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:
  - Vishal Saley
  - Pranjal Aggarwal

Logistics

- Course Website:
  www.cse.iitd.ac.in/~mausam/courses/col772/spring2024
- Join class discussion group on Piazza (access code col772)
  https://piazza.com/iit_delhi/winter2024/col772
- 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:
  - 45% assignments
  - 20% midsem
  - 25% major
  - 10% quiz
  - Extra credit: constructive participation in class and discussion group

Assignments and Project

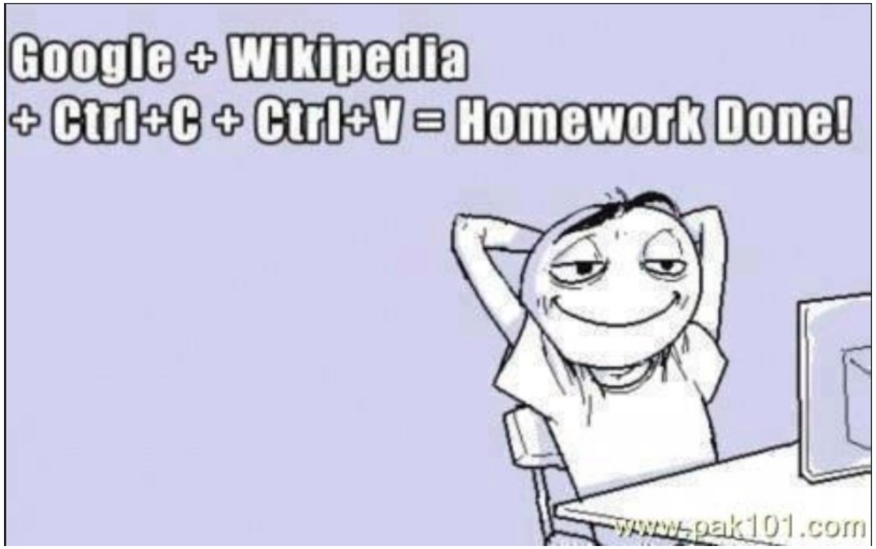
- 3-4 programming assignments
  - assignments done **individually**!
  - late policy: penalty of 10% maximum grade every day for a week

- Project
  - No project this time

- Request HPC approval after A1!
  - End date: 30th May 2024
  - 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
    - Deep Learning
  - 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
  - Deep Learning ☺

Timings

- Tue/Fri 2-3:30
- Office hours
  - By appointment
- Audit criteria
  - B-

Goals of this course

- Learn the issues and techniques of modern NLP
- Build realistic NLP tools
- See the latest trends in field of NLP → deep learning
- See where the holes in the field still are!

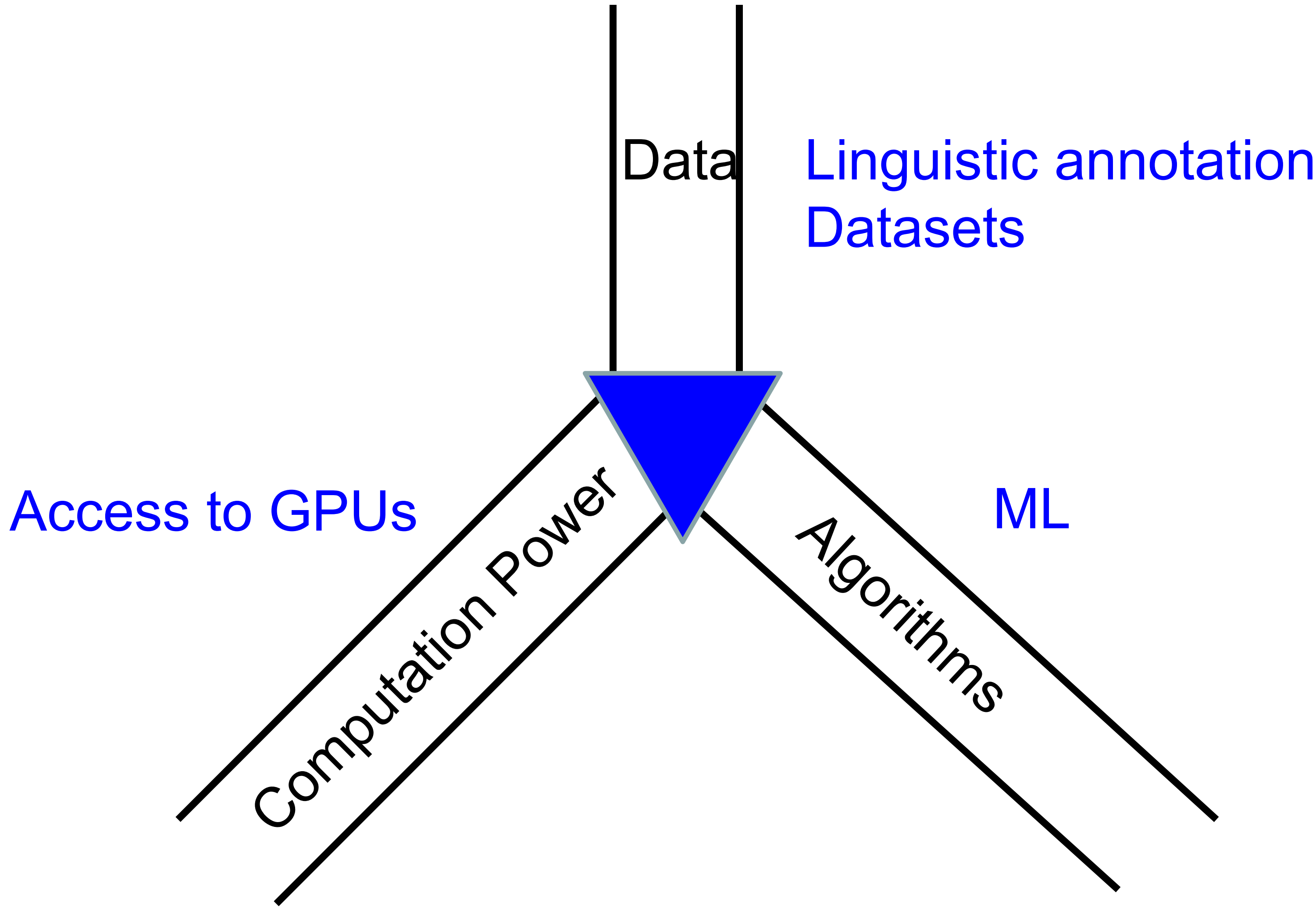
- Computer Engineer
  - very relevant field in the modern age

- 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!

