CS5740: Natural Language Processing

Introduction

Instructor: Yoav Artzi
TAs: Xinya Du and Valts Blukis

Slides adapted from Dan Klein, Luke Zettlemoyer, and Yejin Choi

Technicalities

- People:
  - Instructor: Yoav Artzi
  - TAs: Xinya Du and Valts Blukis
- Webpage (everything is there):
  - http://www.cs.cornell.edu/courses/cs5740/2019sp/
- Discussion group on Piazza
- Assignments on CMS
  - Repositories on Github Classroom

Technicalities

- Office hours posted on website
  - No office hours today and next week – schedule by appointment
- Example assignments and report layout posted on CMS
- Enrollment
  - Not formally enrolled → email your NetID for quizzes and CMS
  - If you are in the system, you should see the course in CMS, if not, you are not in the system
  - This will continue until enrollment stabilizes (February 8)
- Slides always posted after class, often appended to the most recent deck lecture

Procedurals

https://www.cs.cornell.edu/courses/cs5740/2019sp/procedurals.html

Quizzes

- It is not possible to re-take a missed quiz
- A missed quiz gets zero
- Just like an exam: no copying, chatting, and not taking the quiz remotely → all AI violations
- Come on time
  - Late? Enter quietly and sit at the back
  - Quiz starts on time
- Quiz practice
  - Phones, tablets, or laptops
  - http://socrative.com
  - Select “Student Login”
  - Today’s room: NLP19
  - Use NetID to identify
- After the quiz: please put your electronics aside

Tips

- Work together with your partner, don't simply divide the work
- Discuss with each other
  - Beyond your group
  - This is what Piazza is for!

What is this class?

- Depth-first technical NLP course
- Learn the language of natural language processing
- What this class is not?
  - It is not a tutorial to NLTK, PyTorch, etc.
  - There are many resources online that already do that well

Class Goals

- Learn about the issues and techniques of modern NLP
- Be able to read current research papers
- Build realistic NLP tools
- Understand the limitation of current techniques

Main Themes

- Linguistic Issues
  - What are the range of language phenomena?
  - What are the knowledge sources that let us make decisions?
  - What representations are appropriate?
- 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

Main Models

- Generative Models
- Discriminative Models
  - Neural Networks
- Graphical Models

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:
  - Complex:

What is NLP?

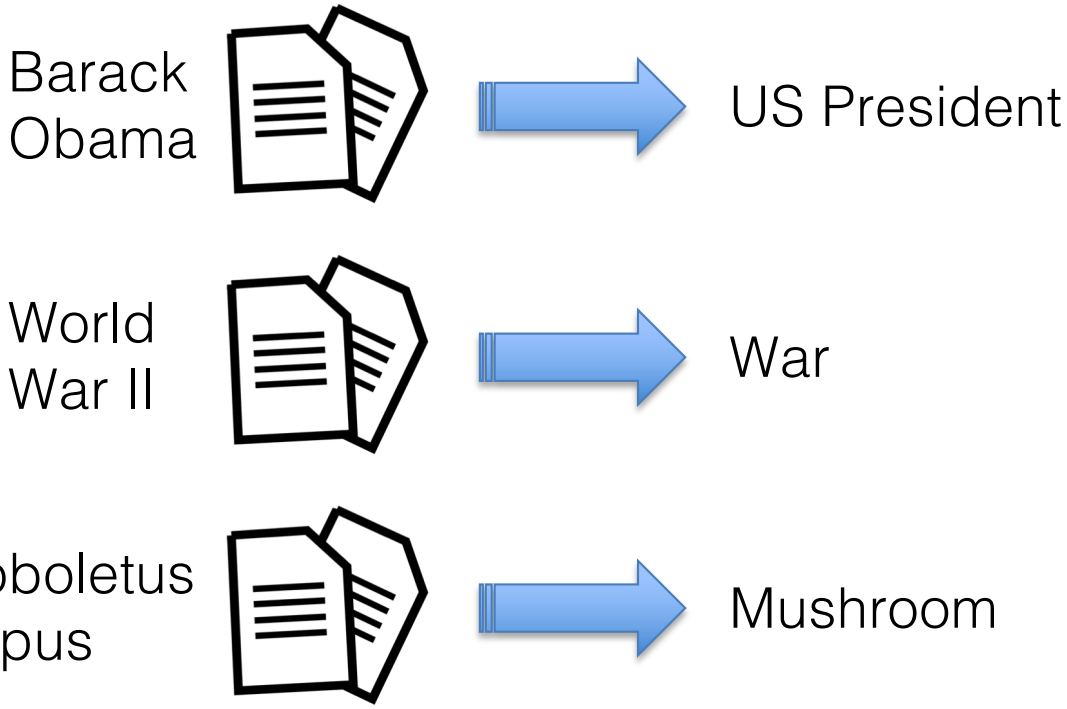
- Fundamental goal: deep understanding of broad language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)

