_of_death		org:country_of_headquarters
per:countries_of_residence		org:stateorprovince_of_headquarters
per:stateorprovinces_of_residence		org:city_of_headquarters
per:cities_of_residence		org:shareholders
per:schools_attended		org:website

QA / NL Interaction

- **Question Answering:**
  - More than search
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"

- **Natural Language Interaction:**
  - Understand requests and act on them
  - "Make me a reservation for two at Quinn's tonight"

Summarization

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

- **Very context-dependent!**

- **An example of analysis with generation**

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

Machine Translation

"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes

"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959
Video Anniversary of the Tibetan rebellion: China on guard

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency vs fidelity

2013 Google Translate: French

2013 Google Translate: Russian

English -- Russian

- ***The spirit is willing but the flesh is weak. (English)***
- ***The vodka is good but the meat is rotten. (Russian)***

And more recently (FB: Bangla → English)

- The conversation went like this:
- End of the day conversation:
- Me: What did you do today?
- Ma: I watched Kedarnath.
- Me: What did you learn from watching Kedarnath?
- Ma: That Sushant Singh Rajput is very strong.
- Me: How did you know this?
- Ma: He was lugging people around on his back.
- Me: What else did you learn?
- Ma (thinks for a bit): That Saif Ali Khan is very "paaka" (Paaka refers to someone who does things before it is appropriate for them to do so age-wise).
- Me: How do you know?
- Ma: Such a grown up daughter at such a young age!

Our words at the end of the day:

Me: what did you do today?

Mother: Kedarnath saw.

Me: what did you learn to see?

Mother: there is a lot of force on the good lion rajput.

Me: what do you know?

Many many happy returns of the day.

Me: and what do you know?

Mother (for a while thought): Saif Ali Khan was very ripe.

Me: what do you know?

Mother: such a big girl at such a young age.

Language Comprehension?

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

(A)   afraid that he might lose the fossil
(B)   worried about the implications of his finding
(C)   concerned that he might not get credit for his work
(D)   uncertain about the authenticity of the fossil
(E)   excited about the magnitude of his discovery

Jeopardy! World Champion

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

http://www.youtube.com/watch?v=qpKoIfTukrA

Why is NLP hard?

- Ambiguity
- Ambiguity
- Ambiguity
- Sparsity

Ambiguities

- **Headlines: Why are these funny?**
  - Ban on Nude Dancing on Governor's Desk
  - Iraqi Head Seeks Arms
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Local High School Dropouts Cut in Half
  - Red Tape Holds Up New Bridges
  - Clinton Wins on Budget, but More Lies Ahead
  - Hospitals Are Sued by 7 Foot Doctors
  - Kids Make Nutritious Snacks