Three Way Crossing in NLP



MOTIVATION

The Dream

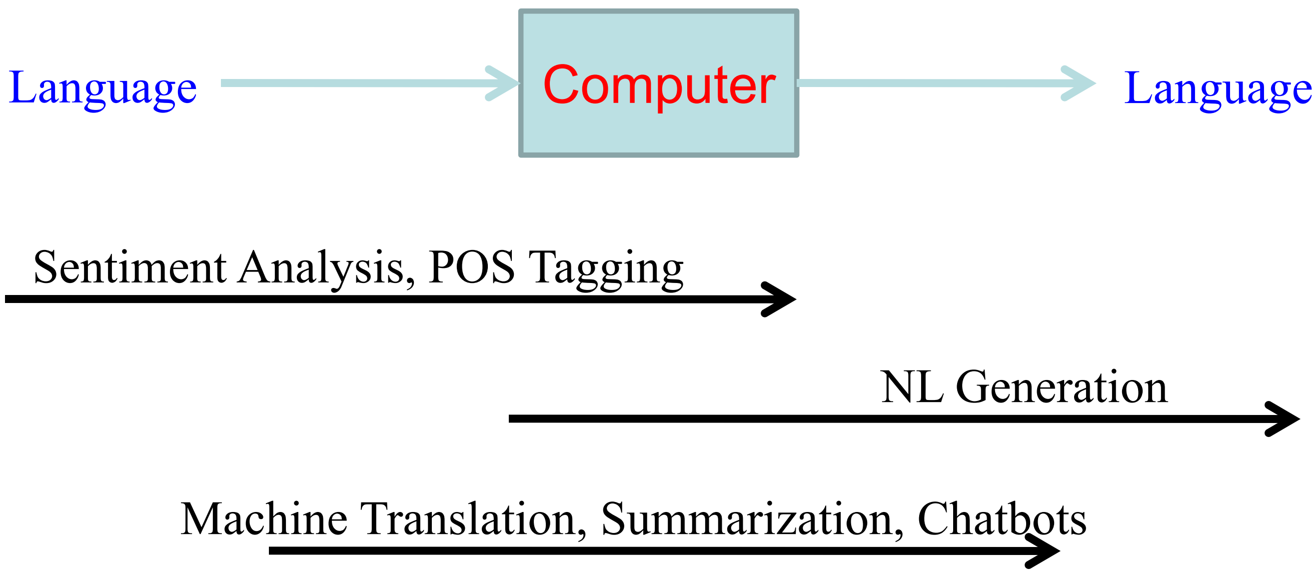
- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle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

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

What is NLP?

Computer processing of human language



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!

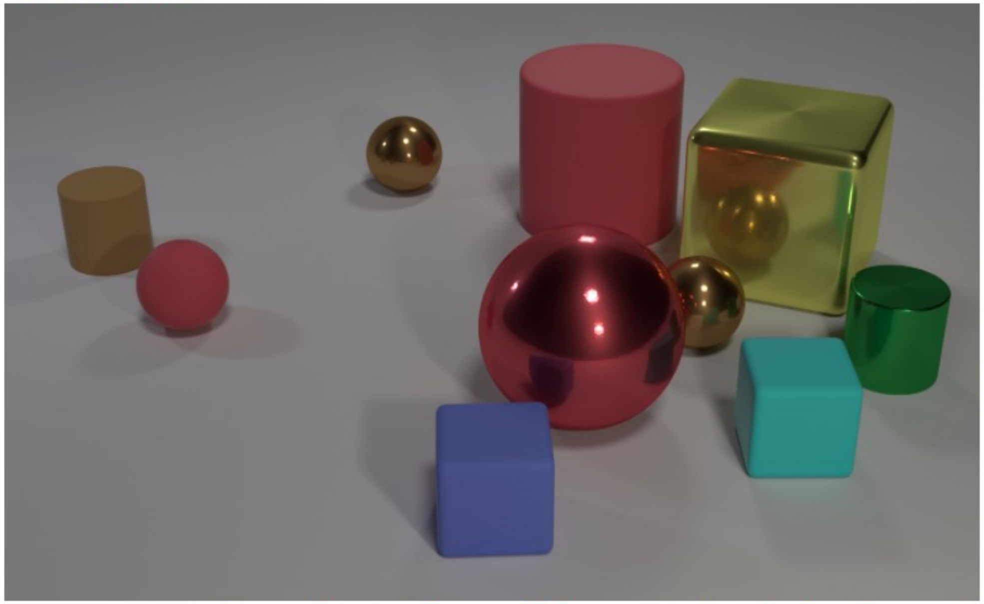
- 1960s and 1970s: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
    - “Little John was looking at his toy box. Finally, he found it. The box was in the pen. John was very happy.”
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)

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

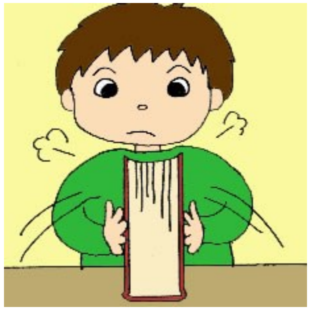
NLP: machine learning and empiricism

“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
- 2020s: Pre-trained language models, Foundation models

Three/Four? Generations of NLP

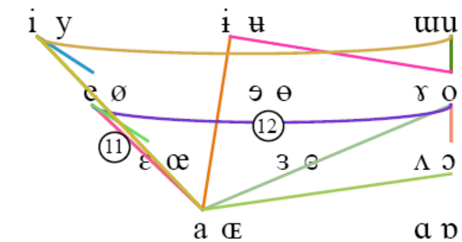
- Hand-crafted Systems – Knowledge Engineering [1950s–]
  - Rules written by hand; adjusted by error analysis
  - Require experts who understand both the systems and domain
  - Iterative guess-test-tweak-repeat cycle



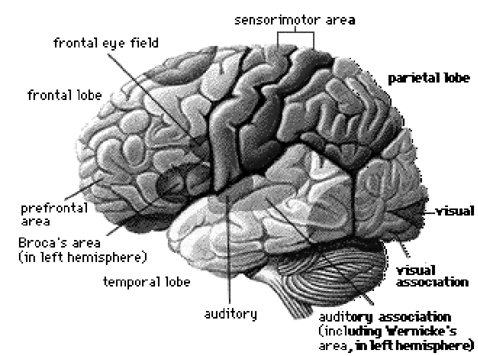
- Automatic, Trainable (Machine Learning) Systems with engineered features [1985s–2012]
  - The tasks use statistical models with hand-coded features
  - More robust techniques based on rich annotations
  - Perform better than rules (Parsing 90% vs. 75% accuracy)
- Automatic, Trainable Neural architectures with no/limited engineered features [2012--]
  - The tasks use deep learning
  - Perform much better than hand-coded features

What is Nearby NLP?

