CSE P 517
Natural Language Processing
Winter 2021

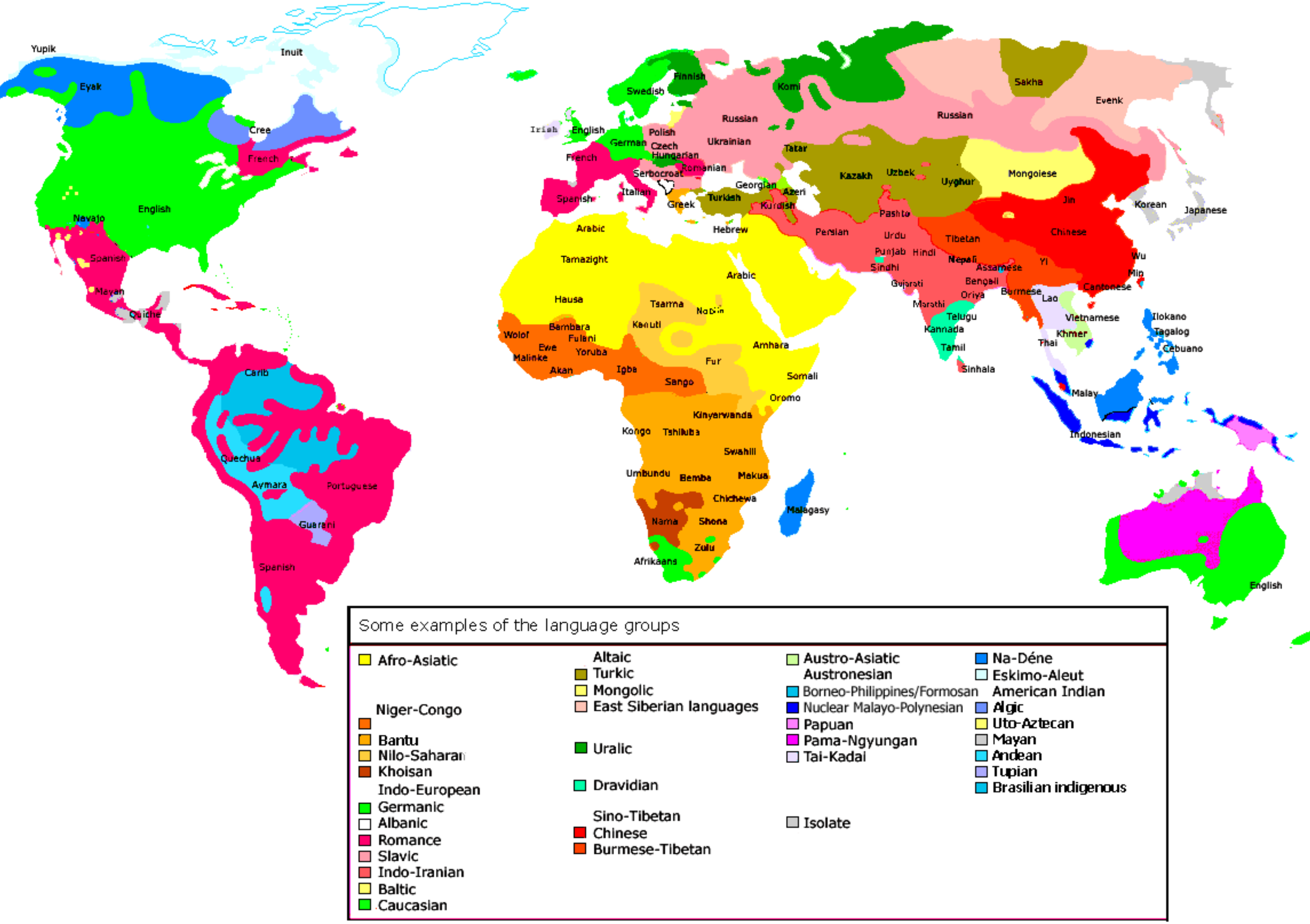
Introduction

Yejin Choi

Slides adapted from Dan Klein, Luke Zettlemoyer

What is NLP?

- Fundamental goal: *deep* understand 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, sentiment analysis, question answering…
  - Unknown: human-level comprehension (is this just NLP?)


Yupik
Inuit
Eyak
Cree
French
English
Navajo
Spanish
Mayan
Quiche
Carib
Quechua
Aymara
Portuguese
Guarani
Spanish
Irish
English
Swedish
Finnish
Komi
Russian
Polish
German
Czech
Hungarian
Ukrainian
French
Romanian
Serbocroat
Tatar
Italian
Spanish
Georgian
Azeri
Greek
Turkish
Kurdish
Kazakh
Uzbek
Uyghur
Mongolese
Sakha
Evenk
Russian
Jin
Korean
Japanese
Arabic
Hebrew
Persian
Pashto
Urdu
Tibetan
Chinese
Tamazight
Punjab
Hindi
Nepali
Sindhi
Assamese
Yi
Wu
Min
Arabic
Gujarati
Bengali
Cantonese
Oriya
Burmese
Lao
Hausa
Tsarma
Nobiin
Marathi
Telugu
Vietnamese
Ilokano
Tagalog
Kanuri
Kannada
Khmer
Thai
Cebuano
Bambara
Wolof
Fulani
Ewe
Yoruba
Amhara
Tamil
Malinke
Fur
Sinhala
Akan
Igbo
Sango
Somali
Malay
Oromo
Kinyarwanda
Kongo
Tshiluba
Indonesian
Swahili
Umbundu
Bemba
Makua
Chichewa
Malagasy
Nama
Shona
Zulu
Afrikaans
English
Some examples of the language groups
Afro-Asiatic
Niger-Congo
Bantu
Nilo-Saharan
Khoisan
Indo-European
Germanic
Albanic
Romance
Slavic
Indo-Iranian
Baltic
Caucasian
Altaic
Turkic
Mongolic
East Siberian languages
Uralic
Dravidian
Sino-Tibetan
Chinese
Burmese-Tibetan
Austro-Asiatic
Austronesian
Borneo-Philippines/Formosan
Nuclear Malayo-Polynesian
Papuan
Pama-Ngyungan
Tai-Kadai
Isolate
Na-Déne
Eskimo-Aleut
American Indian
Algic
Uto-Aztecan
Mayan
Andean
Tupian
Brasilian indigenous