Syntactic Ambiguity
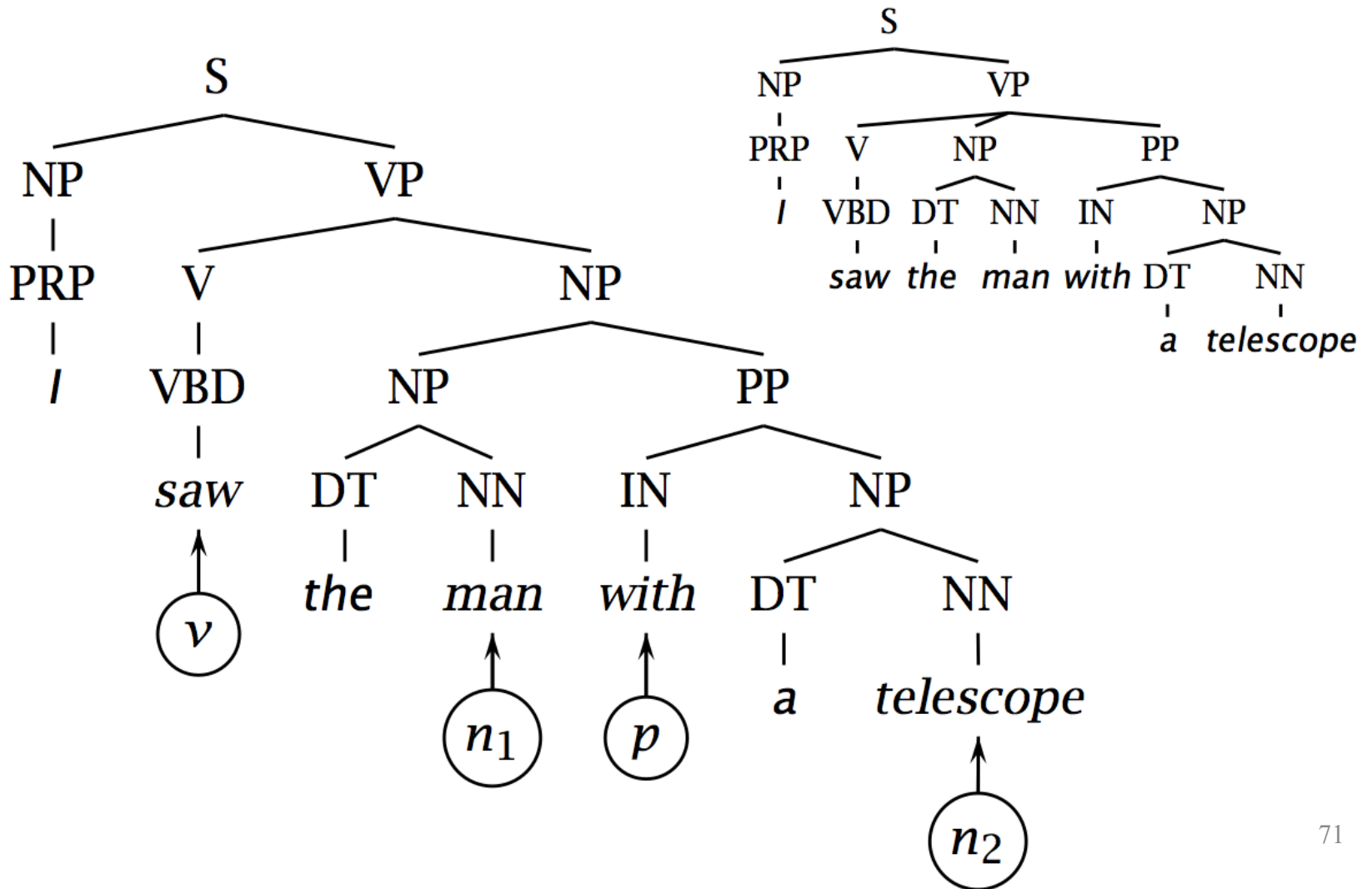
Semantic Ambiguity

At last, a computer that understands you like your mother.

- Direct Meanings:
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
  - It understands you like (it understands) your mother
- But there are other possibilities, e.g. mother could mean:
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
- Context matters, e.g. what if previous sentence was:
  - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients. ☺

[Example from L. Lee]

Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds to the correct parse of

"*This will panic buyers !* "
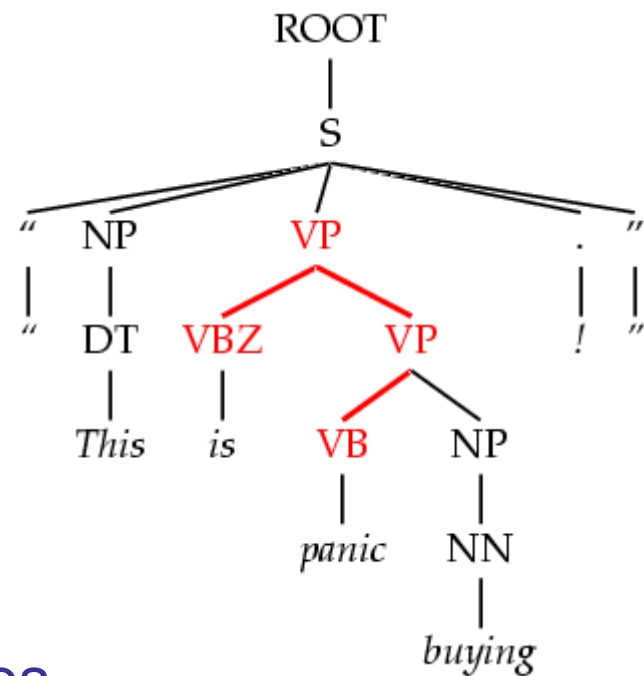
- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