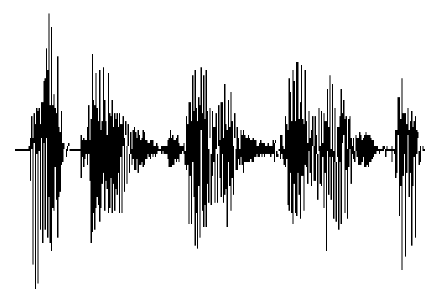
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- 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 Biden’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

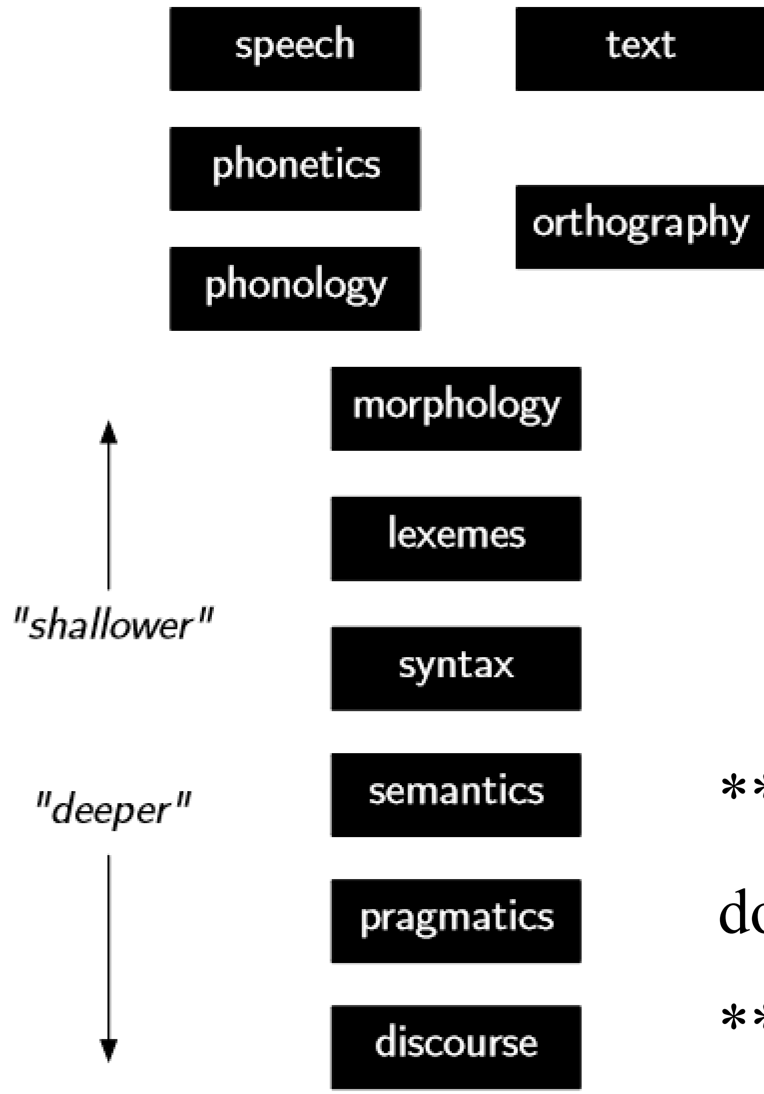
NLP is AI Complete

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