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

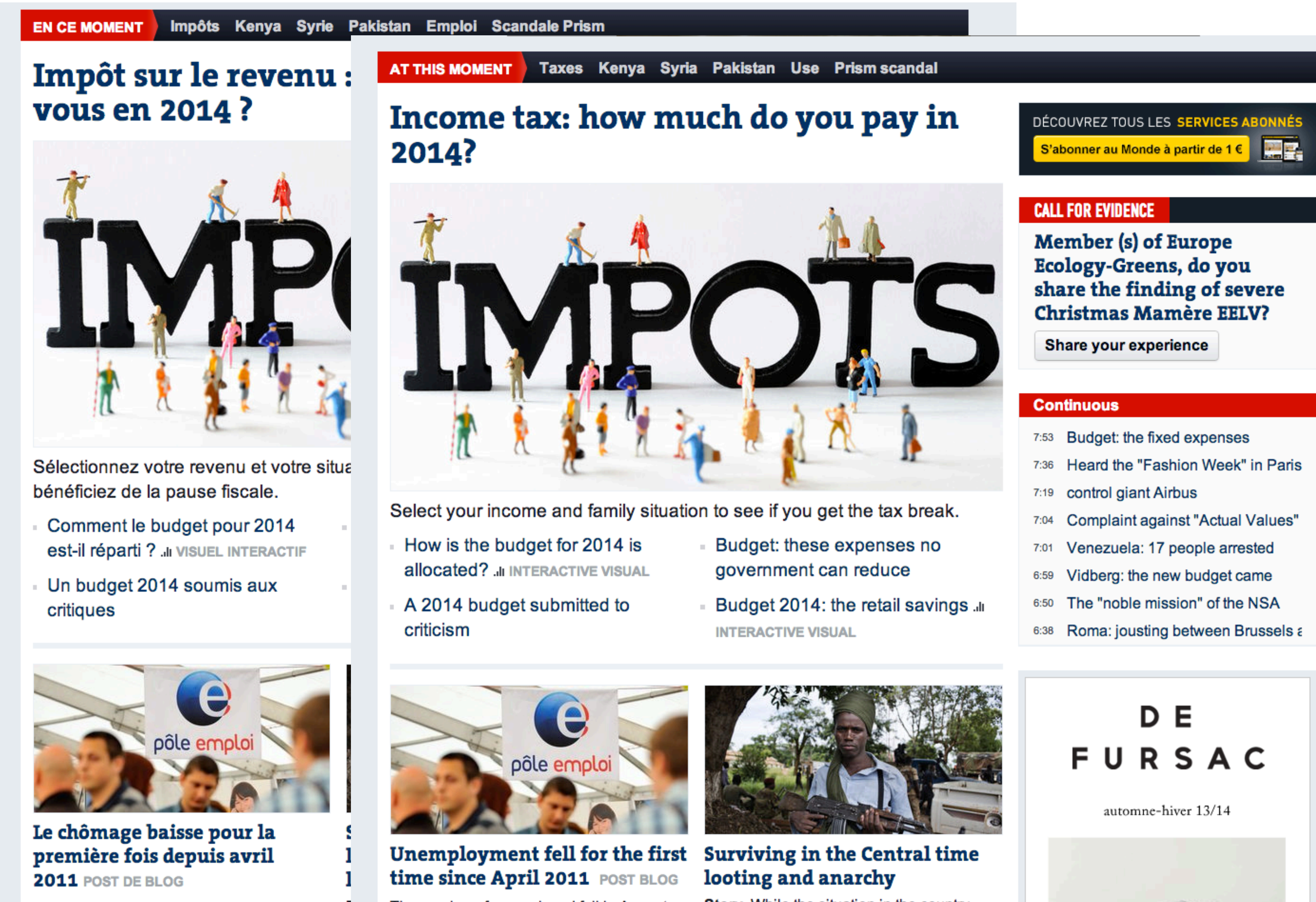
"It is impossible for journalists to enter Tibetan areas"

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."

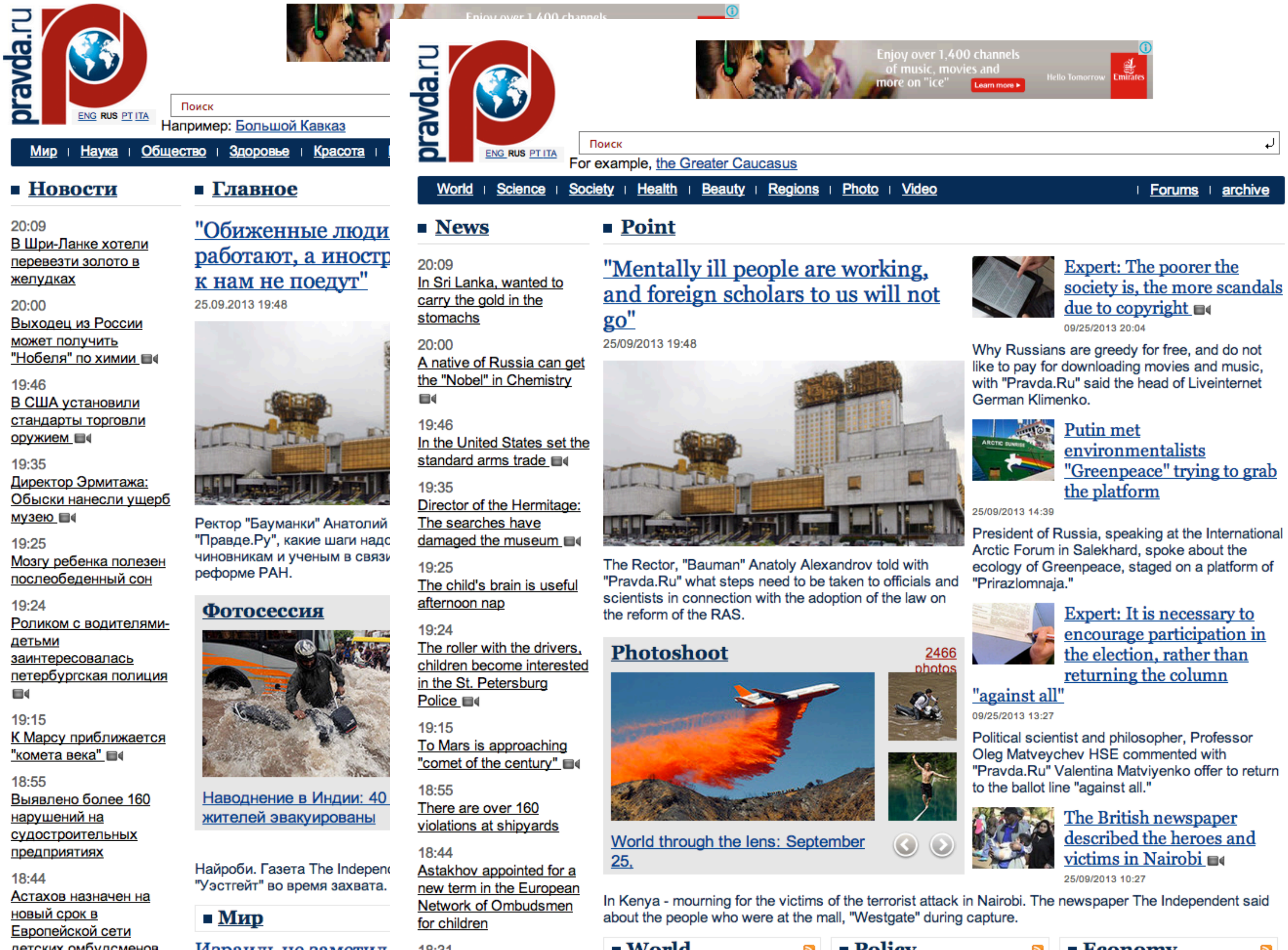
Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959
Video Anniversary of the Tibetan rebellion: China on guard

- 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 (second half of this class) vs fidelity (later)

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.

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

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.

Knowledge Graph: *"things not strings"*

Information Extraction

- From 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	started
Russell T. Lewis	New York Times newspaper	executive vice president	ended
Lance R. Primis	New York Times Co.	president and CEO	started

Information Extraction

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	started
Russell T. Lewis	New York Times newspaper	executive vice president	ended
Lance R. Primis	New York Times Co.	president and CEO	started

Sub-problems:

