CSL 772
Natural Language Processing

Instructor: Mausam

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

Personnel

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

- TAs:
  - Yashoteja Prabhu, yashoteja.prabhu AT gmail.com
  - Ankit Anand, ankit.s.anand AT gmail.com

Timings

- Mon/Thu 5-6:20
  - Can we do earlier in the day?
- Office hours
  - By appointment

Logistics

- Course Website:
  www.cse.iitd.ac.in/~mausam/courses/csl772/autumn2014
- Join class discussion group on Piazza (access code csl772)
  https://piazza.com/iit_delhi/fall2014/csl772/home
- Textbook:
  Dan Jurafsky and James Martin Speech and Language Processing,
  2nd Edition, Prentice-Hall (2008).
  Course notes by Michael Collins: www.cs.columbia.edu/~mcollins/
- Grading:
  - 30% assignments
  - 20% project
  - 10% Minor 1
  - 10% Minor 2
  - 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

Academic Integrity

- Cheating → negative penalty (and possibly more)
  - Exception: if one person identified as cheater
    - Non-cheater gets a zero
  - http://www.willa.me/2013/12/the-top-five-unsanctioned-software.html

- Collaboration is good!!! Cheating is bad!!!
  - 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
    - Machine learning
    - Probabilistic graphical models
    - Basic linguistics background
    - Decent 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 Prerequsites
  - A willingness to learn whatever background you are missing

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
- Relatively few theorems and even fewer 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 Modeling Techniques
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient 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…

MOTIVATION

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 (is this just NLP?)

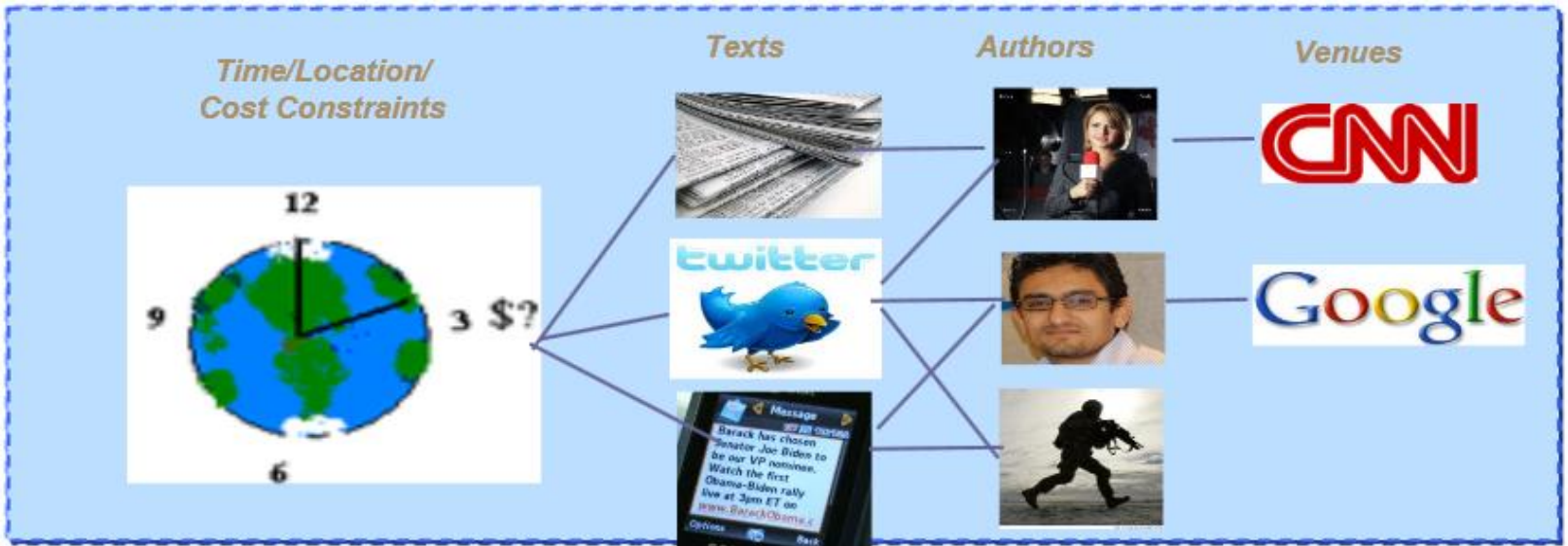
Why Should You Care?

Three trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

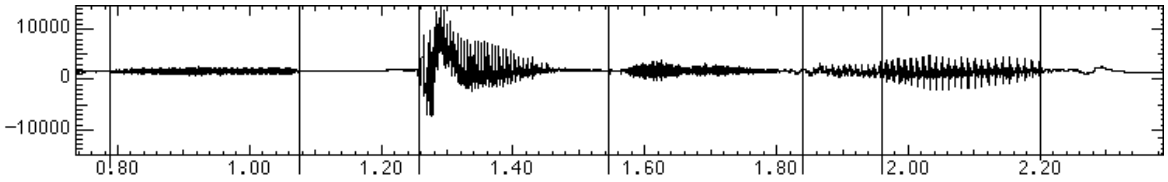
NLP for Big Data

- Huge Size
  - Google processes 20 PB a day (2008)
  - Wayback Machine has 3 PB + 100 TB/month (3/2009)
  - Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
  - eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 80% data is unstructured (IBM, 2010)
- More importantly, Heterogeneous



Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



“Speech Lab”

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)
  - http://www2.research.att.com/~ttsweb/tts/demo.php