**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, fan-frickin'-tastic,

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

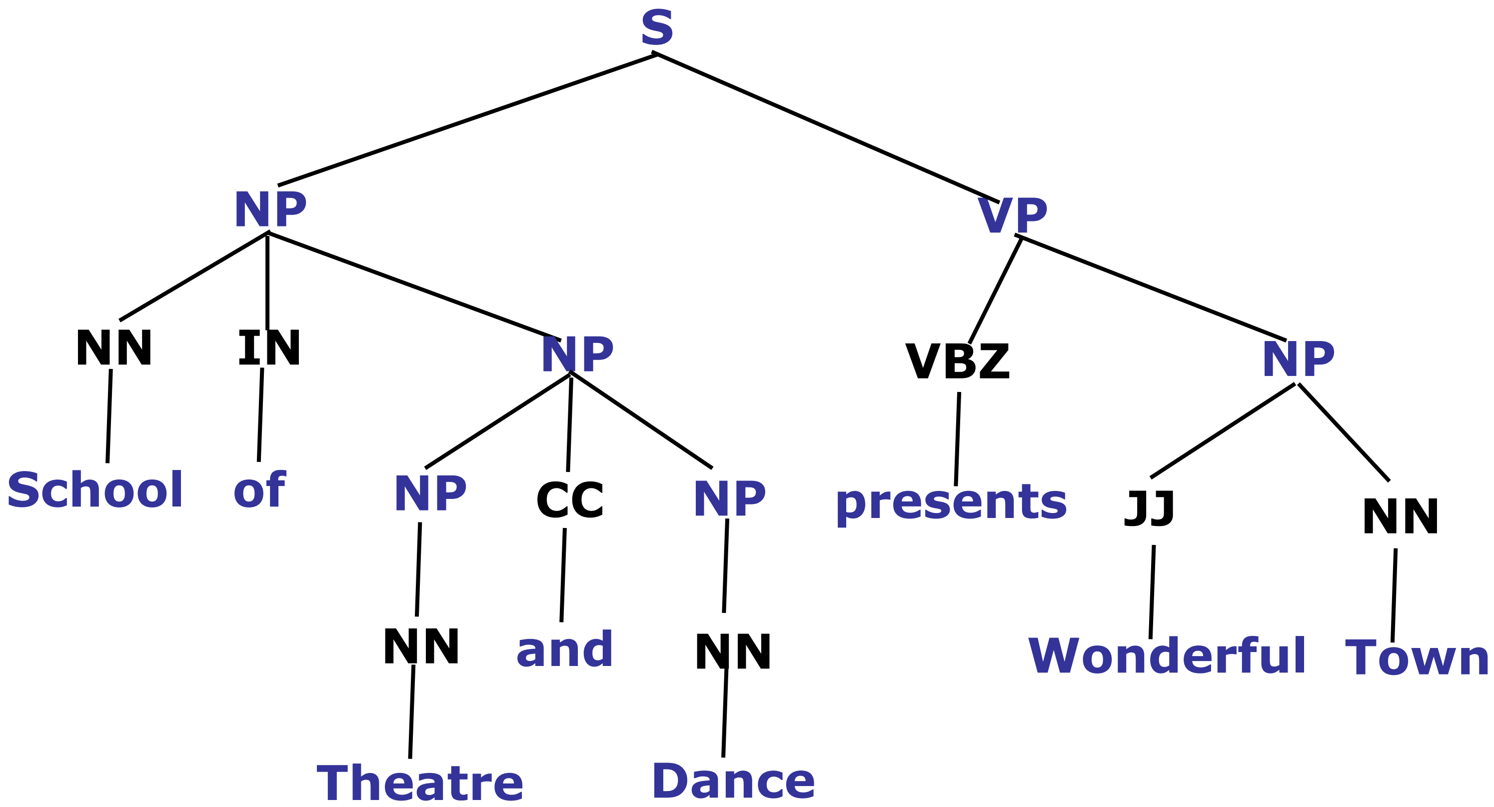
ikr smh he asked fir yo last name

so he can add u on fb lololol

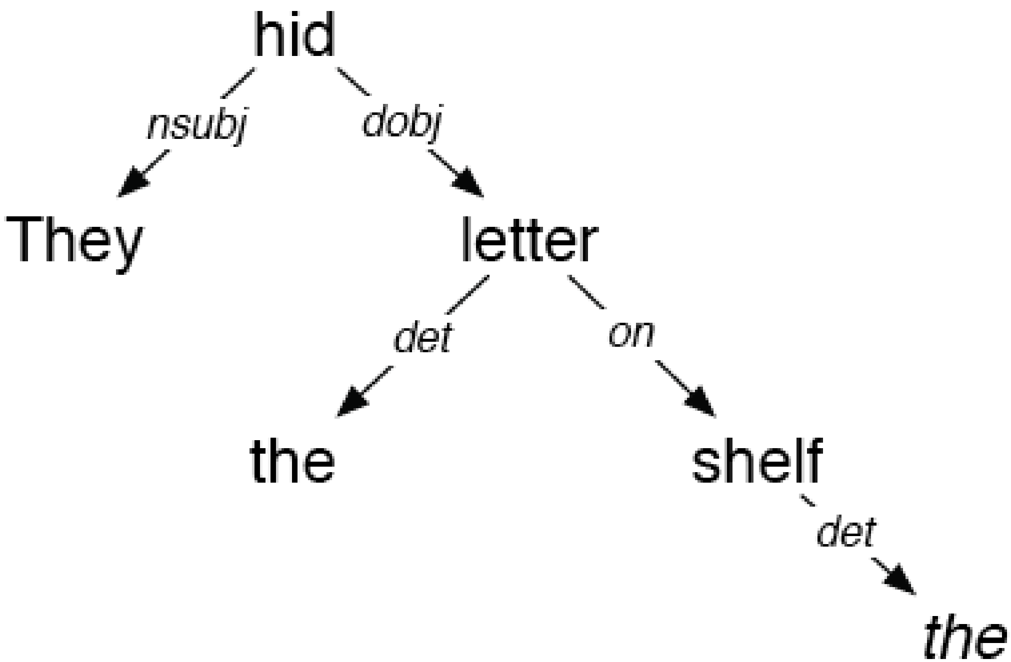
Morphology+Syntax

A ship-shipping ship, shipping shipping-ships.