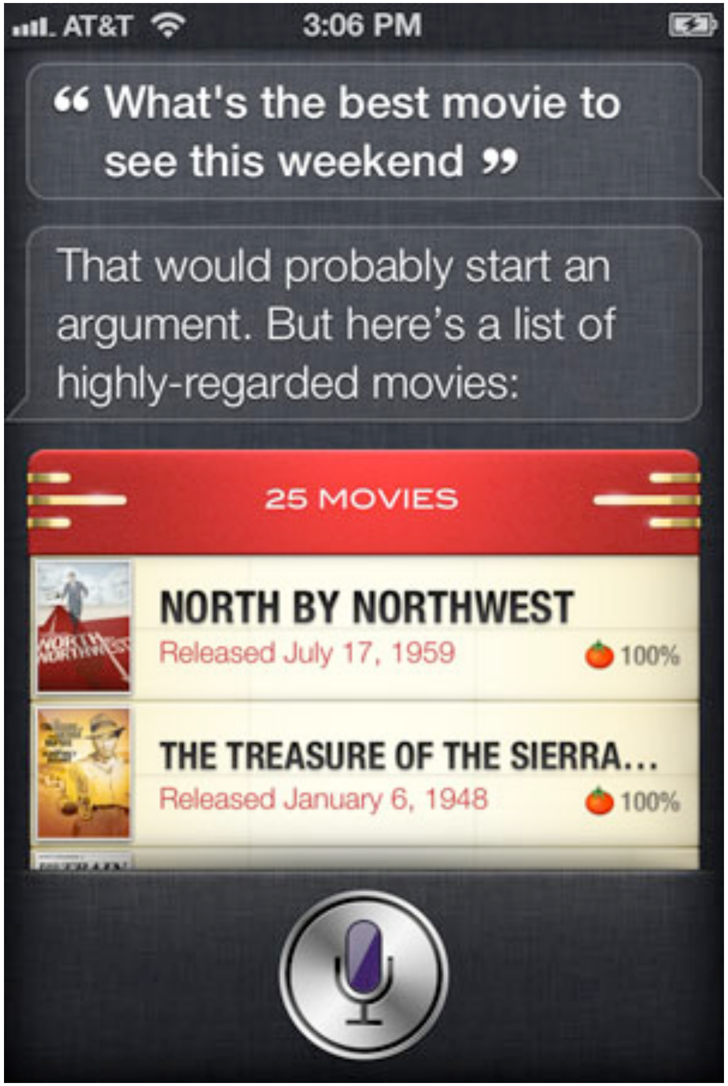
1) Named entity recognition: finding named entities X and their types T(X)
   persons: "Russell T. Lewis", "Lance R. Primis"
   companies: "New York Times Newspaper", "New York Times Co."
2) Relation extraction: the relation R(X,Y) between named entities X, Y
   Works_for(Russell T. Lewis, New York Times Newspaper)
3) Coreference resolution: which text spans refer to the same named entity?
   {Russell T.Lewis, He, He} are an equivalence set.

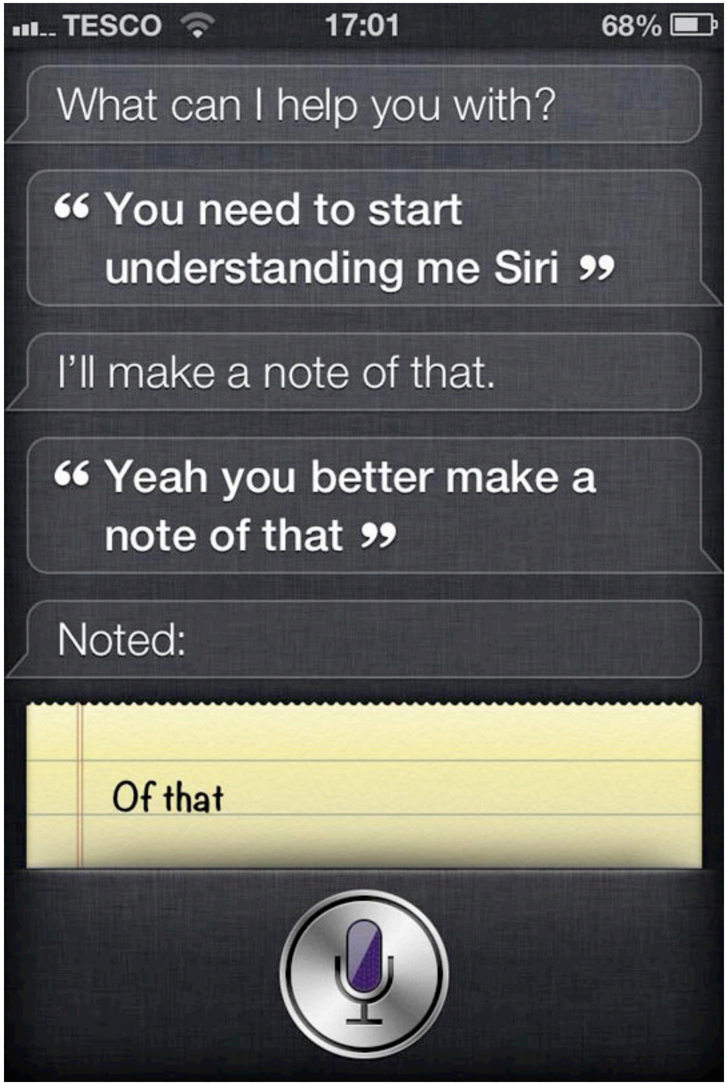
- Is this easy or hard?
- Easier if the model exploits the redundancy of information!

Question Answering

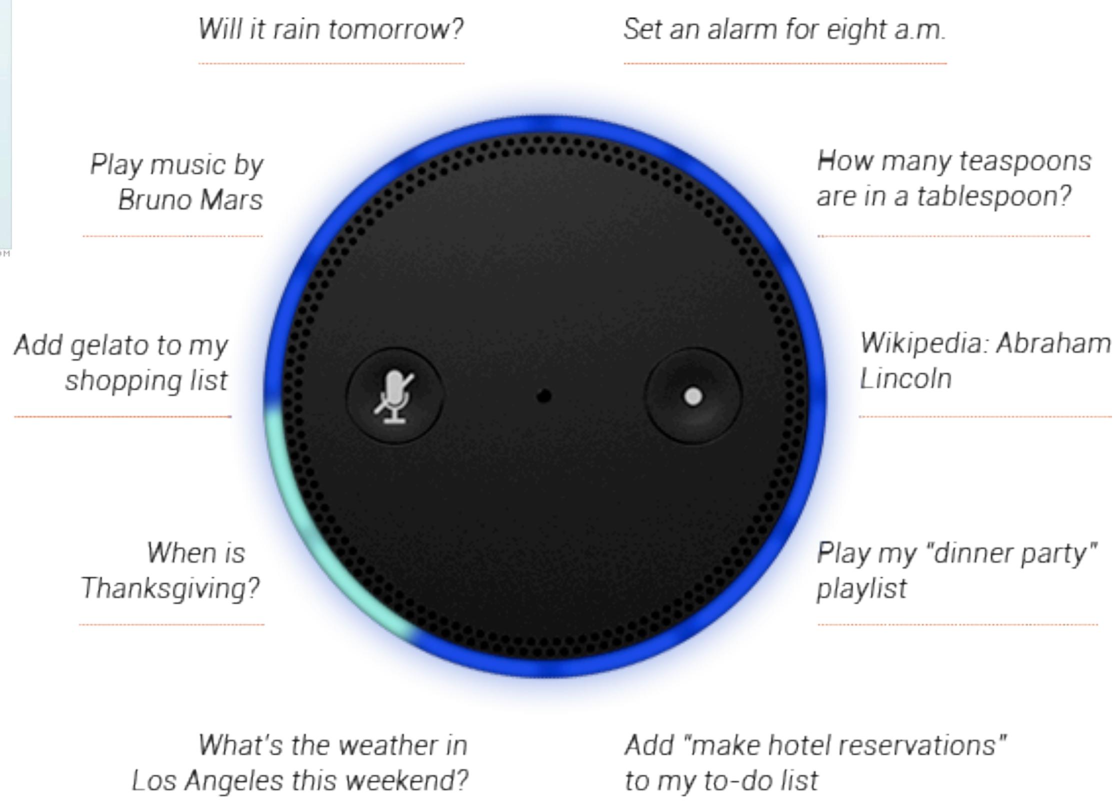
- Question Answering:
  - More than search
  - Can be really easy: “What’s the capital of Wyoming?”
  - Can be harder: “How many US states’ capitals are also their largest cities?”
  - Can be open ended: “What are the main issues in the global warming debate?”
- Natural Language Interaction:
  - Understand requests and act on them
  - “Make me a reservation for two at Quinn’s tonight”

Human-Machine Interactions





Will this Be Part of All Our Home Devices?