Ambiguities (contd)

- Get the cat with the gloves.
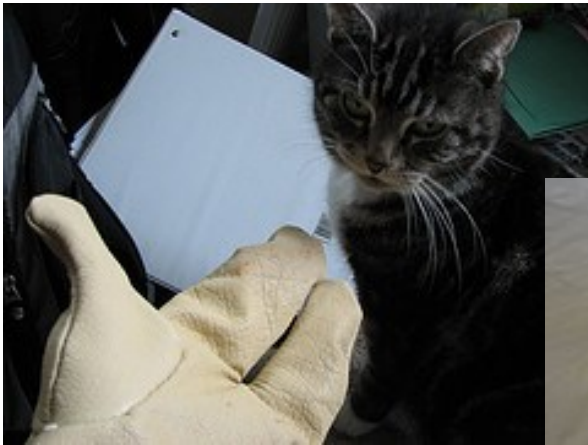
Ambiguity

- ## Find at least 5 meanings of this sentence:
  - ### I made her duck

Speech
and Language Processing -
Jurafsky and Martin

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Speech
and Language Processing -
Jurafsky and Martin

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category:** "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** "make" can mean "create" or "cook"

Speech and Language Processing - Jurafsky and Martin

Ambiguity is Pervasive

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
 - I'm ate or duck
  - I mate or duck

Abbrevations out of context

- Medical Domain: 33% of abbreviations are ambiguous (Liu et al., 2001), major source of errors in medical NLP (Friedman et al., 2001)

RA	"rheumatoid arthritis", "renal artery", "right atrium", "right atrial", "refractory anemia", "radioactive", "right arm", "rheumatic arthritis", …
PN	"Penicillin"; "Pneumonia"; "Polyarteritis"; "Nodosa"; "Peripheral neuropathy"; "Peripheral nerve"; "Polyneuropathy"; "Pyelonephritis"; "Polyneuritis"; "Parenteral nutrition"; "Positional Nystagmus"; "Periarteritis nodosa", …

- Military Domain
  - *"GA ADT 1, USDA, USAID, Turkish PRT, and the DAIL staff met to create the Wardak Agricultural Steering Committee. "*
  - *"DST" = "District Stability Team" or "District Sanitation Technician"?*