Today

- Prominent applications
  - Try to imagine approaches
  - What's behind current limitations?
- Some history
- Key problems
- As much as time allows: meta NLP + text classification

Text Categorization

- Input: Document
- Output: Category assignment



Information Extraction

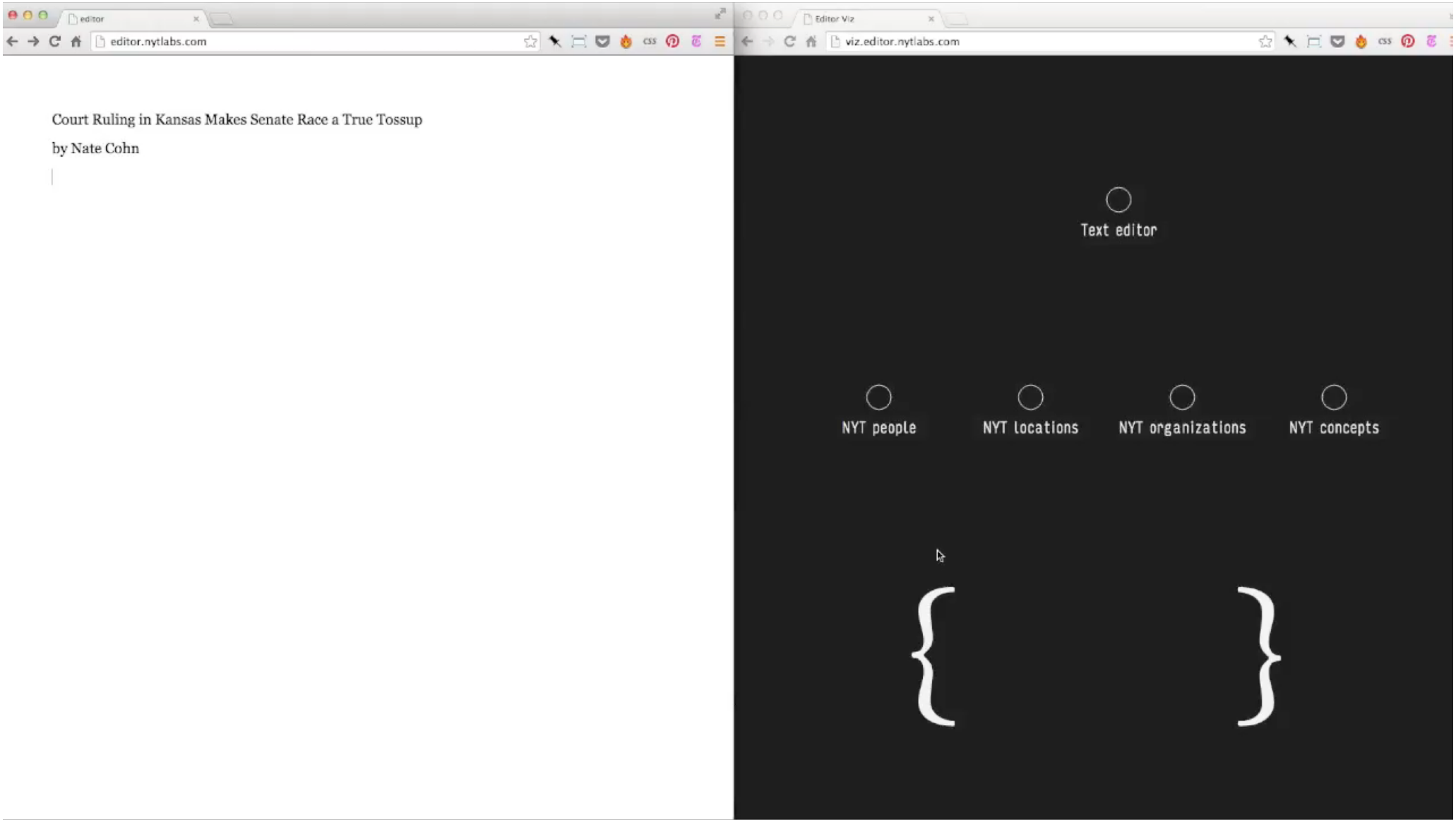
- Unstructured text to database entries

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: good performance on simple templates (e.g., person-role)
- Harder without defining template

Tagging: Back to Text



http://nytlabs.com/projects/editor.html

Machine Translation

Japan's economy turns red for the first time since Fukushima

After reaching a record deficit in 2014, Japan posted a trade surplus for the first time since the 2011 nuclear accident.