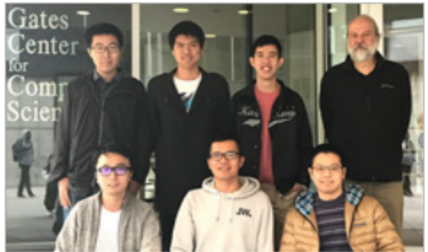


amazon alexa prize

Amazon Selects Teams to Compete for Inaugural $2.5 Million Alexa Prize

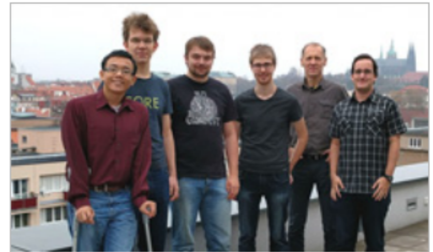
Carnegie Mellon University

CMU Magnus



Carnegie Mellon University

Ruby Star



Czech Technical University in Prague

Alquist

Heriot-Watt University

What's Up Bot

Princeton University

Pixie

Rensselaer Polytechnic Institute

Wise Macaw

University of California, Berkeley

Eigen

University of California, Santa Cruz

SlugBot

University of Edinburgh

Edina

University of Montreal

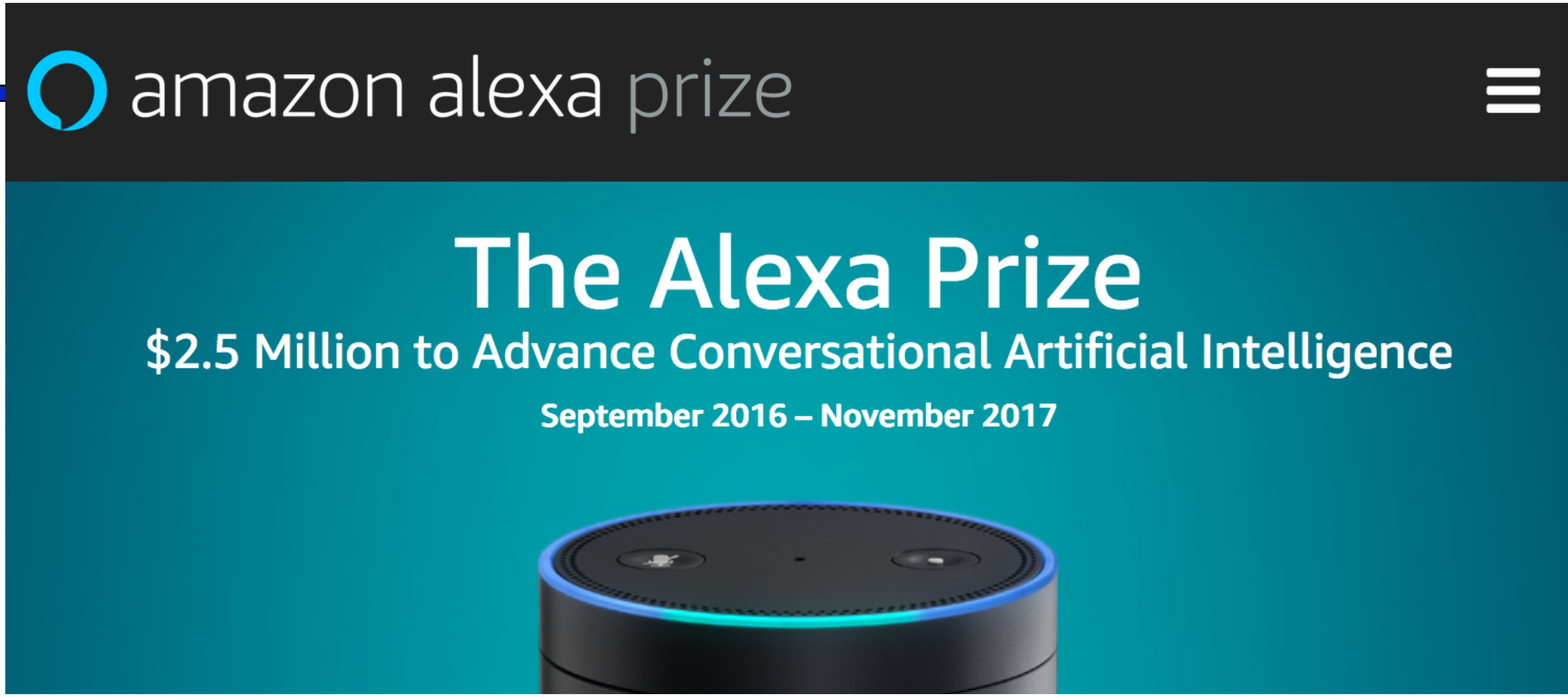
MILA Team

University of Trento

Roving Mind

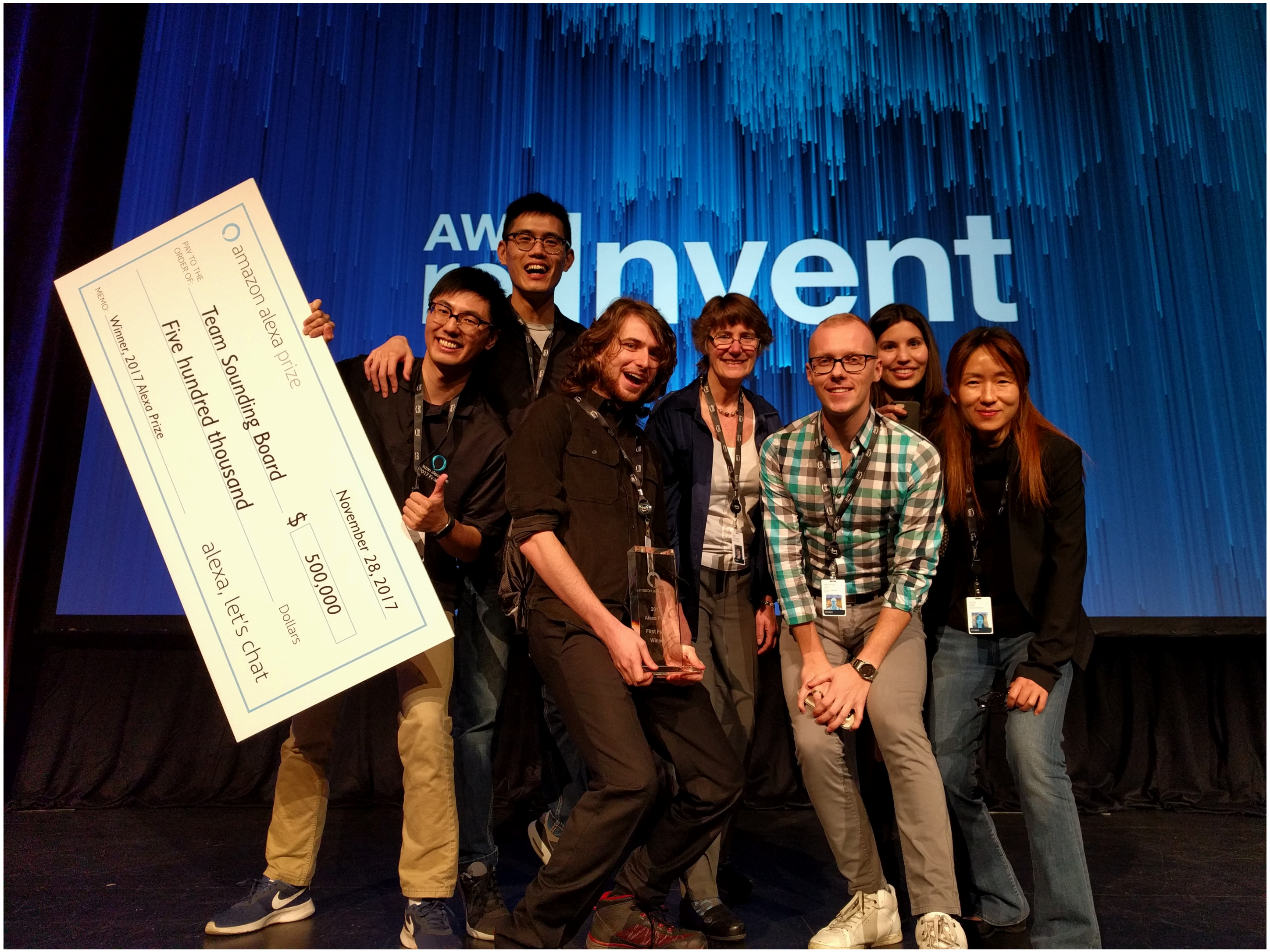
University of Washington

Sounding Board



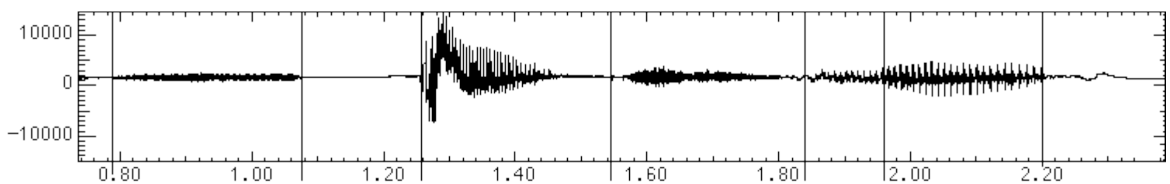
- UW Sounding Board among 3 Finalists!
- Final competition in Las Vegas in Nov
- Unclear if any team will make the 20 min goal
- How not to win:
  - Brute force more data, more depth
  - Add RL and pray magic will arise