Uncertainty: Ambiguity Example
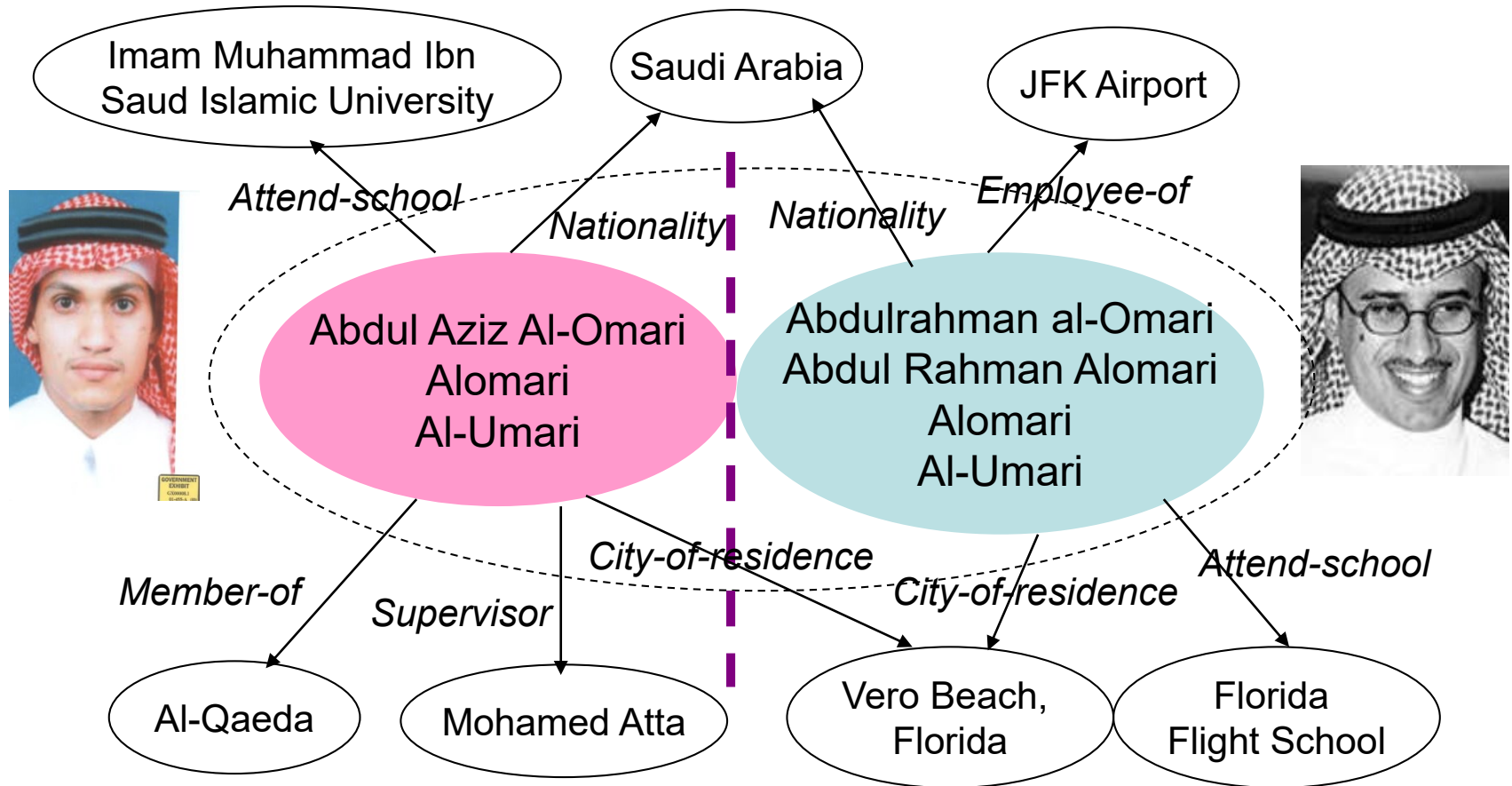
Corpora

- **A corpus is a collection of text**
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora

- **Examples**
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged "balanced" text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what
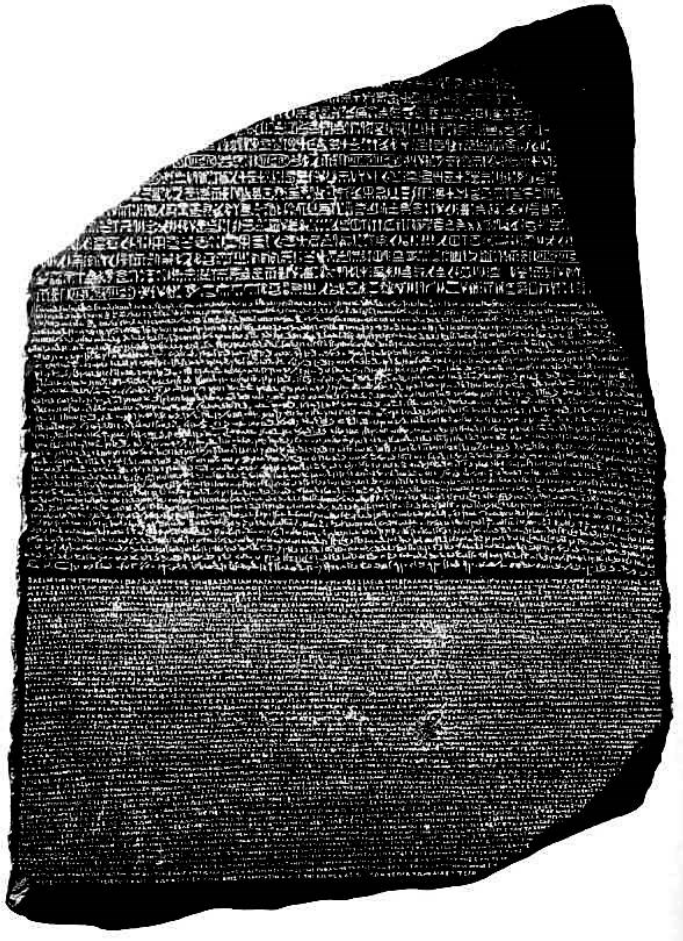
Problem: Sparsity