L'économie japonaise sort du rouge pour la première fois depuis Fukushima

Après avoir atteint un déficit record en 2014, le Japon dégage un excédent commercial pour la première fois depuis l'accident nucléaire de 2011.

Machine Translation

Le Monde.fr

La Bourse de Shanghai dégringolait de plus de 6 % mardi 25 août à l'ouverture, après s'être déjà effondrée de presque 8,5 % la veille, dans un marché affolé par l'affaiblissement persistant de l'économie chinoise et miné par des inquiétudes sur la conjoncture mondiale.

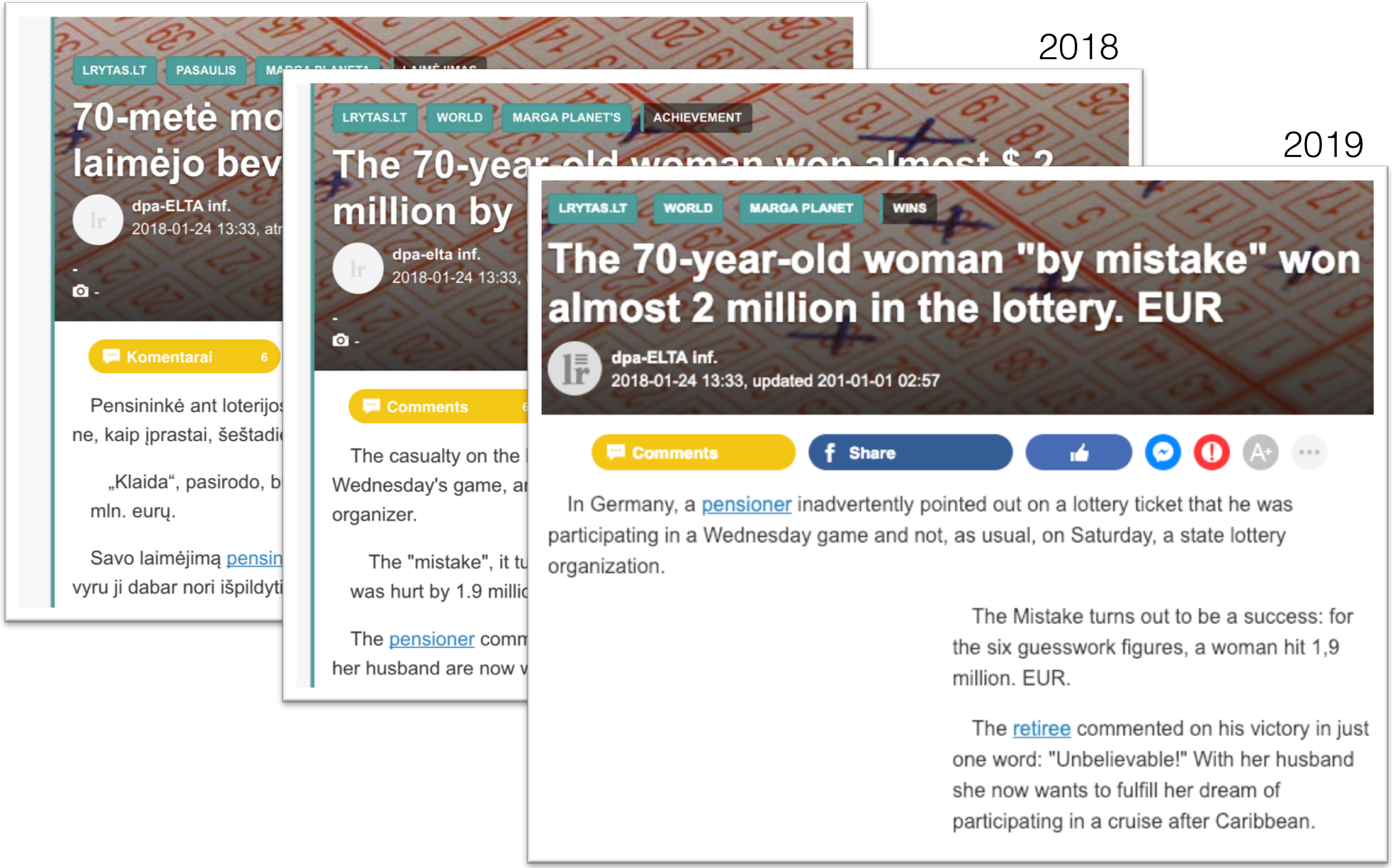
Dans les premiers échanges, l'indice composite chutait de 6,41 % soit 205,78 points, à 3 004,13 points. La Bourse de Shenzhen plongeait quant à elle de [illegible] a ouvert en