Announced at AWS re:INVENT


amazon alexa prize
PAY TO THE ORDER OF
Team Sounding Board
November 28, 2017
$ 500,000
Five hundred thousand
Dollars
MEMO
Winner, 2017 Alexa Prize
alexa, let's chat

Speech Recognition

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

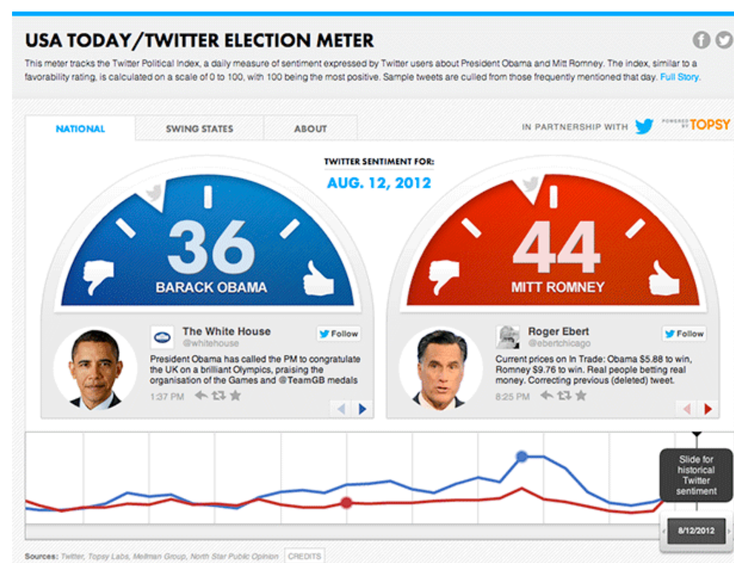


"Speech Lab"

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)

Analyzing public opinion, making political forecasts

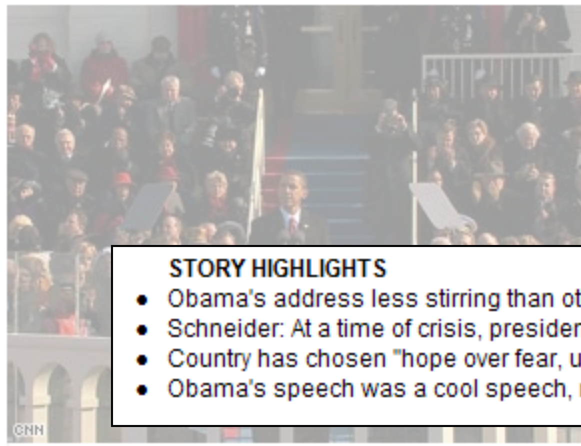
- Today: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- Past: "Sentiment Analysis" research started in 2002
- Future: computational social science and NLP for digital humanities (psychology, communication, literature and more)
- Challenge: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages



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 »

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 ... egin

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

... 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.

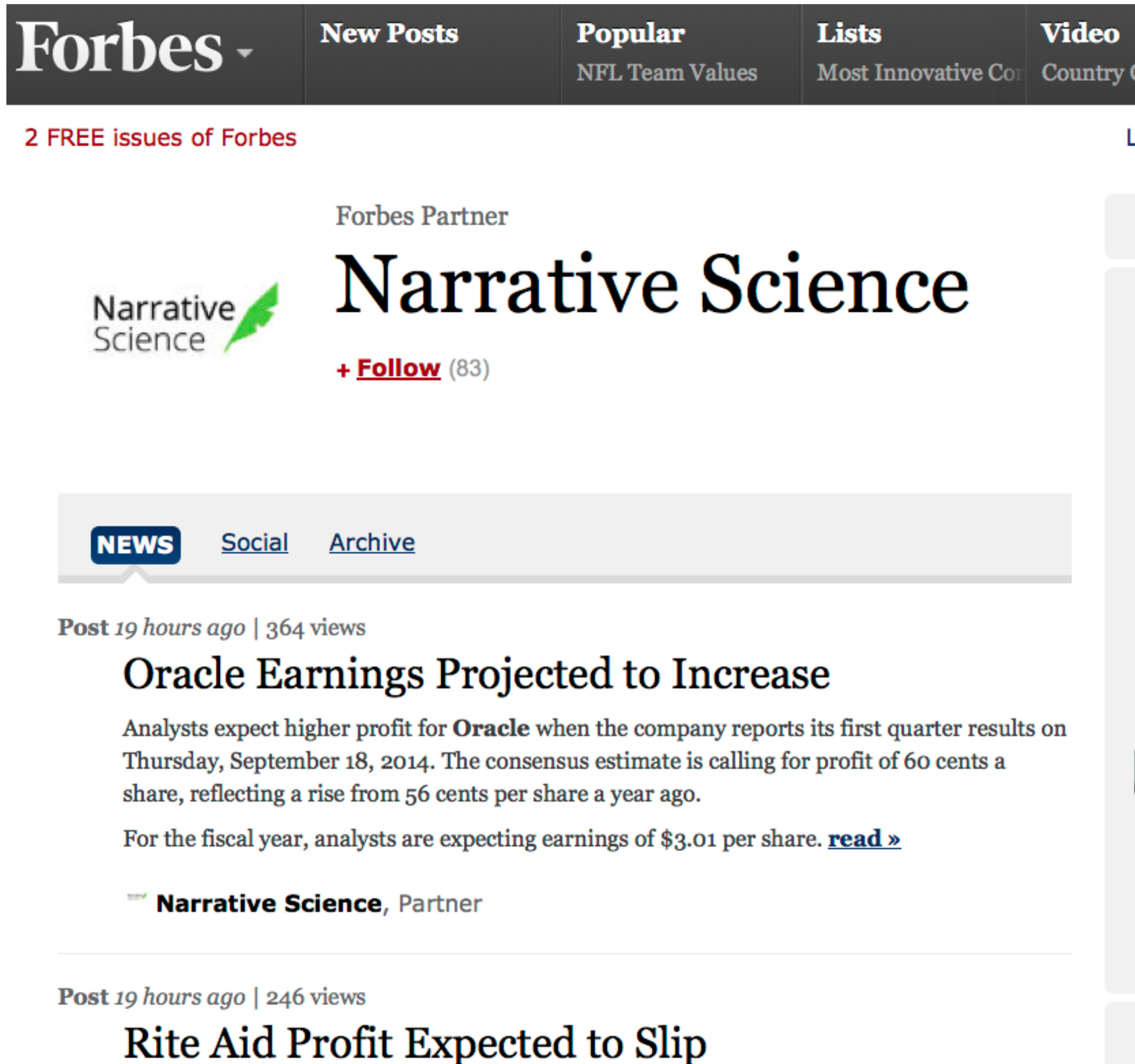
Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

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

Writer-bots for earthquake & financial reports

Some of the formulaic news articles are now written by computers.

•Definitely far from "Op-ed"
•Can we make the generation engine statistically learned rather than engineered?



Bot or human?

Despite an expected dip in profit, analysts are generally optimistic about **Steelcase** as it prepares to reports its third-quarter earnings on Monday, December 22, 2014. The consensus earnings per share estimate is 26 cents per share.