Constituency Parsing



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

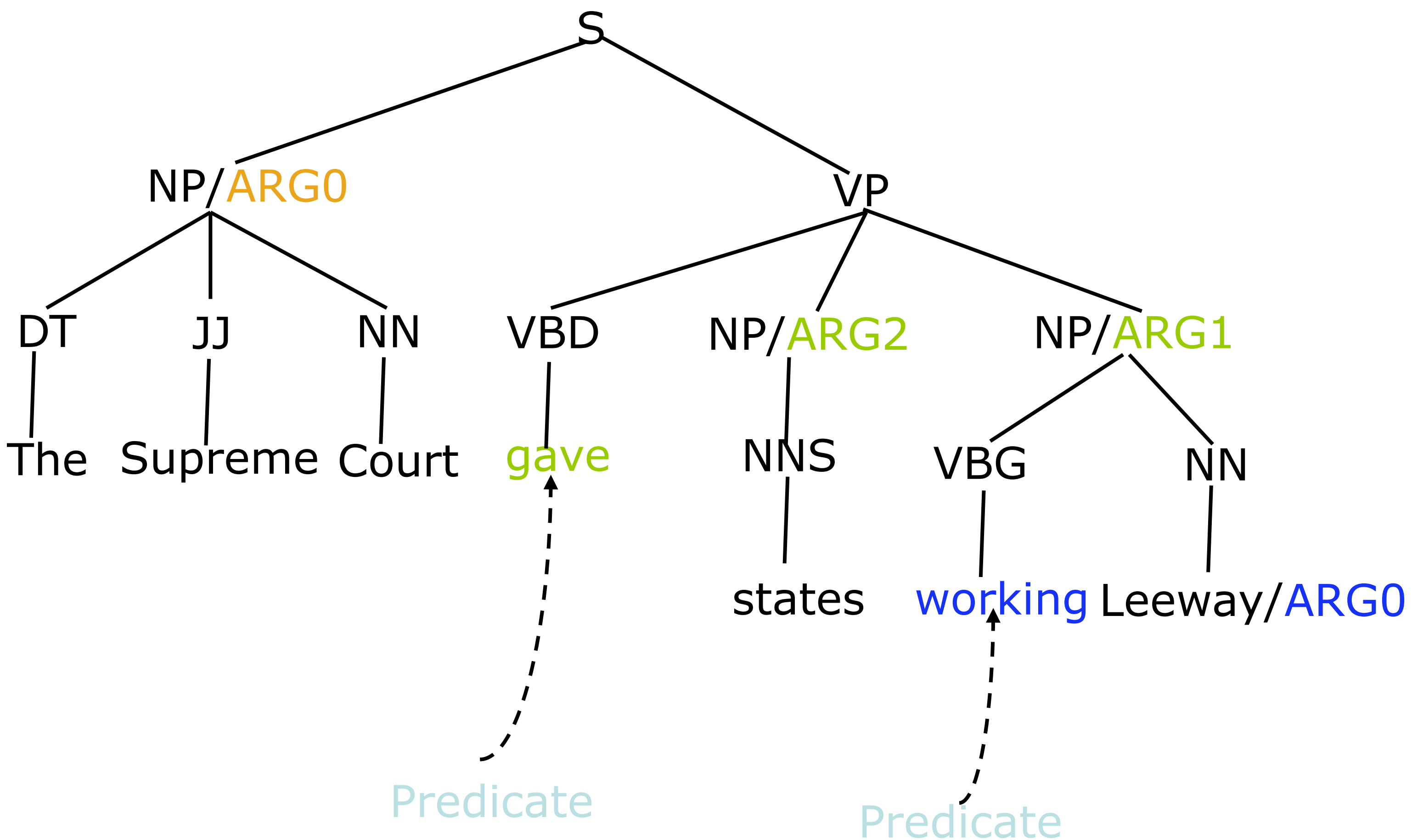
Semantics

- 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.

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!

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

Knowledge Base Linking (Wikification)

Query = "James Parsons"

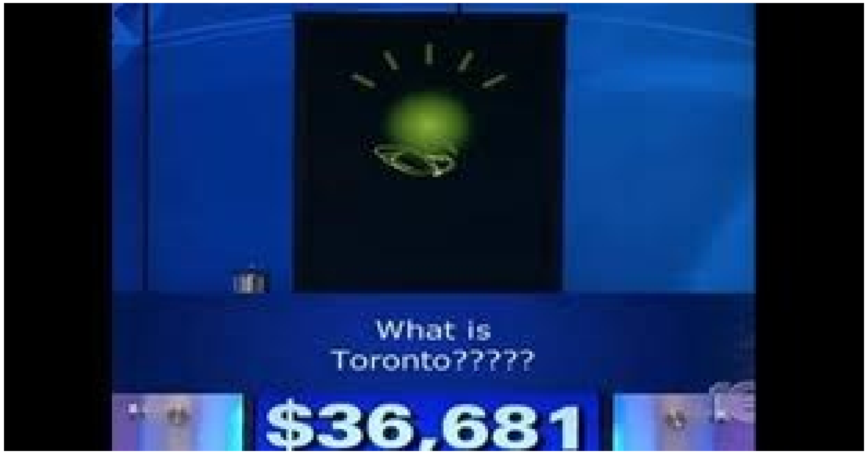
Google Knowledge Graph (2012)

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’’

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



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.

Summarization

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

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

- Very context-dependent!

- An example of analysis with generation

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"

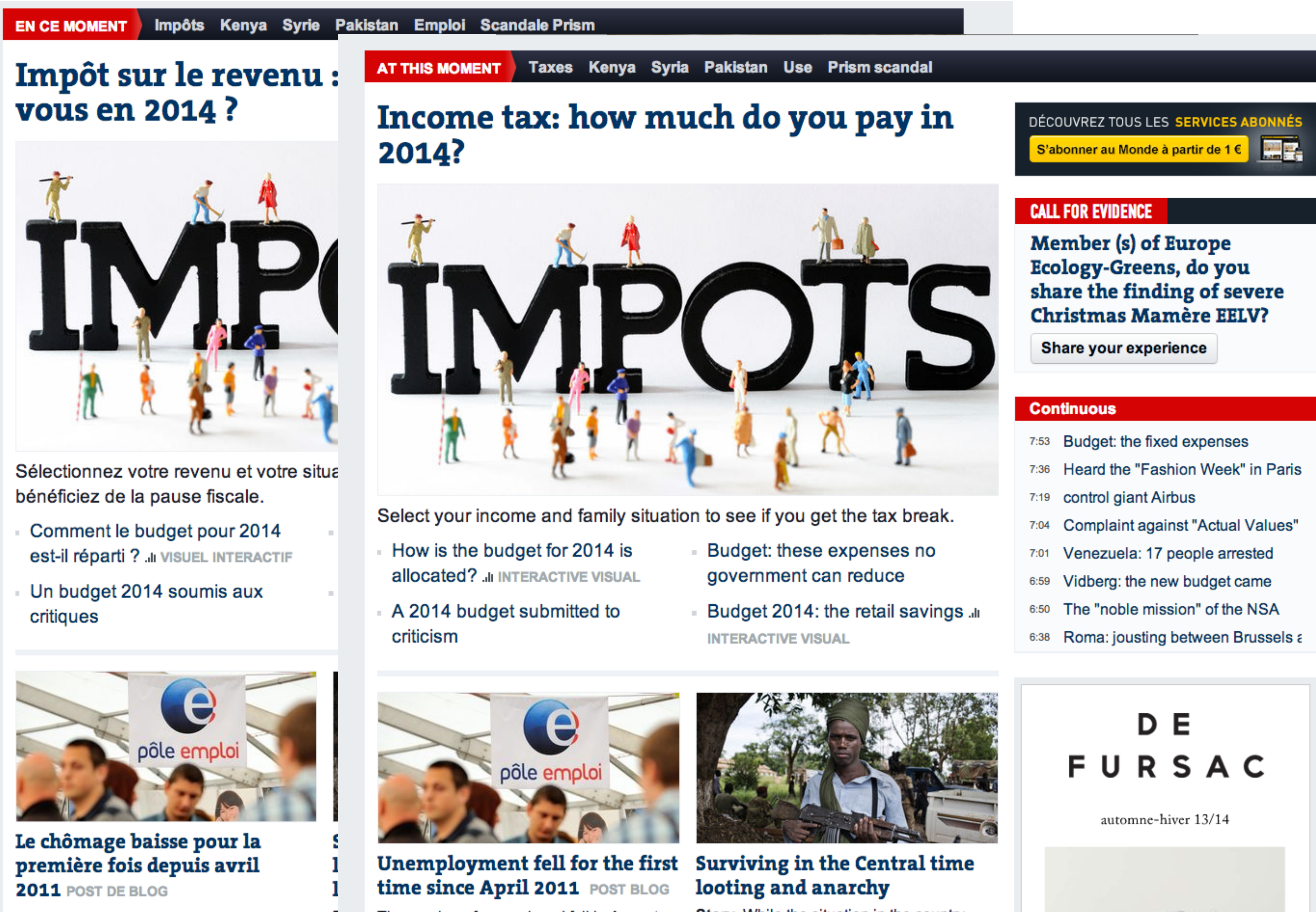
Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

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



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 Indepen "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.