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

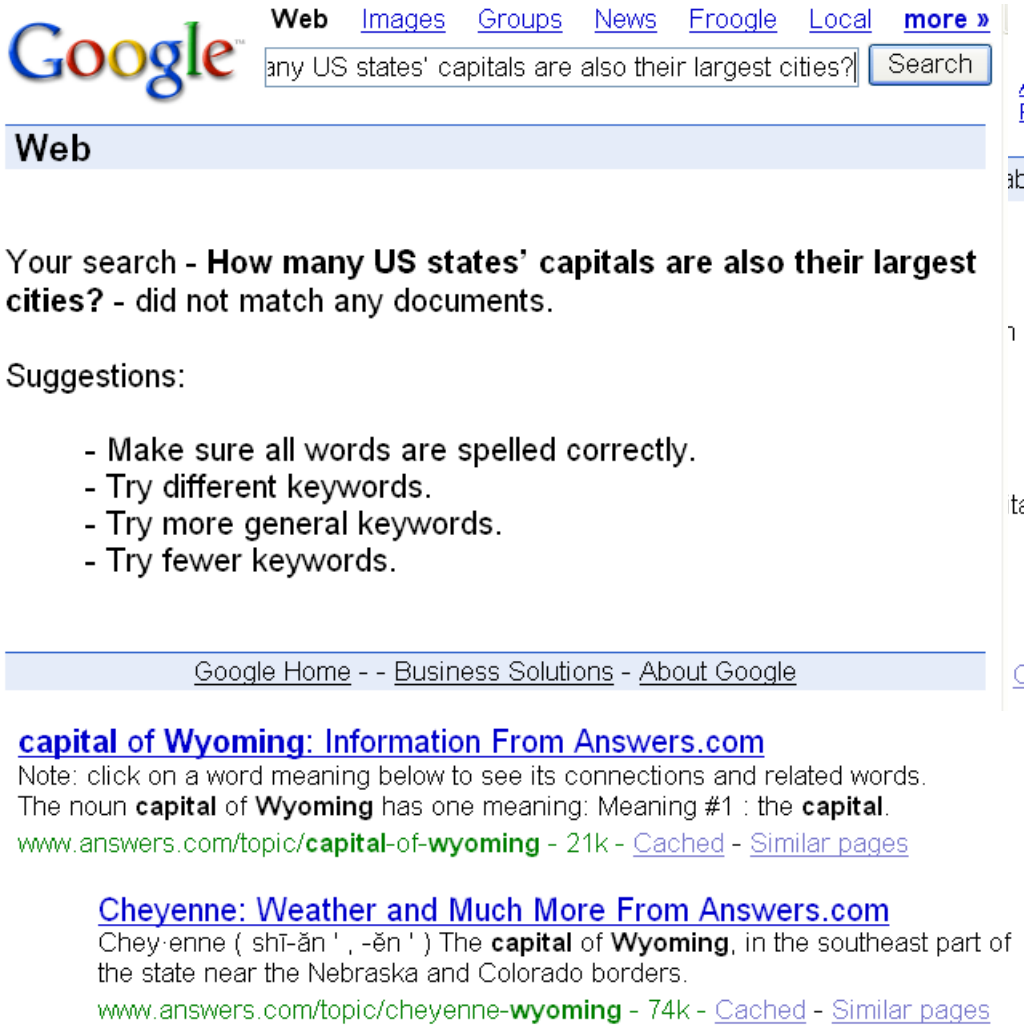
- SOTA: perhaps 80% accuracy for pre-defined tables, 90%+ for single easy fields
- But remember: information is redundant!

Recently!


Home
Tips & Tricks
Features
Search Stories
Playground
Blog
Help
Ginevra de' Benci
1478
The Virgin a...
1508
Adoration o the M...
1481
Feedback
rdo da Vinci
Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. Wikipedia
Born: April 15, 1452, Anchiano
Died: May 2, 1519, Clos Lucé
Buried: Château d'Amboise
Parents: Caterina da Vinci, Piero da Vinci
Structures: Vebjørn Sand Da Vinci Project
The Knowledge Graph
Learn more about one of the key breakthroughs behind the future of search.
See it in action
Discover answers to questions you never thought to ask, and explore collections and lists.

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’’



Hot Area!

WolframAlpha computational... knowledge engine

Oscar for best actress 1958

Examples Random

Assuming year of award ceremony | Use year of film release instead

Input interpretation:

Academy Awards	actress in a leading role
	1958 (year of award ceremony)

Result:

Joanne Woodward in The Three Faces of Eve

Other nominees:

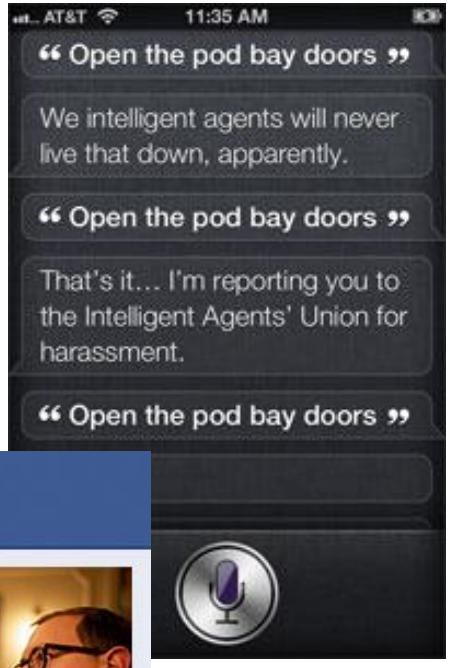
Lana Turner in Peyton Place | Elizabeth Tayl... | Deborah Kerr in Heaven Knows, Mr. Allison | ... the Wind

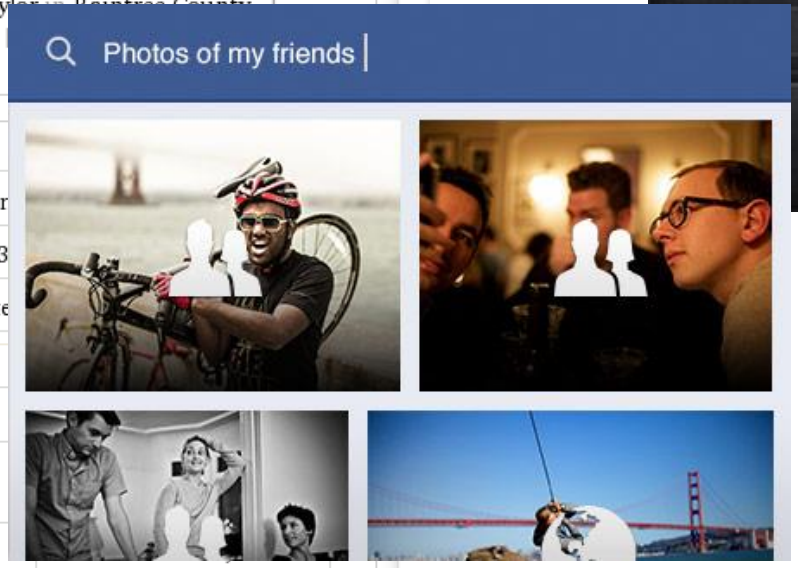
Information about Joanne Woodward:

full name	Joanne Gignilliat Trimmier
date of birth	Thursday February 27, 193
place of birth	Thomasville, Georgia, Unite

Academy Awards and nominations:

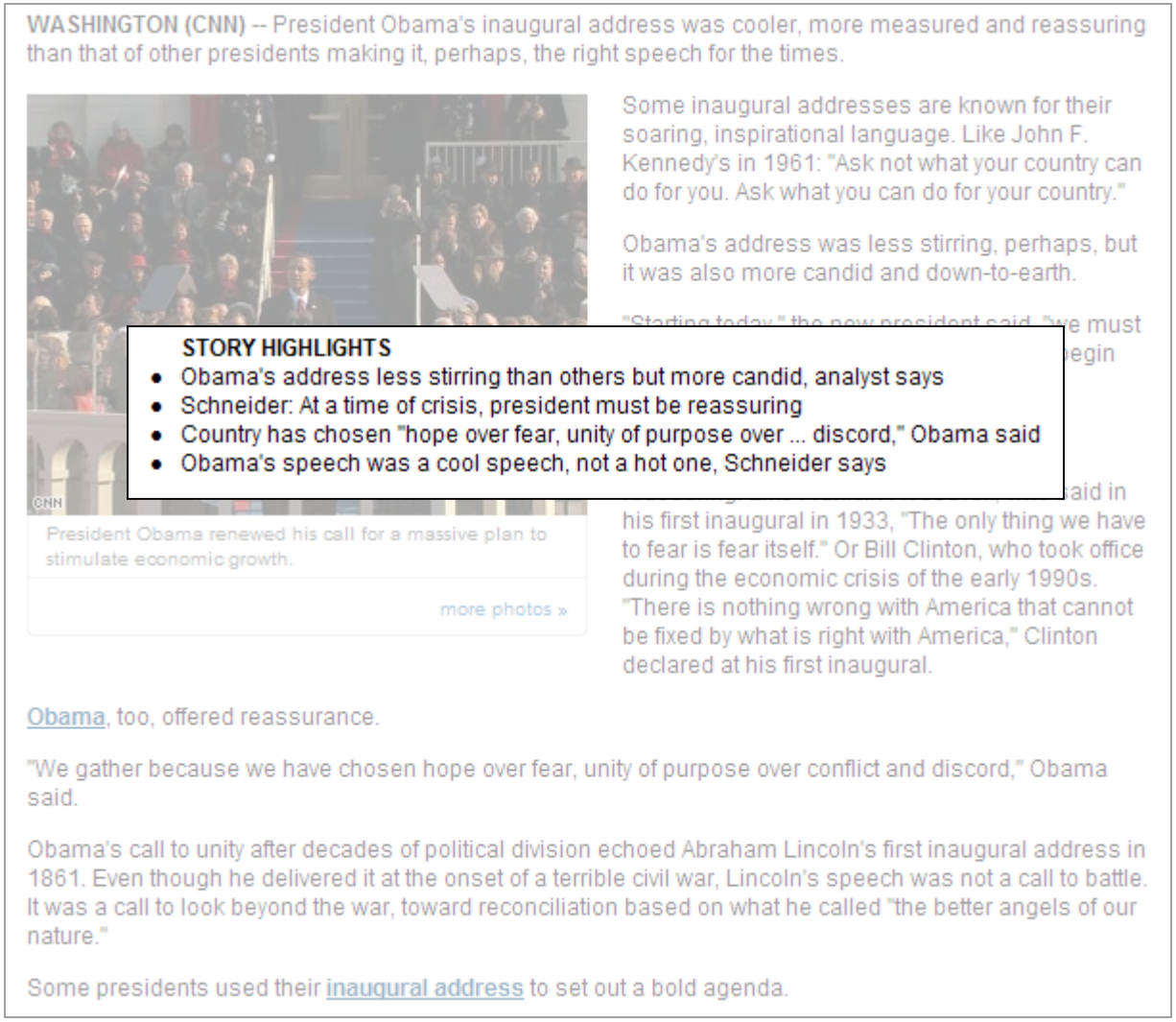
year	category	
1991 (age: 61 years)	actress in a leading role	
1974 (age: 44 years)	actress	Winter Dreams
1969 (age: 39 years)	actress	Rachel, Rachel





Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation



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.

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

more photos »

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

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

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.

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 nature."

Some presidents used their inaugural address to set out a bold agenda.

2013: Summly → Yahoo!

CEO Marissa Mayer announced an update to the app in a blog post, saying, "The new Yahoo! mobile app is also smarter, using Summly's natural-language algorithms and machine learning to deliver quick story summaries. We acquired Summly less than a month ago, and we're thrilled to introduce this game-changing technology in our first mobile application."

Launched 2011, Acquired 2013 for $30M

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

"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

EN CE MOMENT Impôts Kenya Syrie Pakistan Emploi Scandale Prism

Impôt sur le revenu : vous en 2014 ?

Sélectionnez votre revenu et votre situa bénéficiez de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ? VISUEL INTERACTIF
- Un budget 2014 soumis aux critiques

Le chômage baisse pour la première fois depuis avril 2011 POST DE BLOG

AT THIS MOMENT Taxes Kenya Syria Pakistan Use Prism scandal

Income tax: how much do you pay in 2014?

Select your income and family situation to see if you get the tax break.

- How is the budget for 2014 is allocated? INTERACTIVE VISUAL
- A 2014 budget submitted to criticism
- Budget: these expenses no government can reduce
- Budget 2014: the retail savings INTERACTIVE VISUAL

Unemployment fell for the first time since April 2011 POST BLOG

Surviving in the Central time looting and anarchy

CALL FOR EVIDENCE

Member (s) of Europe Ecology-Greens, do you share the finding of severe Christmas Mamère EELV?

Share your experience

Continuous

7:53 Budget: the fixed expenses
7:36 Heard the "Fashion Week" in Paris
7:19 control giant Airbus
7:04 Complaint against "Actual Values"
7:01 Venezuela: 17 people arrested
6:59 Vidberg: the new budget came
6:50 The "noble mission" of the NSA
6:38 Roma: jousting between Brussels a

2013 Google Translate: Russian

pravda.ru

ENG RUS PT ITA

Поиск

Например: Большой Кавказ

Мир | Наука | Общество | Здоровье | Красота |

Новости

20:09
В Шри-Ланке хотели перевезти золото в желудках

20:00
Выходец из России может получить "Нобеля" по химии

19:46
В США установили стандарты торговли оружием

19:35
Директор Эрмитажа: Обыски нанесли ущерб музею

19:25
Мозгу ребенка полезен послеобеденный сон

19:24
Роликом с водителями-детьми заинтересовалась петербургская полиция

19:15
К Марсу приближается "комета века"

18:55
Выявлено более 160 нарушений на судостроительных предприятиях

18:44
Астахов назначен на новый срок в Европейской сети детских омбудсменов

Главное

"Обиженные люди работают, а иностр к нам не поедут"

25.09.2013 19:48

Ректор "Бауманки" Анатолий "Правде.Ру", какие шаги надо чиновникам и ученым в связи реформе РАН.

Фотосессия

Наводнение в Индии: 40 жителей эвакуированы

Найроби. Газета The Independ "Уэстгейт" во время захвата.

Мир

pravda.ru

ENG RUS PT ITA

Поиск

For example, the Greater Caucasus

World | Science | Society | Health | Beauty | Regions | Photo | Video | Forums | archive

News

20:09
In Sri Lanka, wanted to carry the gold in the stomachs

20:00
A native of Russia can get the "Nobel" in Chemistry

19:46
In the United States set the standard arms trade

19:35
Director of the Hermitage: The searches have damaged the museum

19:25
The child's brain is useful afternoon nap

19:24
The roller with the drivers, children become interested in the St. Petersburg Police

19:15
To Mars is approaching "comet of the century"

18:55
There are over 160 violations at shipyards

18:44
Astakhov appointed for a new term in the European Network of Ombudsmen for children

Point

"Mentally ill people are working, and foreign scholars to us will not go"

25/09/2013 19:48

The Rector, "Bauman" Anatoly Alexandrov told with "Pravda.Ru" what steps need to be taken to officials and scientists in connection with the adoption of the law on the reform of the RAS.

Photoshoot

2466 photos

World through the lens: September 25.

Expert: The poorer the society is, the more scandals due to copyright

09/25/2013 20:04

Why Russians are greedy for free, and do not like to pay for downloading movies and music, with "Pravda.Ru" said the head of Liveinternet German Klimenko.

Putin met environmentalists "Greenpeace" trying to grab the platform

25/09/2013 14:39

President of Russia, speaking at the International Arctic Forum in Salekhard, spoke about the ecology of Greenpeace, staged on a platform of "Prirazlomnaja."

Expert: It is necessary to encourage participation in the election, rather than returning the column "against all"

09/25/2013 13:27

Political scientist and philosopher, Professor Oleg Matveychev HSE commented with "Pravda.Ru" Valentina Matviyenko offer to return to the ballot line "against all."

The British newspaper described the heroes and victims in Nairobi

25/09/2013 10:27

In Kenya - mourning for the victims of the terrorist attack in Nairobi. The newspaper The Independent said about the people who were at the mall, "Westgate" during capture.

World

Policy

Economy

English -- Russian

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

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:

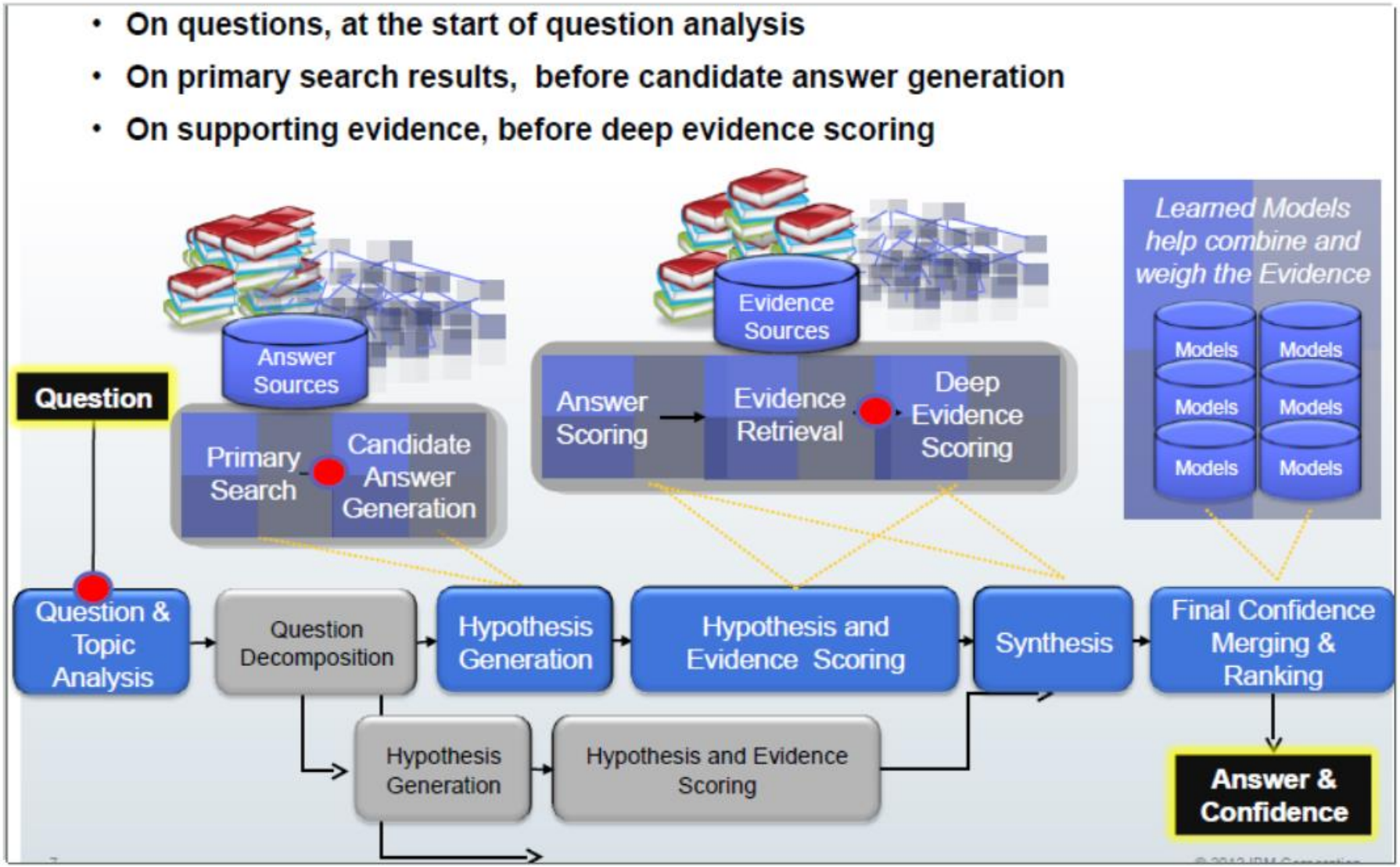
(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery

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

NLP in Watson

- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring



Weblog and Tweet Analytics

- Data-mining of Weblogs, discussion forums, message boards, user groups, tweets, other social media
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what's hot, what topics are people talking about right now).

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.

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
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)

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: you decide!

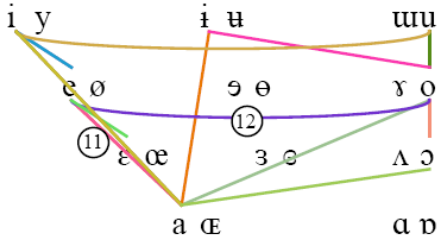
Two Generations of NLP

- 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

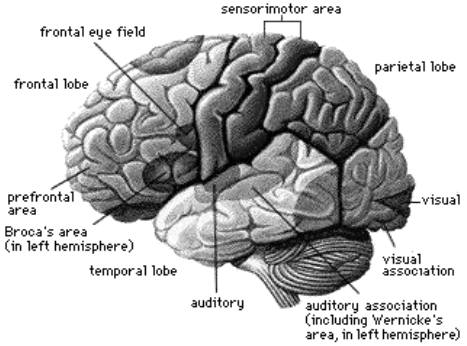
- Automatic, Trainable (Machine Learning) System [1985s–]
  - The tasks are modeled in a statistical way
  - More robust techniques based on rich annotations
  - Perform better than rules (Parsing 90% vs. 75% accuracy)

What is Nearby NLP?

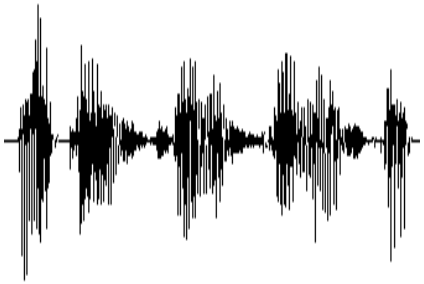
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- 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



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

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 the campaign so far

Caveat

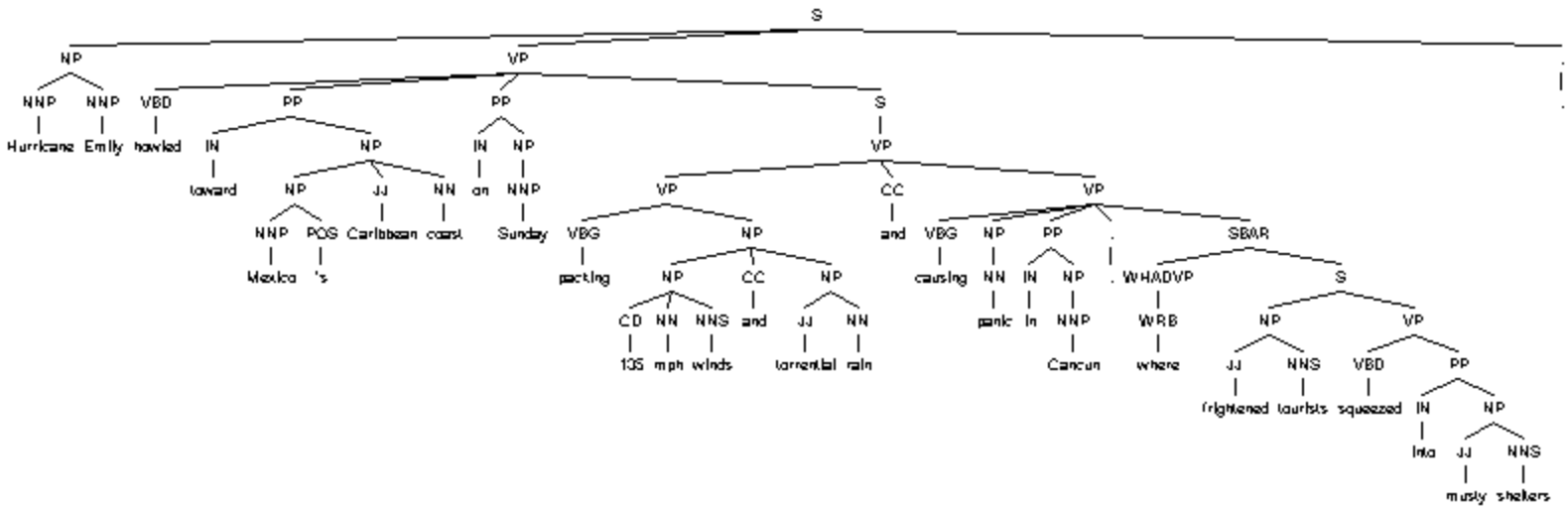
NLP has an AI aspect to it.

- 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

Problem: 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

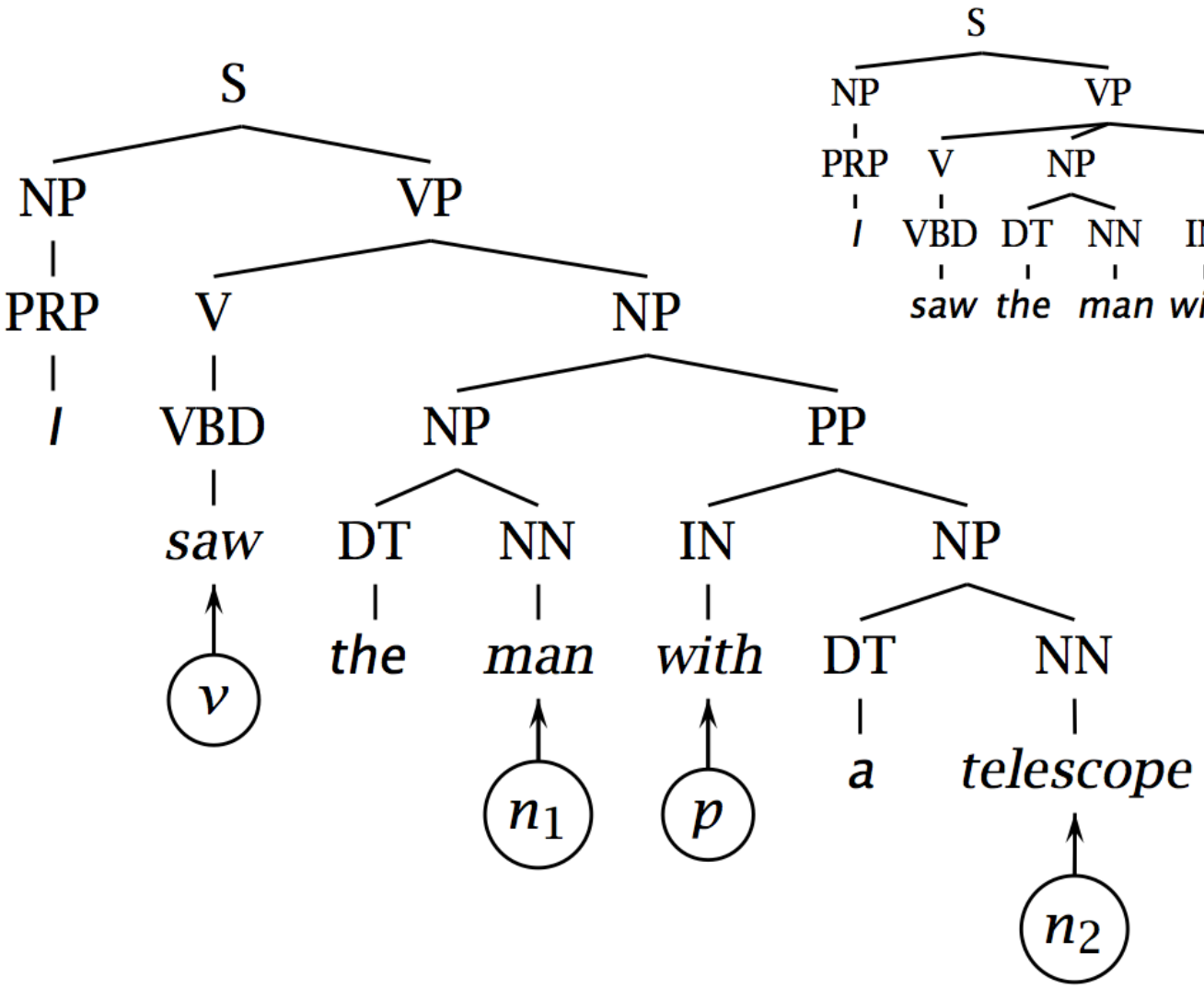
Syntactic Analysis



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

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. ☺

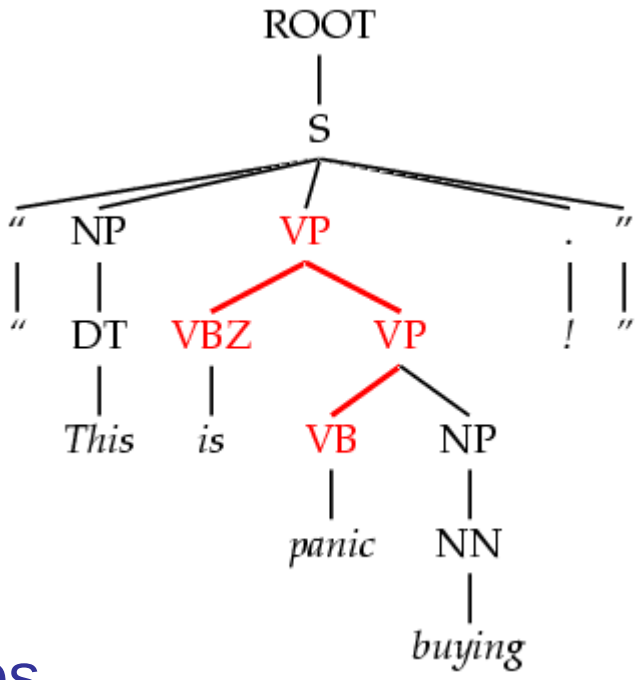
[Example from L. Lee]

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

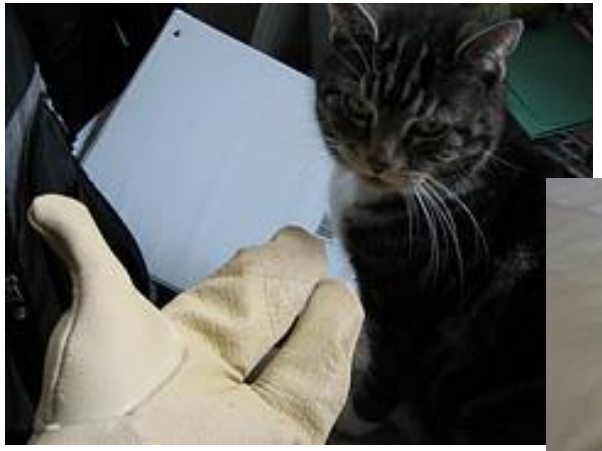
"This will panic buyers ! "



- 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

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

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”

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", "tenal 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

Imam Muhammad Ibn Saud Islamic University

Saudi Arabia

JFK Airport

Attend-school

Nationality

Nationality

Employee-of

Abdul Aziz Al-Omari
Alomari
Al-Umari

Abdulrahman al-Omari
Abdul Rahman Alomari
Alomari
Al-Umari

Member-of

Supervisor

City-of-residence

City-of-residence

Attend-school

Al-Qaeda

Mohamed Atta

Vero Beach, Florida

Florida Flight School

Even More Uncertainty: “Morphing” in Vision

Steganography, Codes, Ciphers and Jargons

- Steganography (concealed writing)
  - Wooden tablets covered with wax
  - Tattoon it on the scalp of a messenger
- **Codes**
- Boxer Seven | Seeks | Tiger5 | At | Red Coral

- **Ciphers**
- BEWARE | THE | IDES | OF | MARCH

- **Jargons**
- The supply drop | will take place | **at** | 0100 hours tomorrow

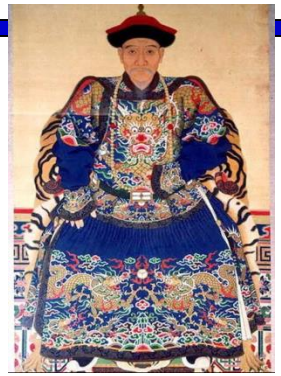
Morphing in Texts

To Ramzi bin al-Shibh

"The first semester commences in on September 11
World Trade Center and Pentagon and Capitol
This summer will surely be hot ...
19 hijackers for private education and four planes .
Regards to Bin Laden
Goodbye.

Abu Abdul Rahman

Morphs in Social Media



“Conquer West King”
(平西王)

=

“Bo Xilai”
(薄熙来)

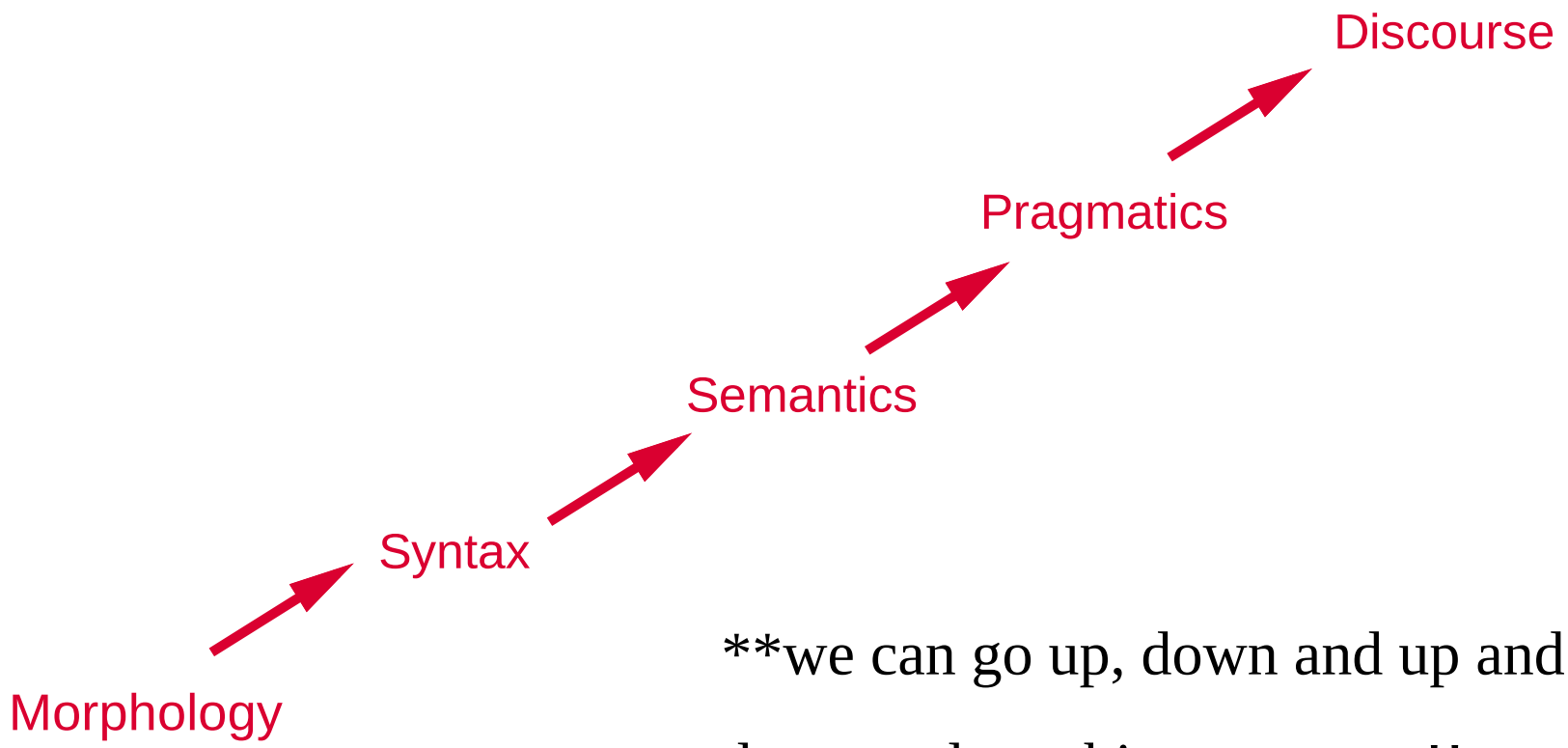
“Baby”
(宝宝)

=

“Wen Jiabao”
(温家宝)

Morph	Target	Motivation
Blind Man (瞎子)	Chen Guangcheng(陈光诚)	Sensitive
First Emperor (元祖)	Mao Zedong (毛泽东)	Vivid
Kimchi Country (泡菜国)	Korea (韩国)	Vivid
Rice Country (米国)	United States (美国)	Pronunciation
Kim Third Fat (金三胖)	Kim Jong-un (金正恩)	Negative
Miracle Brother (奇迹哥)	Wang Yongping (王勇平)	Irony