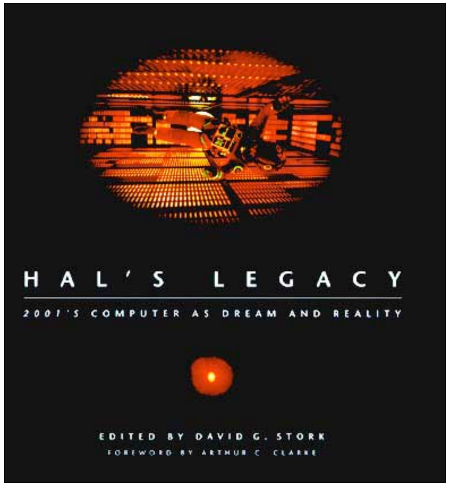
The consensus estimate remains unchanged over the past month, but it has decreased from three months ago when it was 27 cents. Analysts are expecting earnings of 85 cents per share for the fiscal year. Revenue is projected to be 5% above the year-earlier total of $784.8 million at $826.1 million for the quarter. For the year, revenue is projected to come in at $3.11 billion.

The company has seen revenue grow for three quarters straight. The less than a percent revenue increase brought the figure up to $786.7 million in the most recent quarter. Looking back further, revenue increased 8% in the first quarter from the year earlier and 8% in the fourth quarter.

The majority of analysts (100%) rate Steelcase as a buy. This compares favorably to the analyst ratings of three similar companies, which average 57% buys. Both analysts rate Steelcase as a buy.

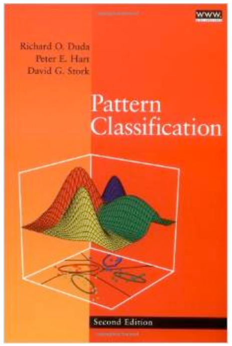
Steelcase is a designer, marketer and manufacturer of office furniture. Other companies in the furniture and fixtures industry with upcoming earnings release dates include: HNI and Knoll.

Language and Vision

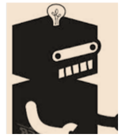


"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"

-- David Stork (HAL's Legacy, 2001) on A. Rosenfeld's vision

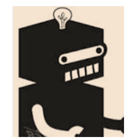


What begins to work (e.g., Kuznetsova et al. 2014)

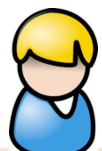


The flower was so vivid and attractive.

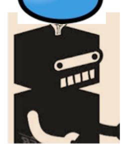
We sometimes do well: 1 out of 4 times, machine captions were preferred over the original Flickr captions:



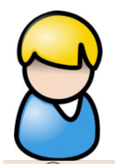
Blue flowers are running rampant in my garden.



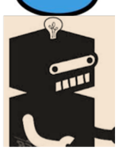
Spring in a white dress.



Blue flowers have no scent. Small white flowers have no idea what they are.



Scenes around the lake on my bike ride.



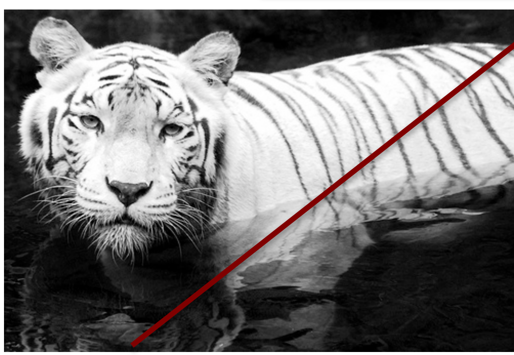
This horse walking along the road as we drove by.

But many challenges remain
(better examples of when things go awry)

The couch is definitely bigger than it looks in this photo.

Yellow ball suspended in water.

Incorrect Object Recognition



My cat laying in my duffel bag.

Incorrect Scene Matching

A high chair in the trees.

Incorrect Composition

Table of Content

- Definition of NLP
- Historical account of NLP

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

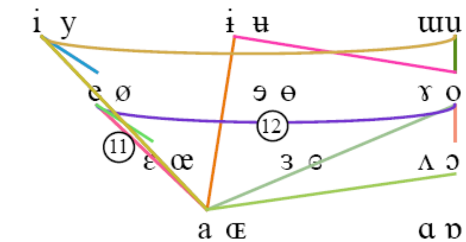
NLP: machine learning and empiricism

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

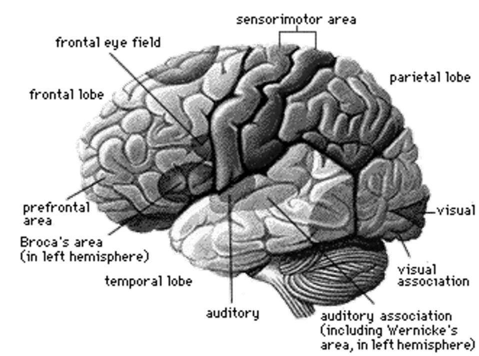
- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: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

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

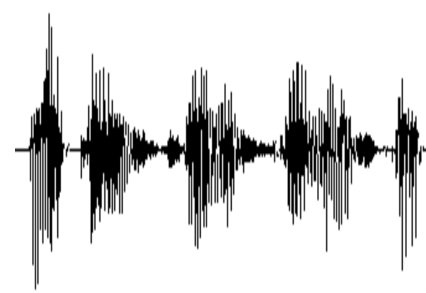


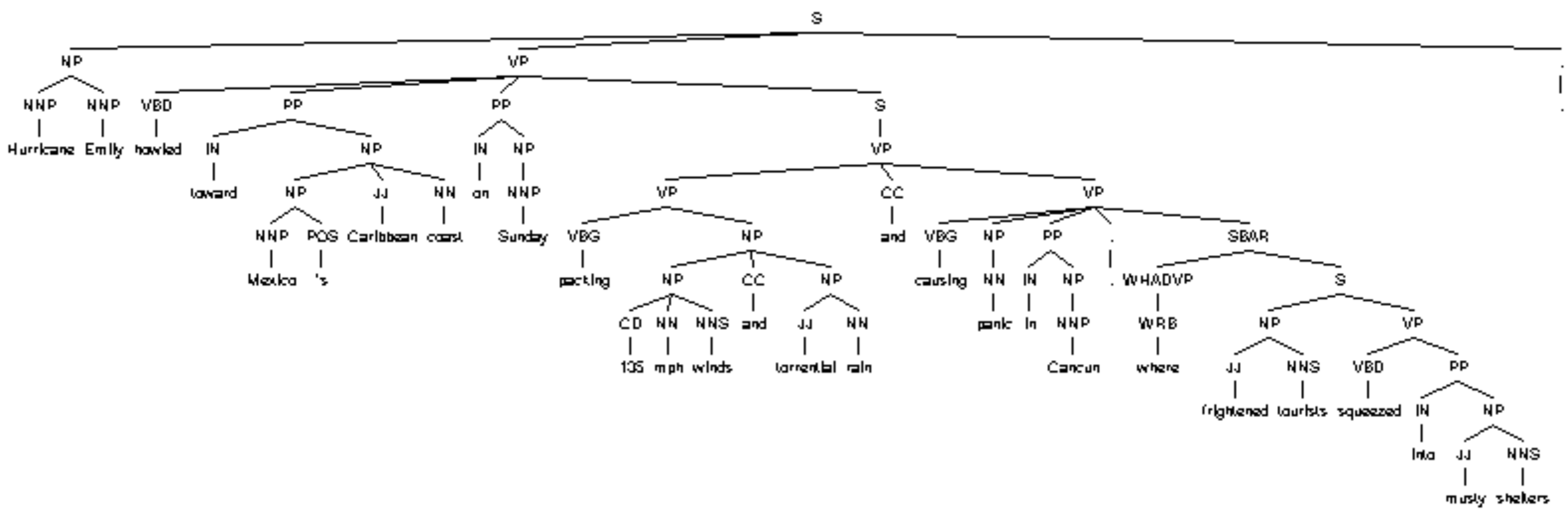
Table of Content

- Definition of NLP
- Historical account of NLP
- Unique challenges of NLP

Problem: Ambiguities

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Ban on Nude Dancing on Governor's Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?