The Shanghai Stock Exchange tumbled more than 6% Tuesday, August 25 at the opening, having already collapsed by almost 8.5% yesterday , in a panicked market the persistent weakening of the Chinese economy and undermined by concerns about the global economy.

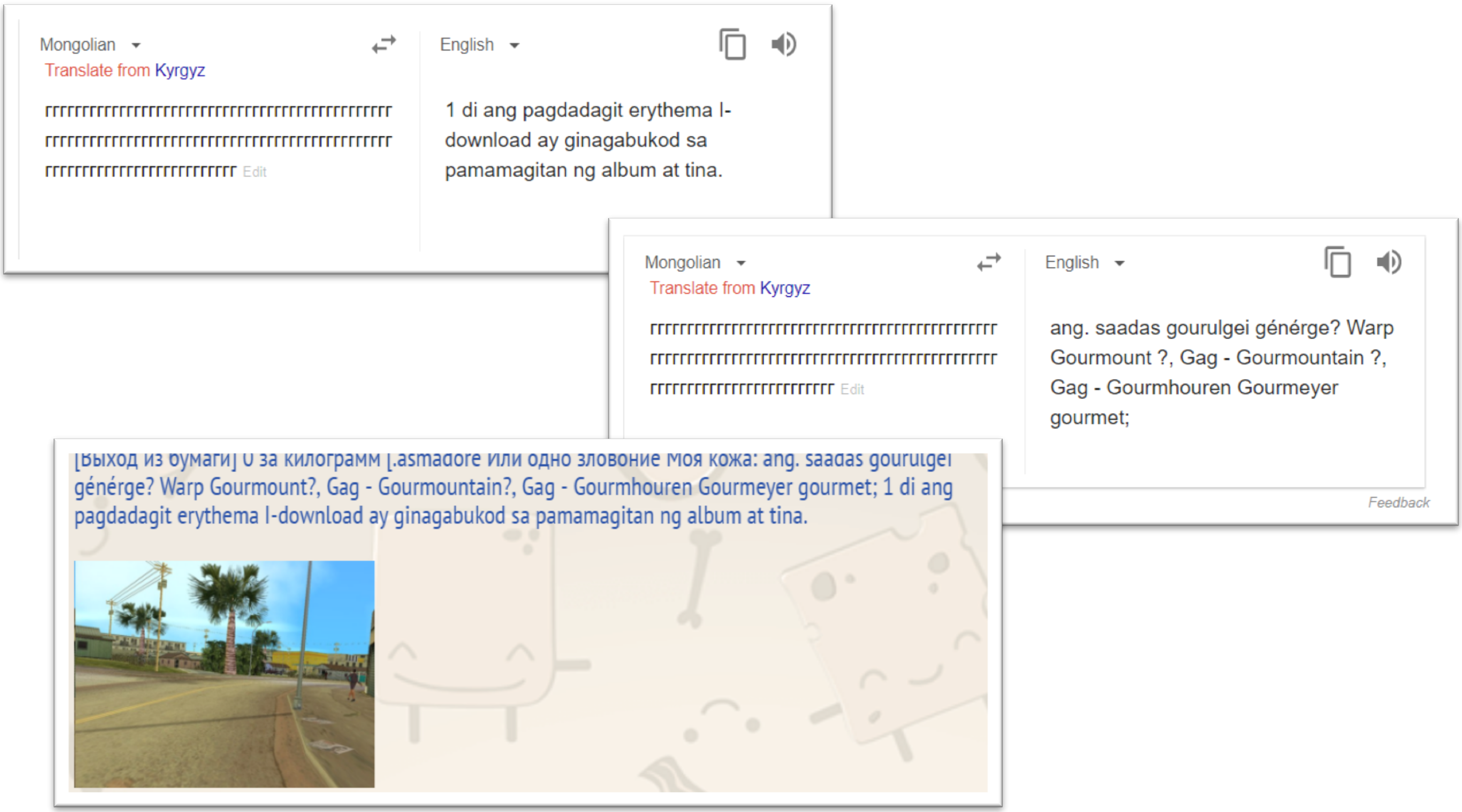
In early trade, the composite index fell by 6.41% or 205.78 points to 3 004.13 points. The Shenzhen Stock Exchange dived for its 6.97% to 1 751.28 points. The Hong Kong Stock Exchange, meanwhile, opened down 0.67%.