- ## However: sparsity is always a problem
  - ### New unigram (word), bigram (word pair)
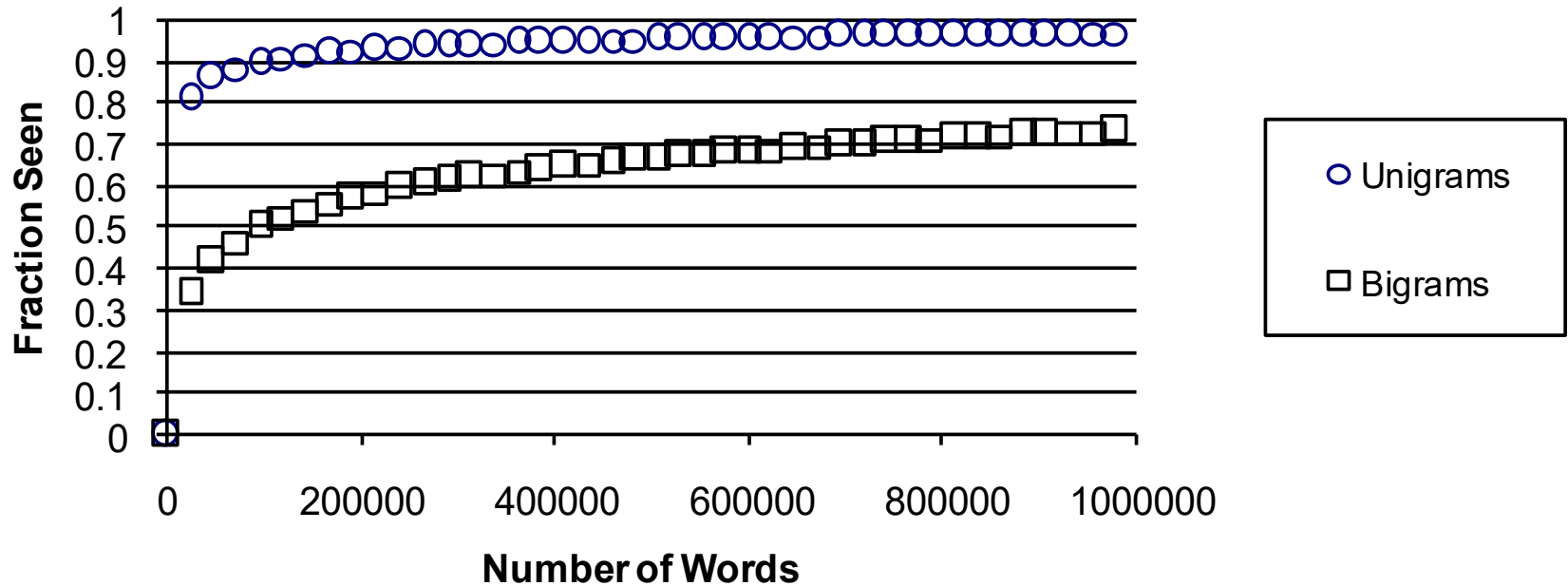
NLP Topics in the Course

- tokenization,
- language models,
- part of speech tagging,
- noun phrase chunking,
- named entity recognition,
- coreference resolution,
- parsing,
- information extraction,
- sentiment analysis,
- question answering,
- text classification,
- document clustering,
- document summarization,

We will make NLP Models

- Model: an abstract, theoretical, predictive construct. Includes:

  - a (partial) representation of the world

  - a method for creating or recognizing worlds

  - a system for reasoning about worlds

    - NLP uses many tools for modeling.

    - Surprisingly shallow models work fine for some applications.

ML Models in the Course

- Naive Bayes,
- Hidden Markov Models,
- Expectation Maximization,
- Conditional Random Fields,
- MaxEnt Classifiers,
- Probabilistic Context Free Grammars,
- Neural Language Models
- Sequence to Sequence models
- Pre-trained Language Models

Disclaimer

- This course will be highly biased
  - won't focus much on linguistics
  - won't focus much on historical perspectives
  - won't be all deep learning

- This course will be highly biased
  - I will teach you what I like
  - I will teach what I can easily learn … ☺