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

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. ☺

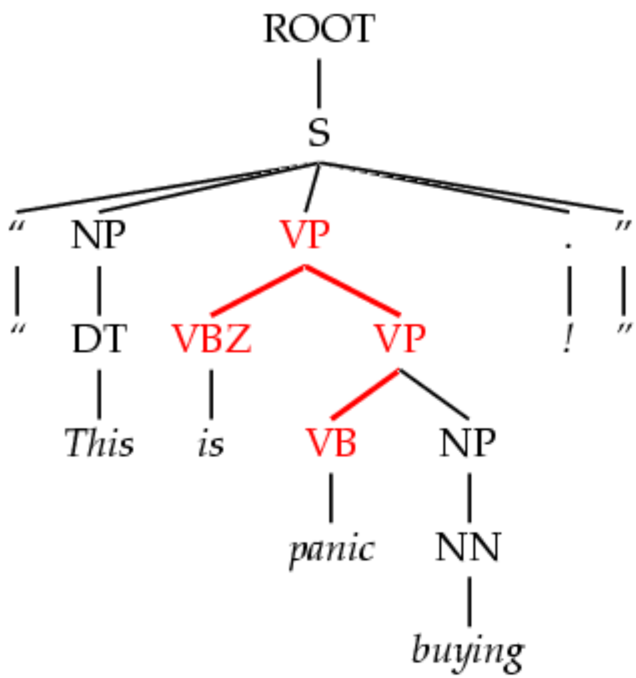
[Example from L. Lee]

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

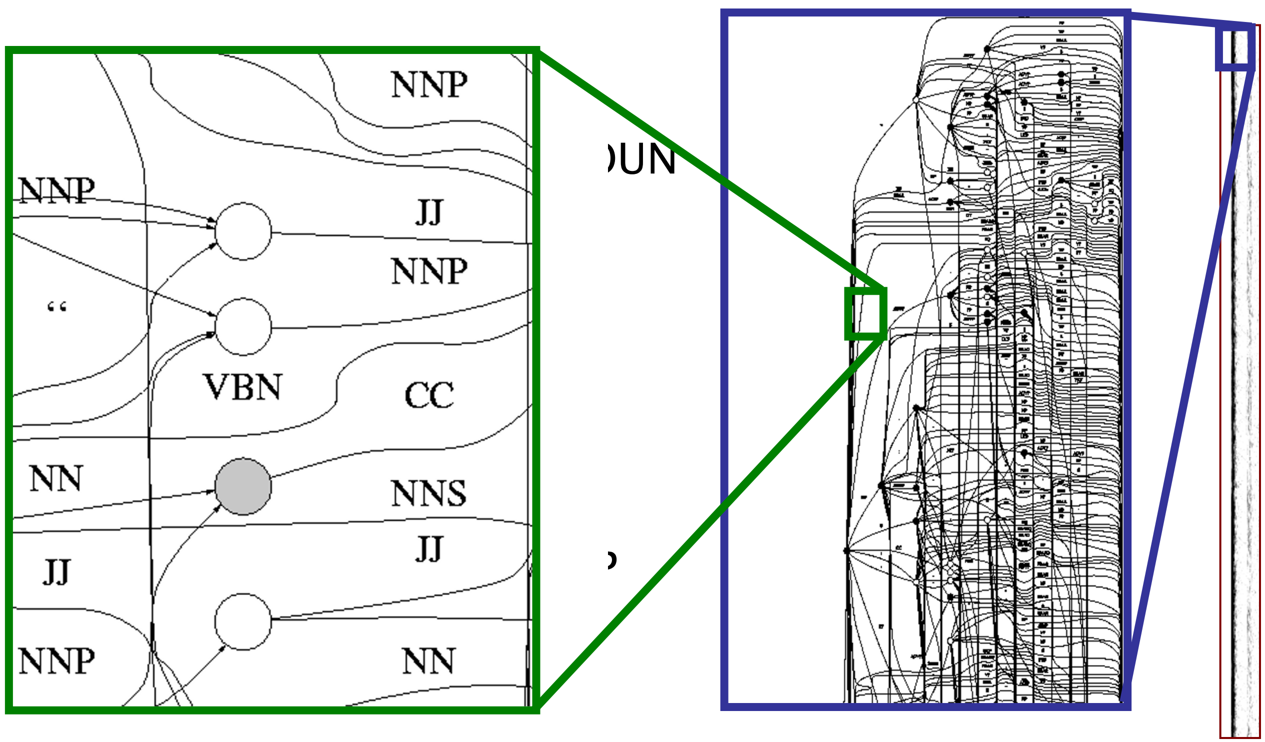
"This will panic buyers ! "



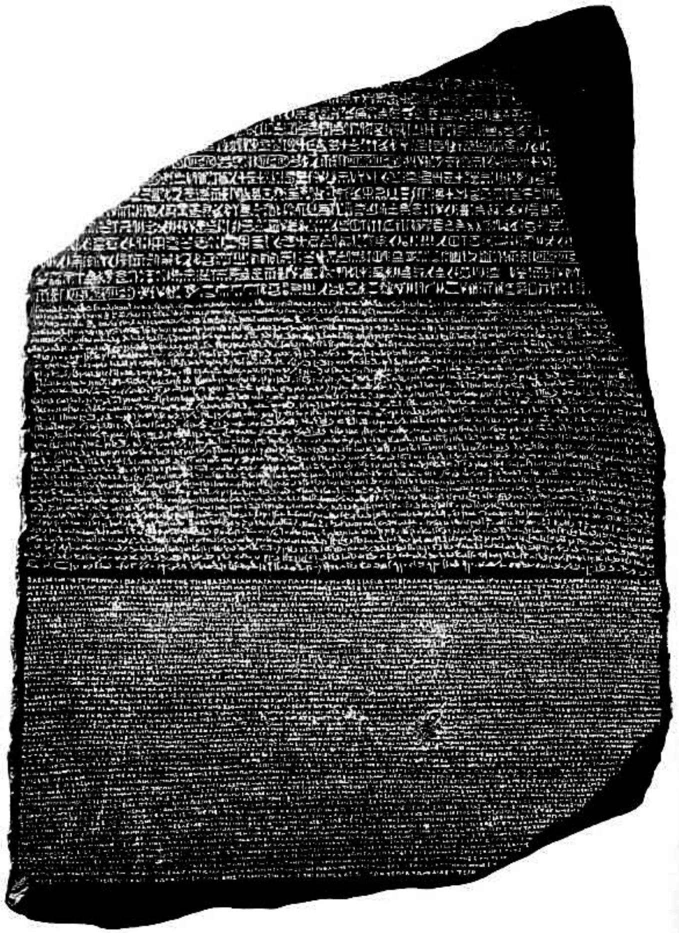
- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be "good" ones (or ruled out pragmatically)
  - …they didn't realize how bad it would be



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)

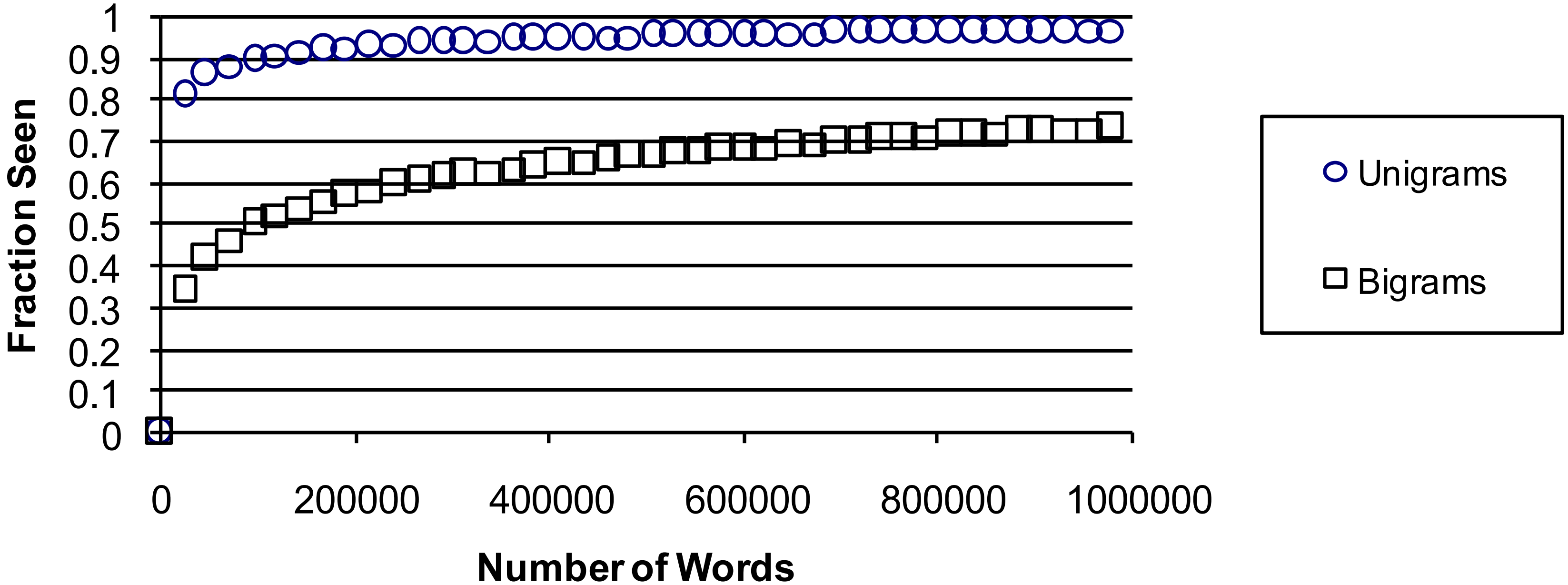
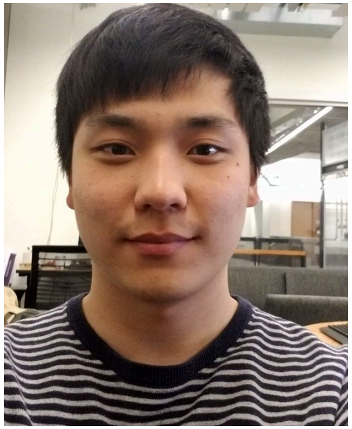


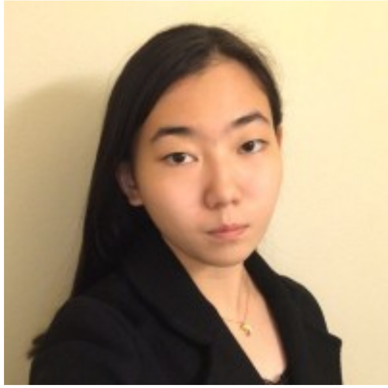
Table of Content

- Definition of NLP
- Historical account of NLP
- Unique challenges of NLP
- Class administrivia

Site & Crew

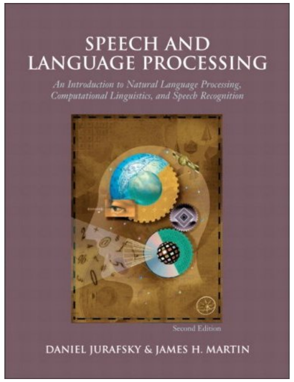
- Site: https://courses.cs.washington.edu/courses/csep517/21wi/
- Canvas: https://canvas.uw.edu/courses/1445050
- Crew:
- Instructor:
  Yejin Choi (office hour: Mon/Wed 8:00 – 8:30pm)
- TA:
  Xiujun Li
  James Park
  Peter West
  Ximing Lu

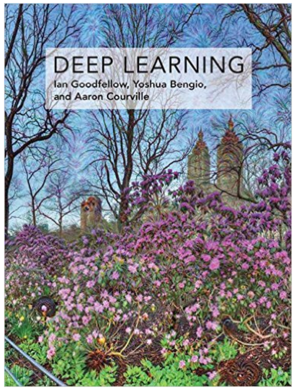


Textbooks and Notes

- Textbook (recommended but not required):
  - Recommended: [JM-ed3] Dan Jurafsky and James H. Martin, “Speech and Language Processing”. draft chapters of the third edition available online at https://web.stanford.edu/~jurafsky/slp3/
  - Optional: [JM-ed2] Dan Jurafsky and James H. Martin, the second edition of the above, 2009.
  - Optional: [GBC] Ian GoodFellow, Yoshua Bengio, and Aaron Courville, "Deep Learning" (online book at http://www.deeplearningbook.org/)





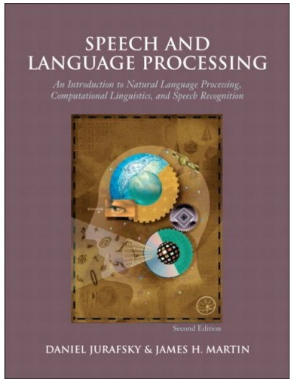
- Lecture slides & notes are required
  - See the course website for details
- Assumed Technical Background:
  - Data structure, algorithms, strong programming skills, probabilities, statistics

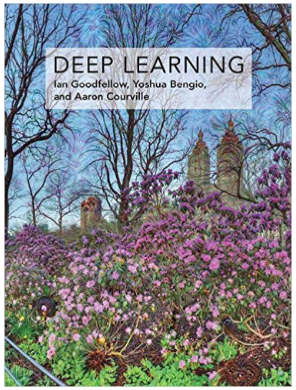
Grading & Policy

- Grading:
  - 5 programming assignments (75%)
  - Self-defined final project (15%)
  - Participation (in-class and online forums) (10%)

- Policy:
  - All homework will be completed individually.
  - Final projects can be done in groups of 1-3 people.
  - Plagiarism will lead to grave consequences.

- Participation and Discussion:
  - Class participation is expected and appreciated!!!
  - Email is great, but please use the message board when possible (we monitor it closely)





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?
  - Statistical Modeling Methods
    - 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…

Approximate Schedule

Week	Dates	Topics & Lecture Slides
1	Jan 4, 6	**Introduction** [slides (old)] **Probabilistic Models of Words:** N-gram Language Models (LMs) [slides (old)]
2	Jan 11, 13	**Probabilitic Models of Sequences**: Hidden Markov Models (HMMs) [slides (old)]
3	Jan 20 (Jan 18: MLK day)	**Probabilitic Models of Sequences:** Hidden Markov Models (HMMs) **& EM** [slides (old)]
4	Jan 25, 27	**Learning (Feature-Rich Models):** Log-Linear Models, Maximum Entropy Models [slides (old)]
5	Feb 1, 3	**Learning (Structural Graphical Models):** Conditional Random Fields (CRFs) [slides (old)]
6	Feb 8, 10	**Neural Language Models:** Basics [slides (old)]
7	Feb 17 (Feb 15: Presidents Day)	**Neural Language Models:** Tranformers
8	Feb 22, 24	**Neural Language Generation:** Summarization, Dialogues, and De-generation
9	Mar 1, 3	**Neural Knowledge and Reasoning:** On Commonsense Intelligence
10	Mar 8, 10	**New Challenges with Neural Models:** Interpretability, Bias, and Ethics

Comparisons with Other Classes

- Compared to ML
  - Typically multivariate, dynamic programming everywhere
  - Structural Learning & Inference
  - Insights into language matters (a lot!)
  - DL: RNNs, LSTMs, Seq-to-seq, Attention, …
- Compared to CompLing classes
  - More focus on core algorithm design, technically more demanding in terms of math, algorithms, and programming

Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics
    - Decent coding skills
    - Data structure and algorithms (dynamic programming!)
    - (Optional) basic linguistics background
  - ML/AI helps if you've taken either before, but not necessary
- Class goals
  - Learn the fundamental concepts and techniques
  - Learn current engineering practices
  - Learn how to advance the field!