Machine Translation

lrytas.lt

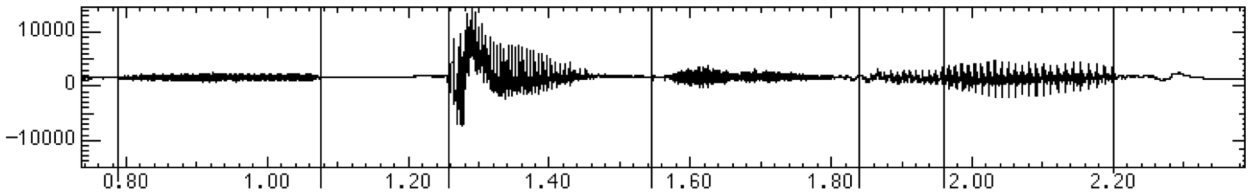


Machine Translation



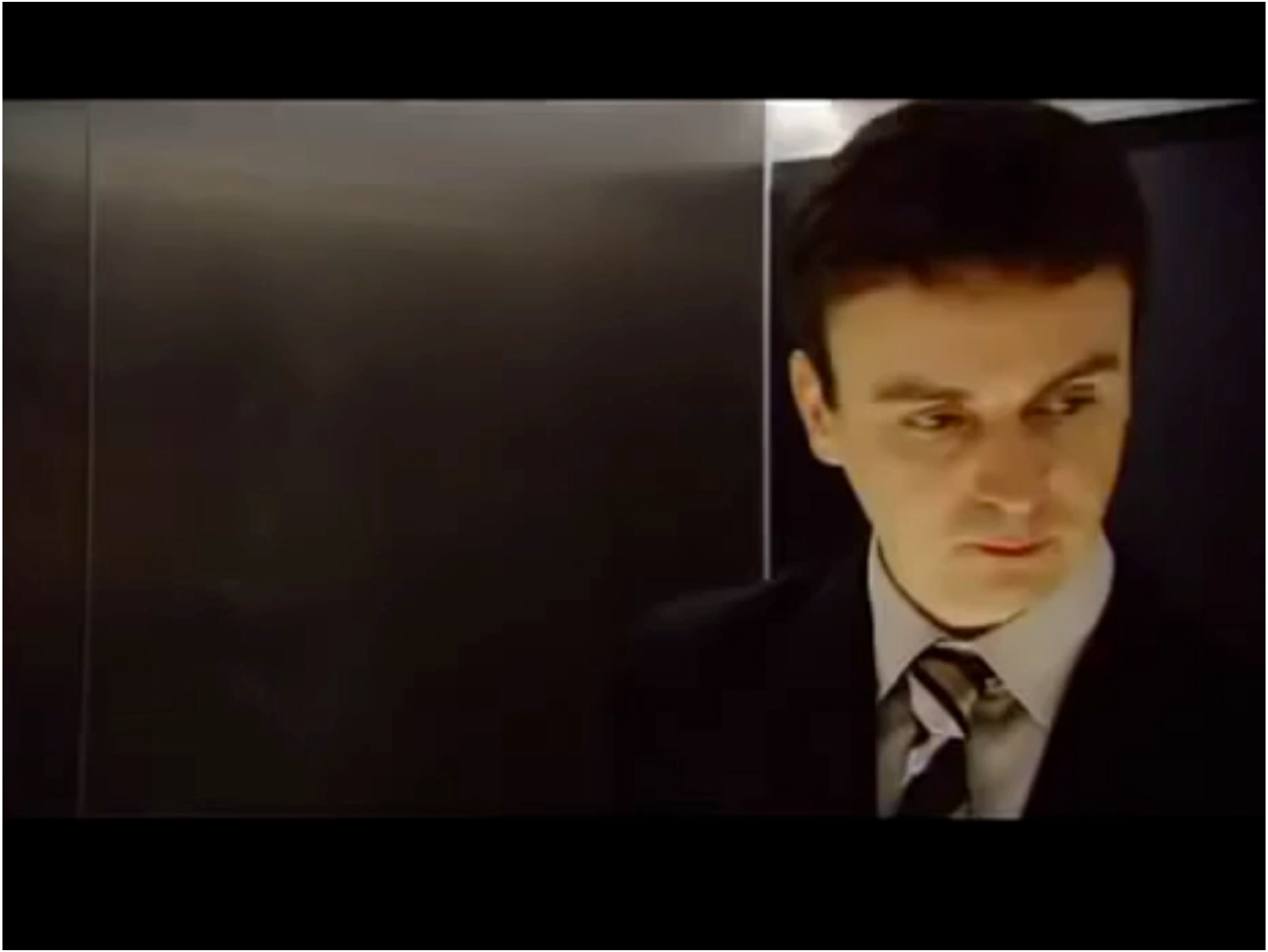
Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 16% PER, Google claims 8% WER