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.

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

■ 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)***

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/Common-Sense

"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

ConvAI/ChatBots/Dialog Systems

Write a tweet about the importance of learning AI and ML learning skills and embracing the new possibilities of this sector.

"Embracing #AI and #MachineLearning is more than a trend; it's a pathway to the future. 🚀 Acquiring these skills opens doors to innovation, creativity, and endless possibilities in tech. Don't miss out on shaping tomorrow. Invest in learning today! 💻 🌐 #FutureTech #LearnAI"

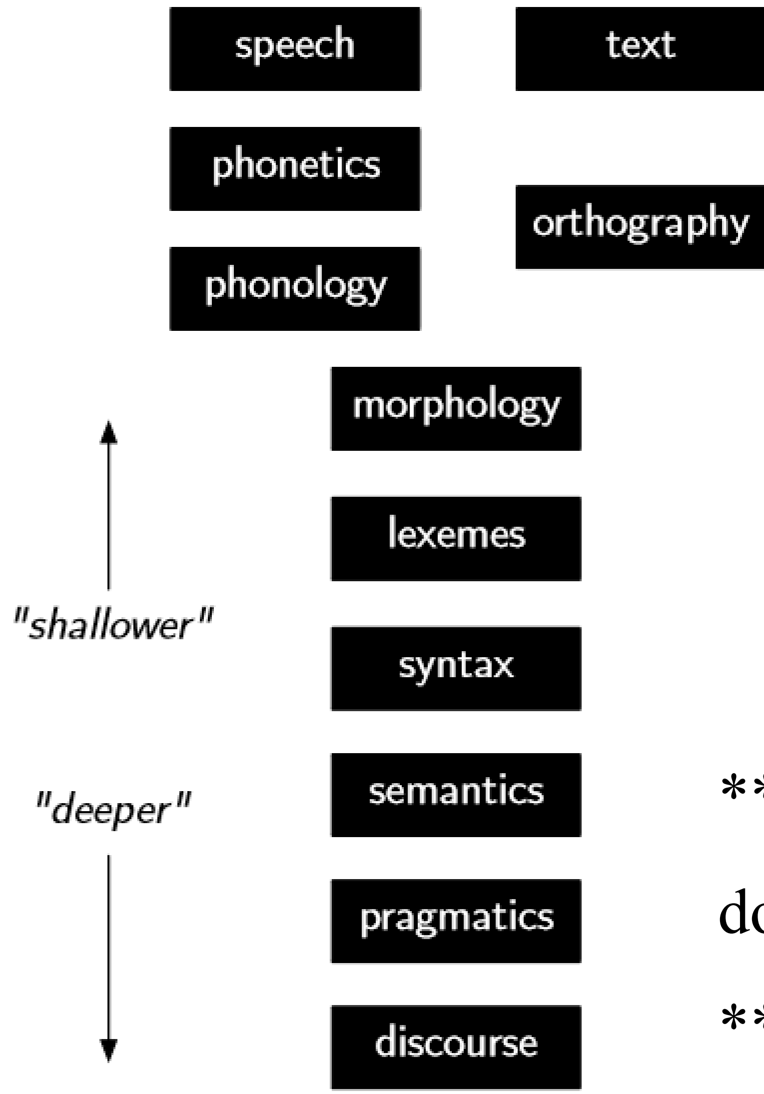
Write a tweet about the importance of learning AI and ML learning skills and embracing the new possibilities of this sector.

My goal is to entice readers to subscriber to my newsletter and download my guides for using automation to become more productive.

"Unlock the power of #AI and #MachineLearning! 🤖 ✨ Enhance productivity and discover new horizons with our expert guides. Subscribe to our newsletter and download FREE tools to automate your life. Don't just keep up with the future – lead it! 🔗 [Your Link Here] #AutomationMagic"

Regenerate

Levels of Linguistic Knowledge



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

**every step is equally complex

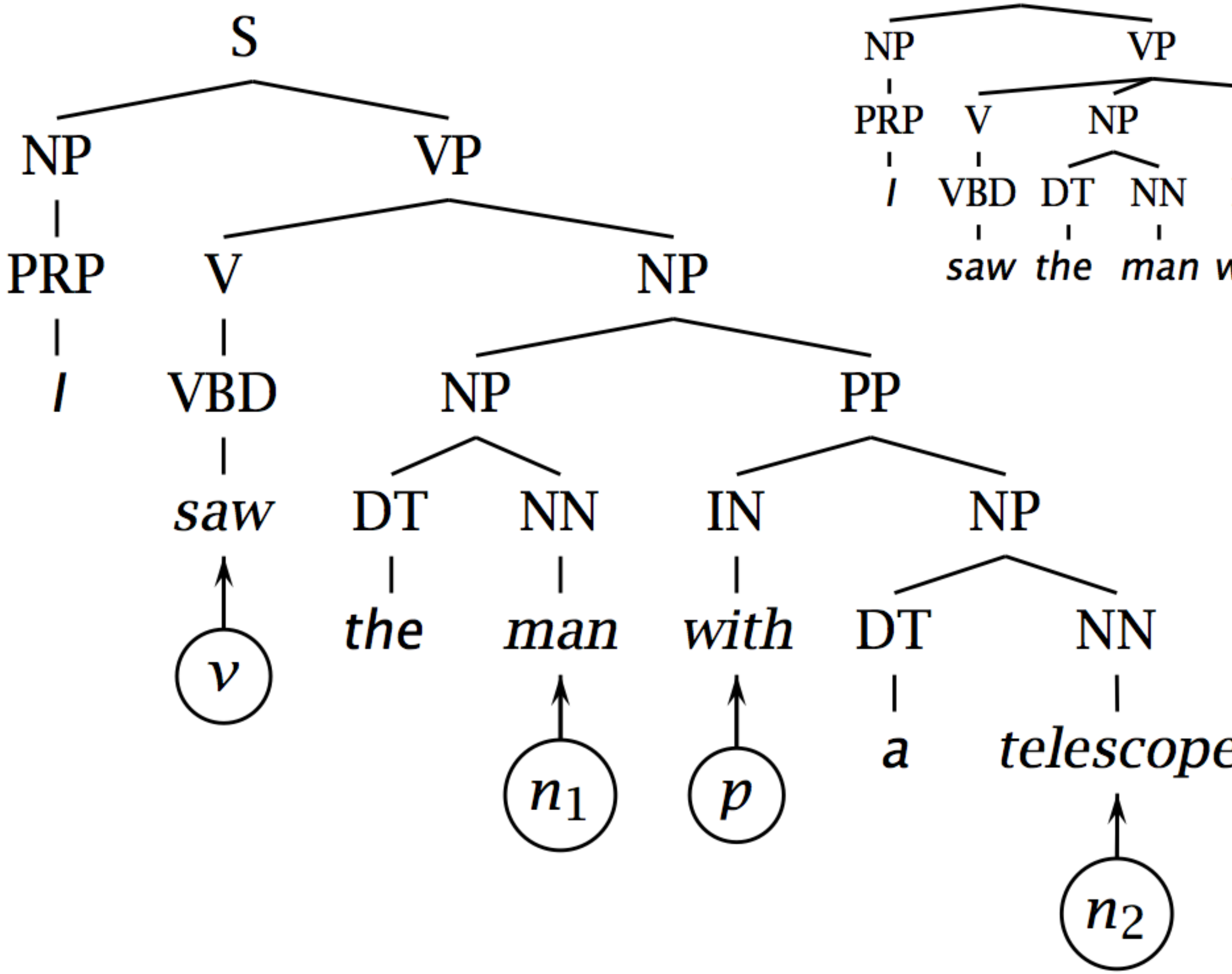
Why is NLP hard?

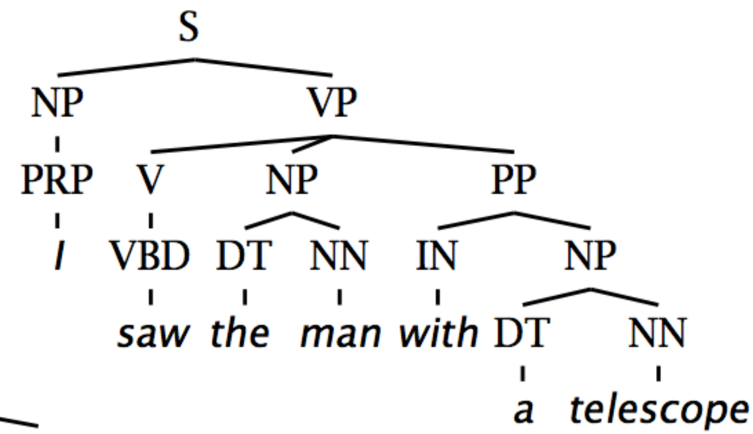
- Ambiguity
- Ambiguity
- Ambiguity
- Sparsity

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

PP Attachment Ambiguity





Coordination Ambiguity

Modifier	% of Total	Example from Surveys (modifiers underlined)
Noun	46.4	(It) targeted the project and election managers
Adjective	23.2	.... define architectural components and connectors
Prep	15.9	Facilitate the scheduling and performing of works
Verb	5.8	capacity and network resources required
Adverb	4.4	(It) might be automatically rejected or flagged
Rel. Clause	2.2	Assumptions and dependencies that are of importance
Number	0.7	zero mean values and standard deviation
Other	1.4	increased by the lack of funding and local resources

https://www.kilgarriff.co.uk/Publications/2006-ChantreeWillisKilgDeRoeck-RANLPbook.pdf

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. ☺

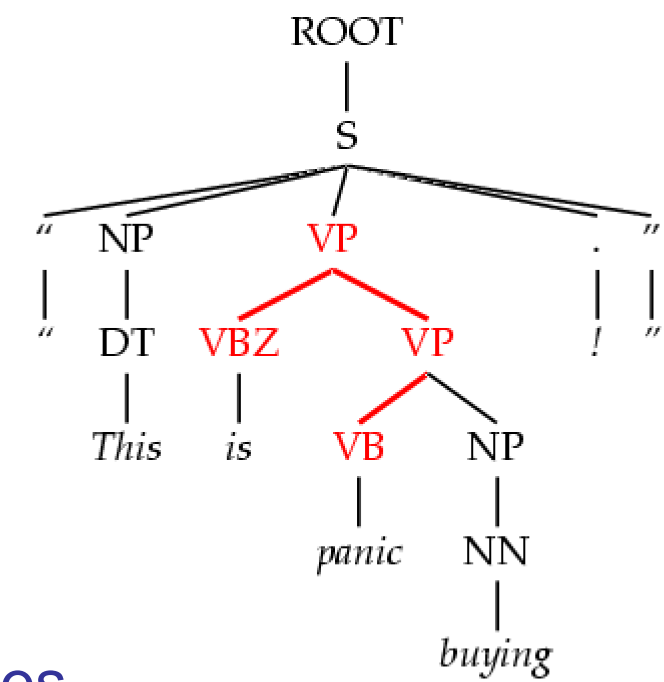
[Example from L. Lee]

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

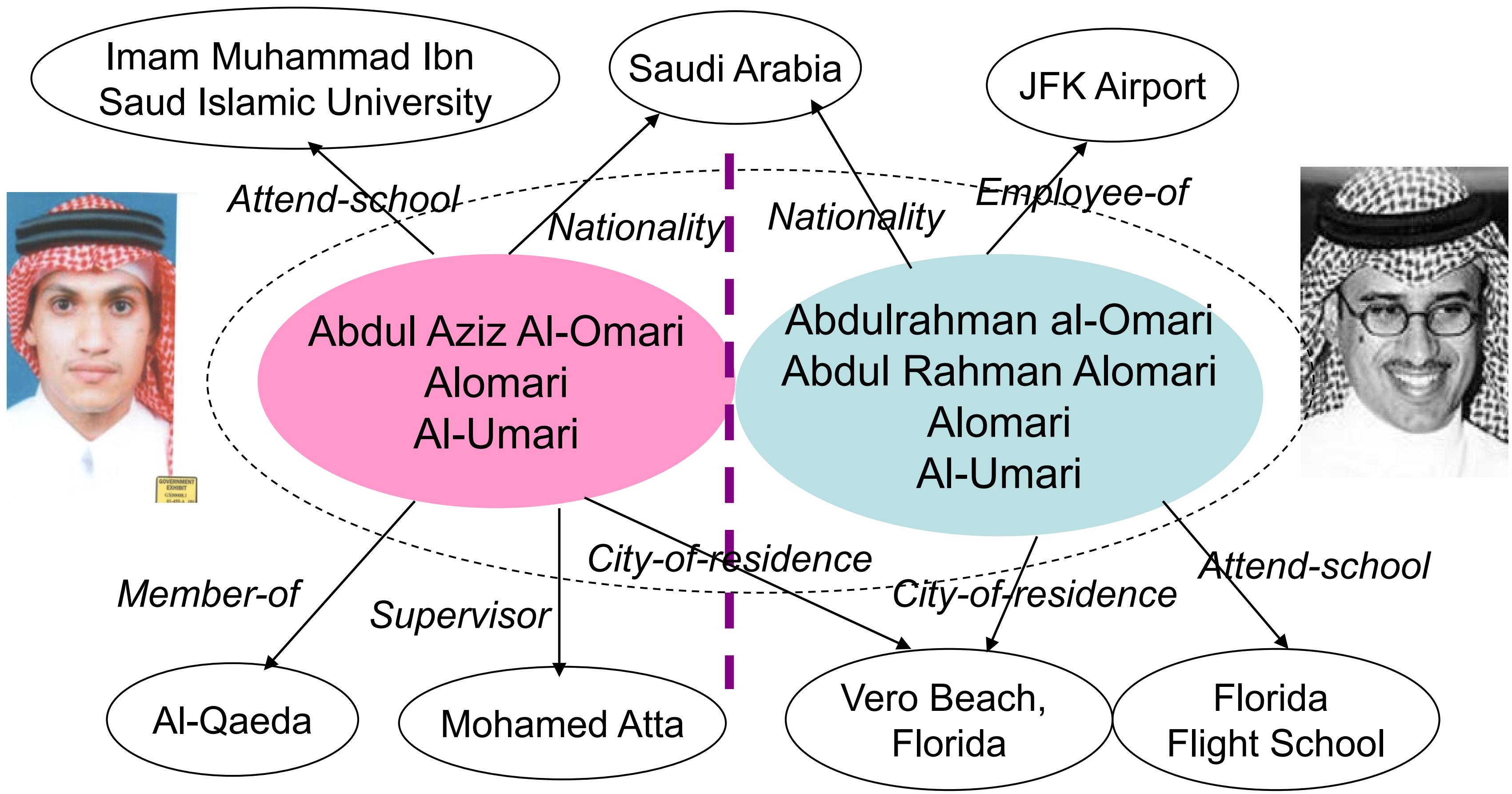
Abbrevations out of context

- 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”, “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”, …

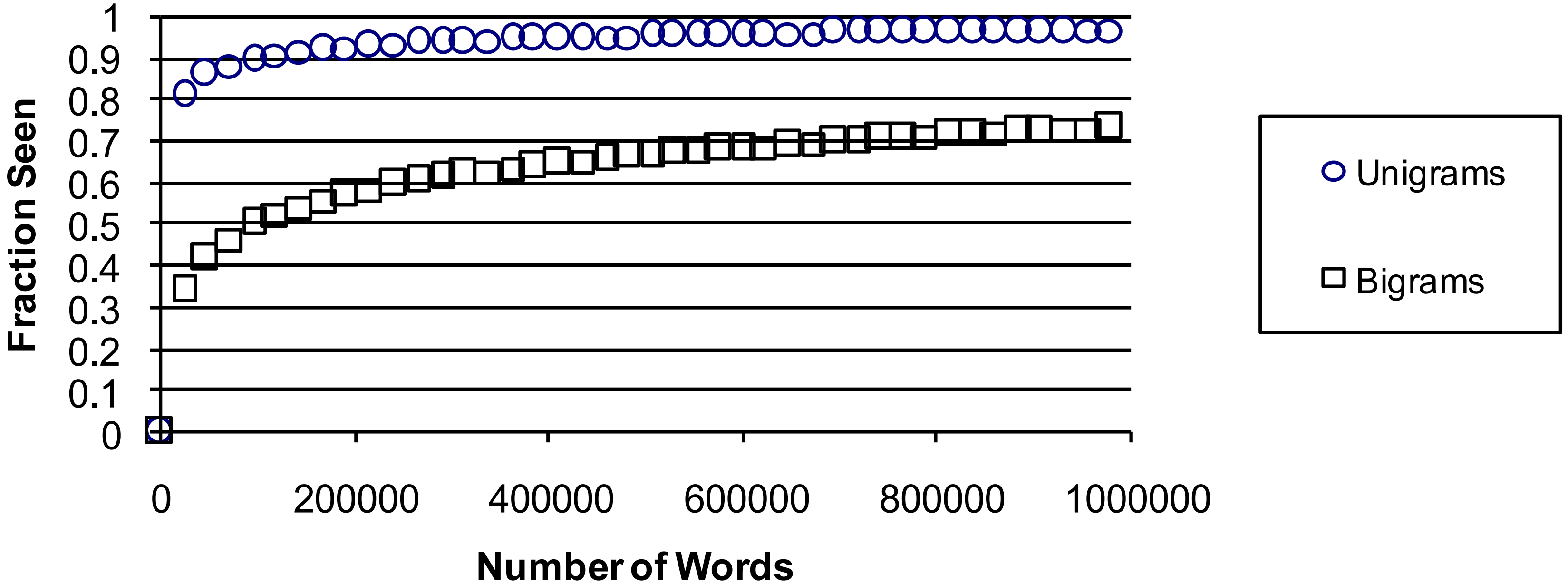
- Military Domain
  - *“GA ADT 1, USDA, USAID, Turkish PRT, and the DAIL staff met to create the Wardak Agricultural Steering Committee. “*
  - “*DST*” = “*District Stability Team*” or “*District Sanitation Technician*”?

Uncertainty: Ambiguity Example



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, MaxEnt Classifiers
- Expectation Maximization,
- Conditional Random Fields,
- Probabilistic Context Free Grammars,

- CNNs, Recurrent NNs, Transformers
- Sequence to Sequence models
- Neural Language Models
- Pre-trained Language Models
- Advanced Ideas

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 … ☺