The Steps in NLP



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

**every step is equally complex

Dealing with Ambiguity

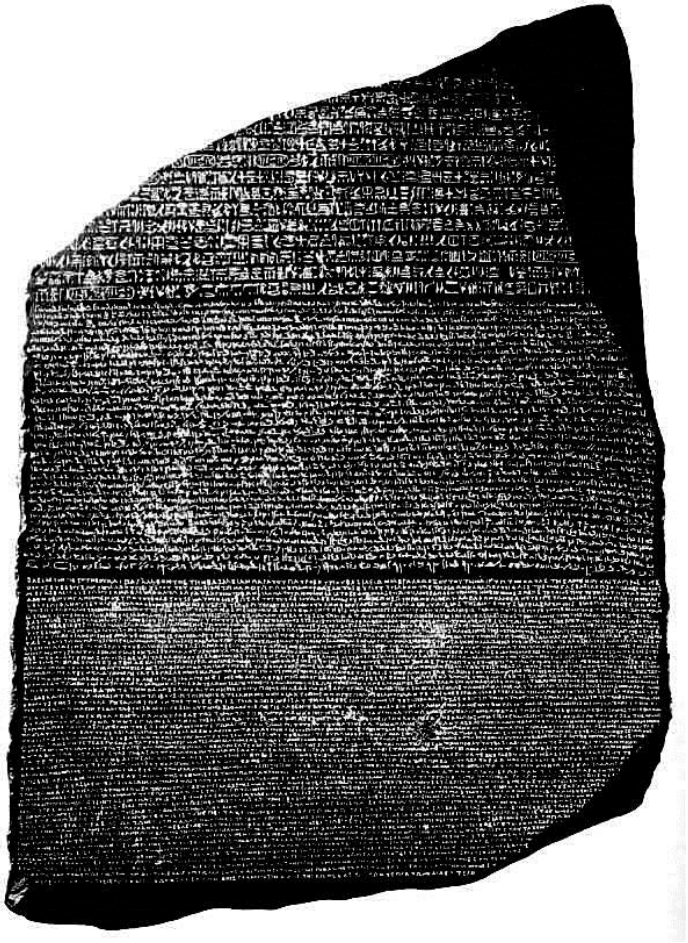
- Four possible approaches:
  1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide among choices at ambiguous levels.
  2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.

Dealing with Ambiguity

3. Probabilistic approaches based on making the most likely choices

4. Don't do anything, maybe it won't matter
   1. *We'll leave when the duck is ready to eat.*
   2. *The duck is ready to eat now.*
      - Does the "duck" ambiguity matter with respect to whether we can leave?

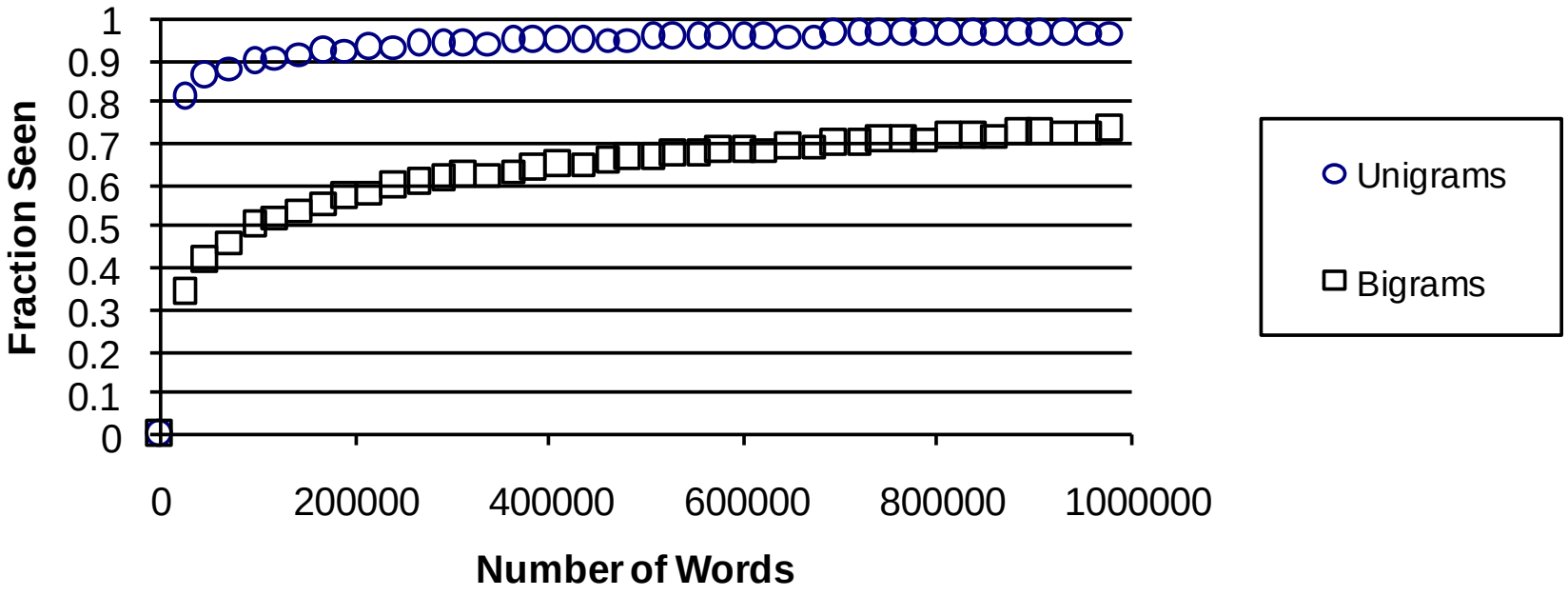
Corpora



- 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)



The Challenge

- 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

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,

ML Topics in the Course

- Naive Bayes,
- Hidden Markov Models,
- Expectation Maximization,
- Conditional Random Fields,
- MaxEnt Classifiers,
- Probabilistic Context Free Grammars,
- …

Disclaimer

- This course will be highly biased
  - won‘t focus much on linguistics
  - won‘t focus much on historical perspectives

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