“speech lab”

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: mechanical and monotone



Personal Assistants

Exciting Times

- We can do a lot of things
- But:
  - Must think beyond the NLP technique or the cool system we build
  - Don’t forget what we still don’t know

NLP History: Pre-statistics

(1) Colorless green ideas sleep furiously.

(2) Furiously sleep ideas green colorless

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)

NLP History: Pre-statistics

- 70s and 80s: more linguistic focus
  - Emphasis on deeper models, syntax and semantics
  - Toy domains / manually engineered systems
  - Weak empirical evaluation

NLP History: ML 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

NLP History: ML 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

"Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, "Every time I fire a linguist, my performance goes up"). The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high-performance broad-domain language understanding systems …"

- Lillian Lee (2001) http://www.cs.cornell.edu/home/llee/papers/cstb/index.html

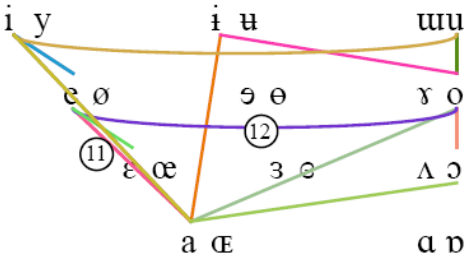
NLP History: ML and Empiricism

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

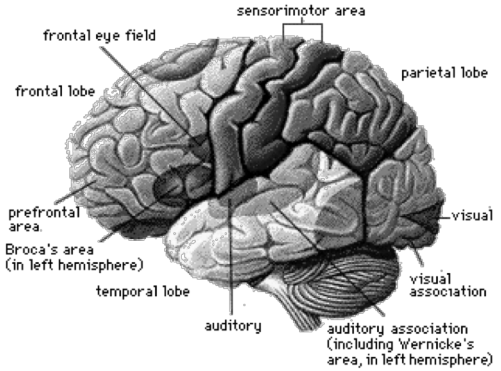
- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: NLP+X, excitement about neural networks (again), and …

Related Fields

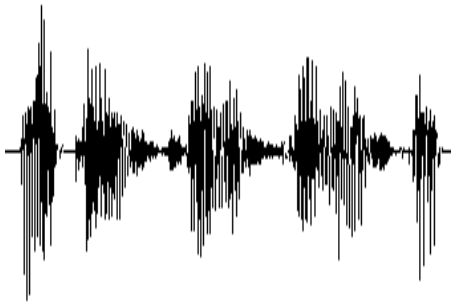
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- 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



Key Problems

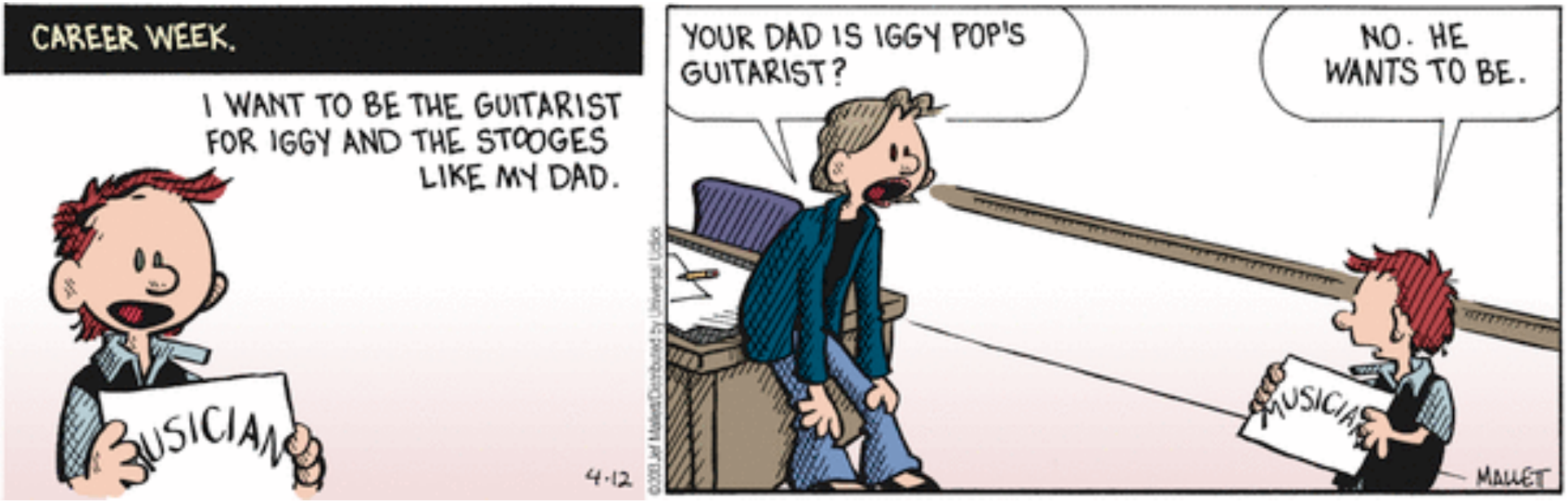
We can understand programming languages.
Why is NLP not solved?

Key Problems

We can understand programming languages.
Why is NLP not solved?

- Ambiguity
- Scale
- Sparsity

Key Problem: Ambiguity

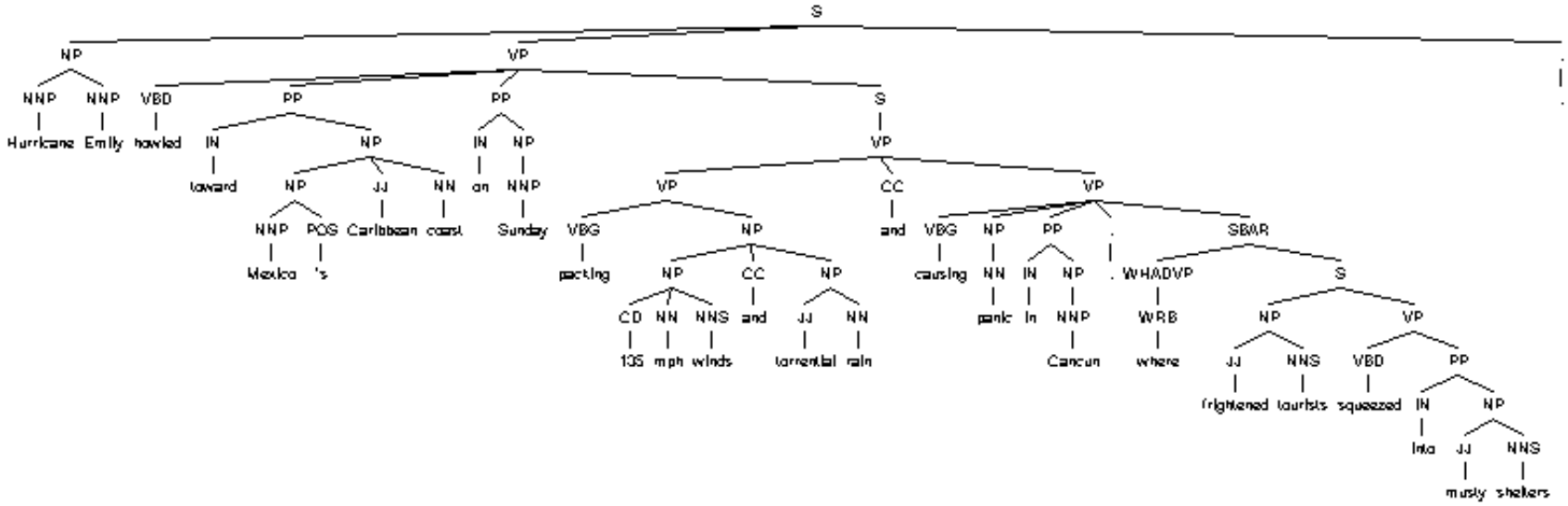


Key Problem: Ambiguity

- Some 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

Syntactic Ambiguity

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.

[Example from Lillian Lee]

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
- 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 Lillian Lee]

Ambiguities in the Wild

Ambiguities in the Wild: Context

Ambiguities in the Wild: Context

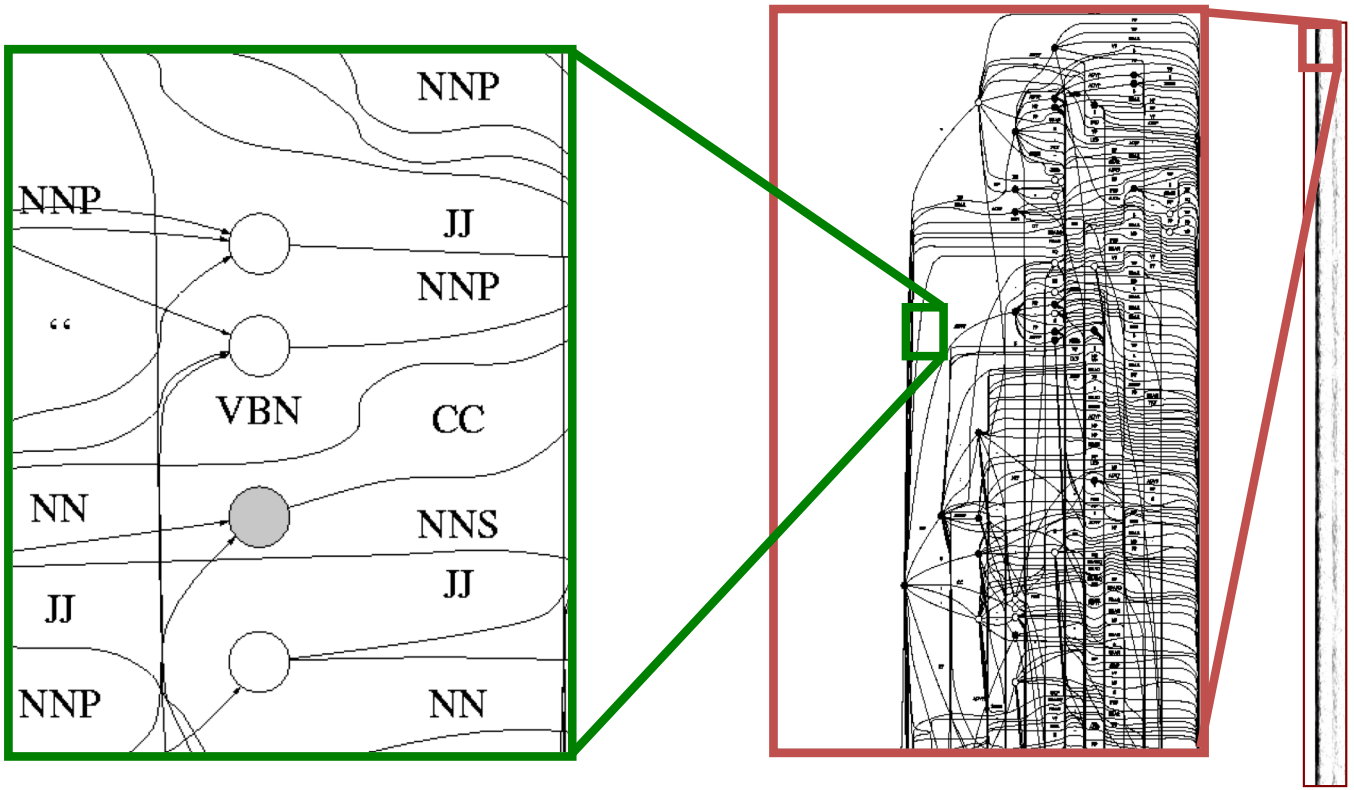
Ambiguities in the Wild: Context

Key Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be "good" ones (or ruled out pragmatically)

Key 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



Key Problem: Sparsity

- 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

Key Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair)