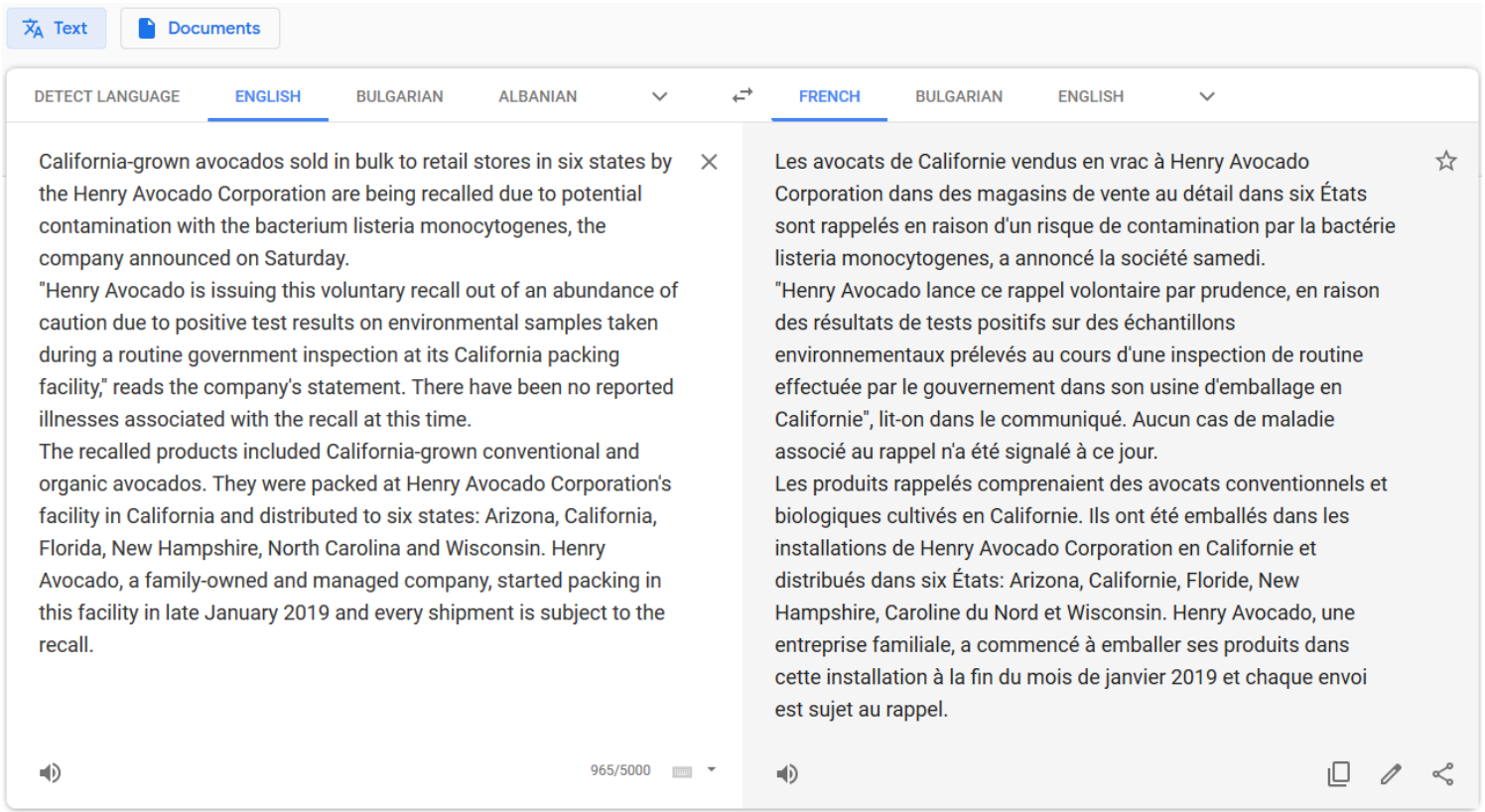
NLP

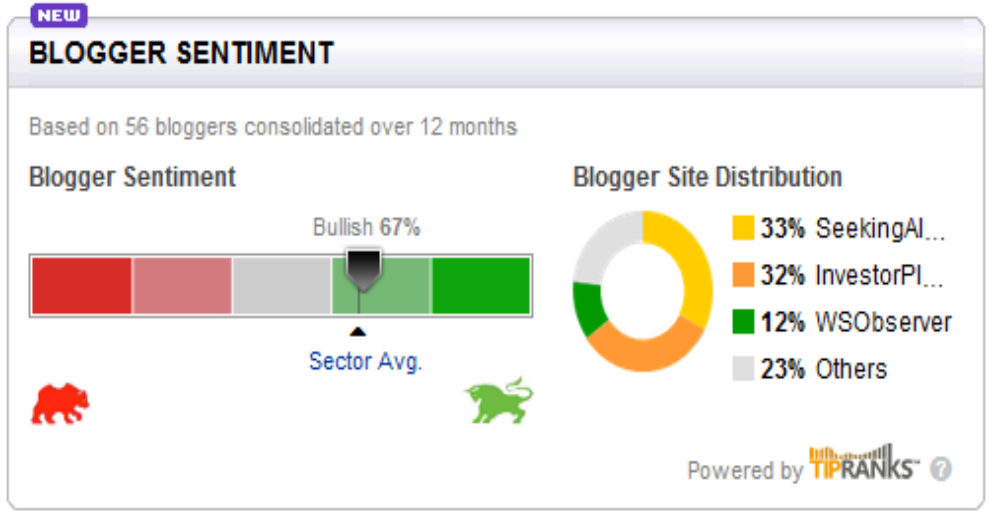
Natural Language Processing

111

Introduction

NLP Applications





http://www.geekwire.com/2013/ibm-takes-watson-cloud/

NLP Applications

- Search engines
  - Google, Bing, DuckDuckGo, Baidu, Yandex
- Question answering
  - IBM Watson
- Natural language assistants
  - Siri, Alexa, Google, Cortana
- Machine translation
  - Google Translate
- Email assistants
  - Autocomplete

What is Natural Language Processing

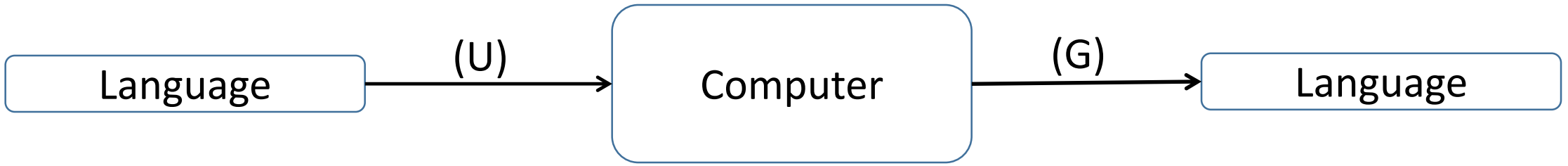
- Natural Language Processing (NLP)
  - The study of the computational treatment of natural (human) language.
  - In other words, building computers that understand (and generate) language.
- Computational Linguistics (CL)
  - The use of computers to study language
  - Applications to social science, literature, finance
- Computers are designed to deal with human language
  - Specific techniques are needed
- NLP draws on research in many fields
  - Linguistics, Theoretical Computer Science, Mathematics, Statistics, Artificial Intelligence, Psychology, Databases, etc.

Language and Communication

- Speaker
  - Intention (goals, shared knowledge and beliefs)
  - Generation (tactical)
  - Synthesis (text or speech)
- Listener
  - Perception
  - Interpretation (syntactic, semantic, pragmatic)
  - Incorporation (internalization, understanding)
- Both
  - Context (grounding)

Basic NLP Pipeline

- (U)nderstanding and (G)eneration



Representation Matters

Representation:
The Earth is fixed center of our Solar System

Representation:
The Sun is fixed center of our Solar System

Geocentric Model

(Anaximander, 6th century BC)

Heliocentric Model

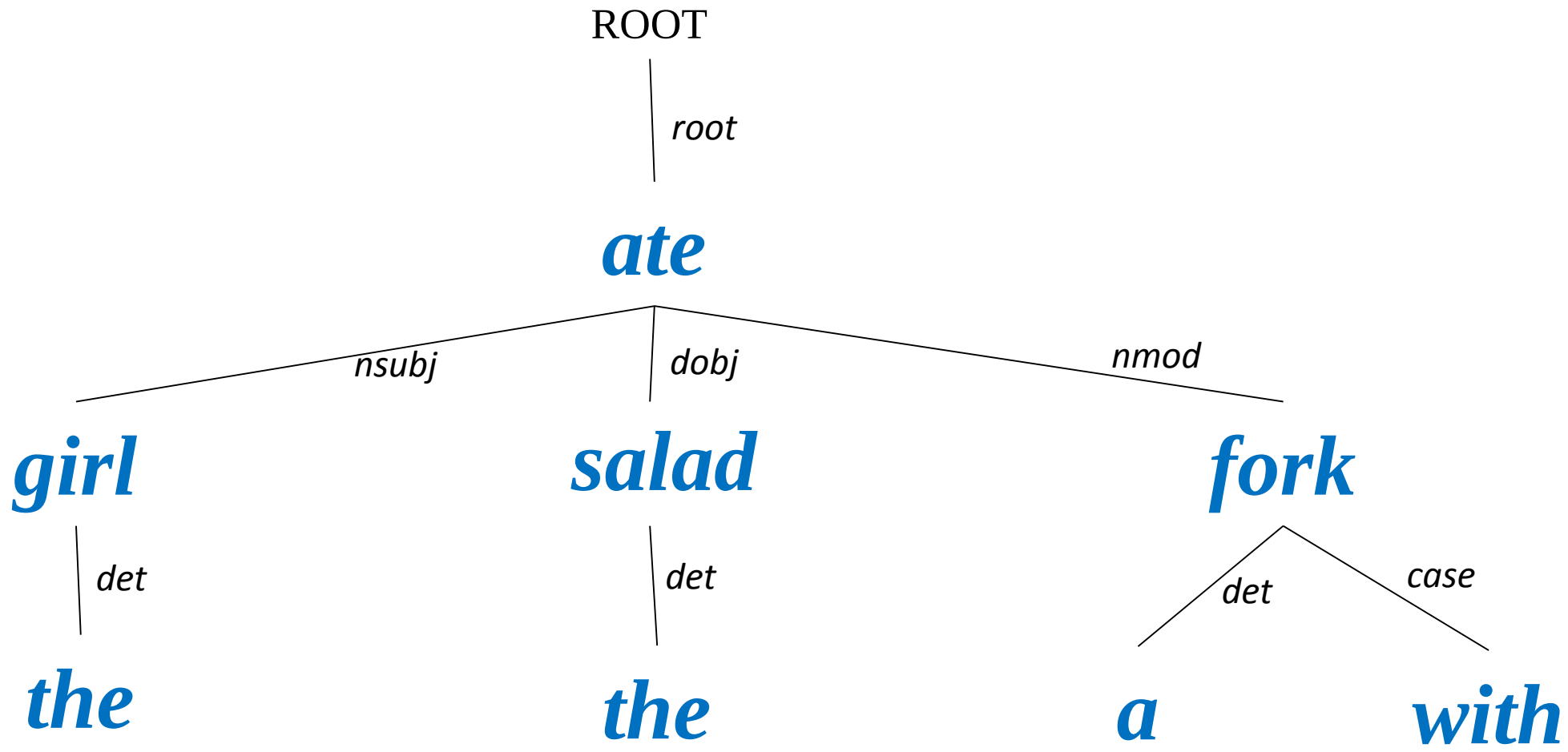
(Copernicus, 1543)

Natural Language Understanding

DET N VBD DET N PRP DET N

The girl ate the salad with a fork.

Sentence Representation



Logical Language Understanding

- Semantic Analysis

  Girl (g_1)
  Salad (s_1)
  Fork (f_1)
  EatingEvent (e_1) $\wedge$ Eater (e_1,g_1) $\wedge$ Eaten (e_1,p_1) $\wedge$ Instrument (e_1,f_1) $\wedge$ Time(e,"1:24 pm")

- World Knowledge

  $\forall$x: Salad (x) $\Rightarrow$ Food (x)

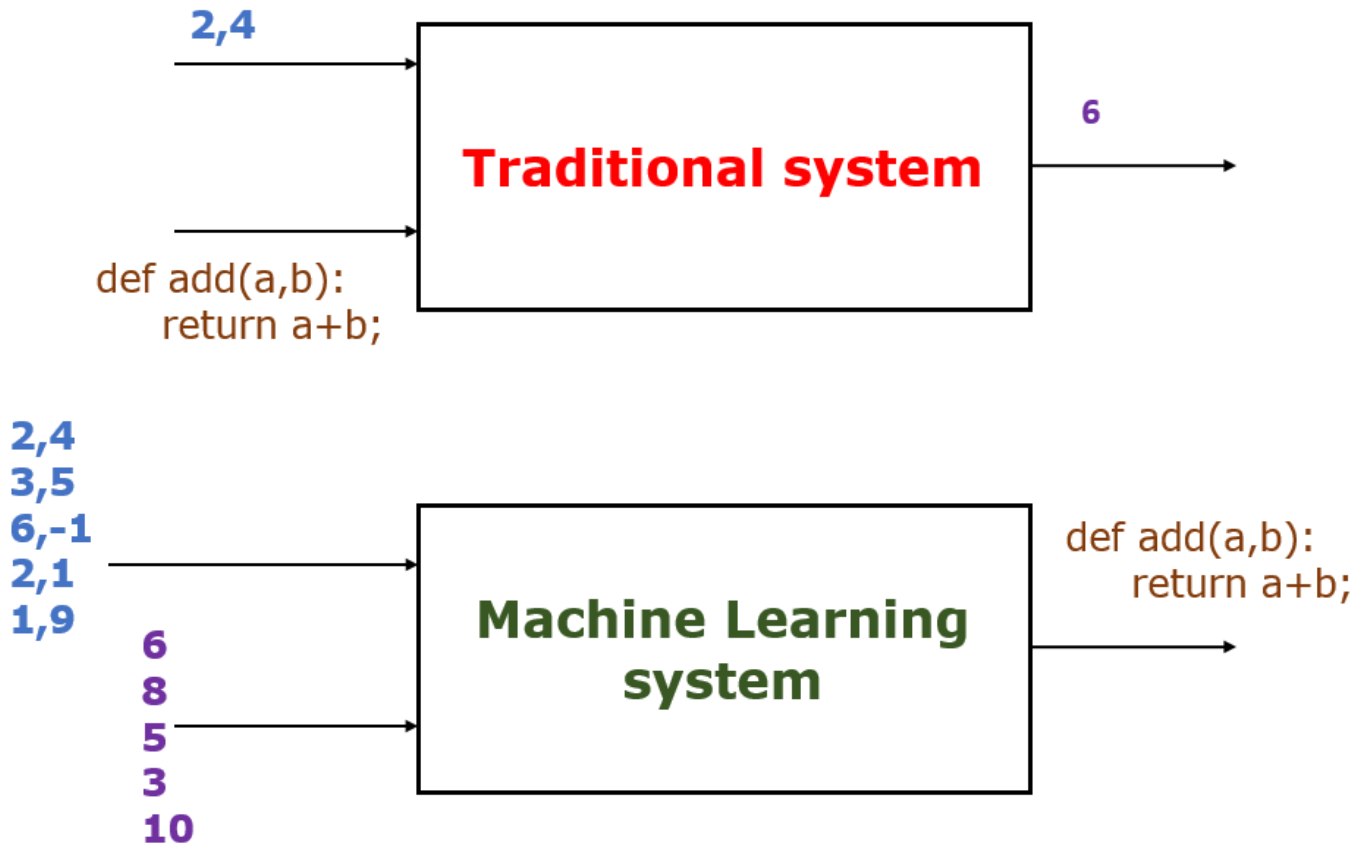
- Inference

  $\forall z,t_0,t_1,y,e$: Hungry (z,t_0) $\wedge$ EatingEvent(e) $\wedge$ Eater (e,z) $\wedge$ Eaten (e,y) $\wedge$ Time (e,t_0)
  $\wedge$ Food (y) $\wedge$ Precedes (t_0,t_1) $\Rightarrow \neg$Hungry (z,t_1)

- Conclusion

  $\neg$Hungry (g_1,now)

Machine Learning



Modern NLP Techniques (2013-2021)

- Deep Learning
- Word Embeddings
- Convolutional Neural Networks
- Recurrent Neural Networks
- Attention
- Transformers
- BERT

Distributed representations

- *acerola* is a significant source of vitamin C.
- the pulp of the *acerola* is very soft
- *acerola* are now found growing in most sub-tropical regions of the world.
- *acerola* can be eaten fresh or used to make jams or jellies.

Word Embeddings

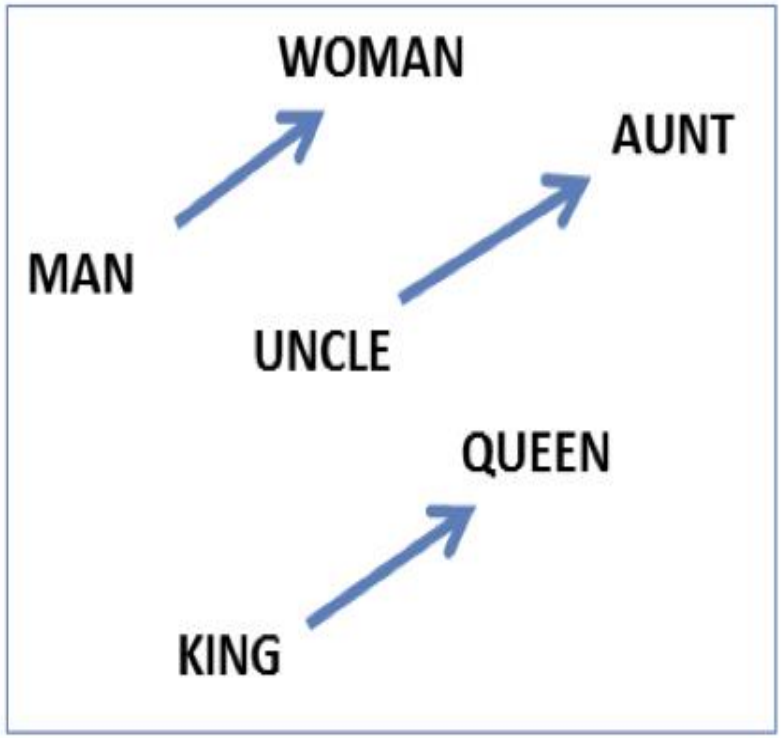
pharmaceuticals

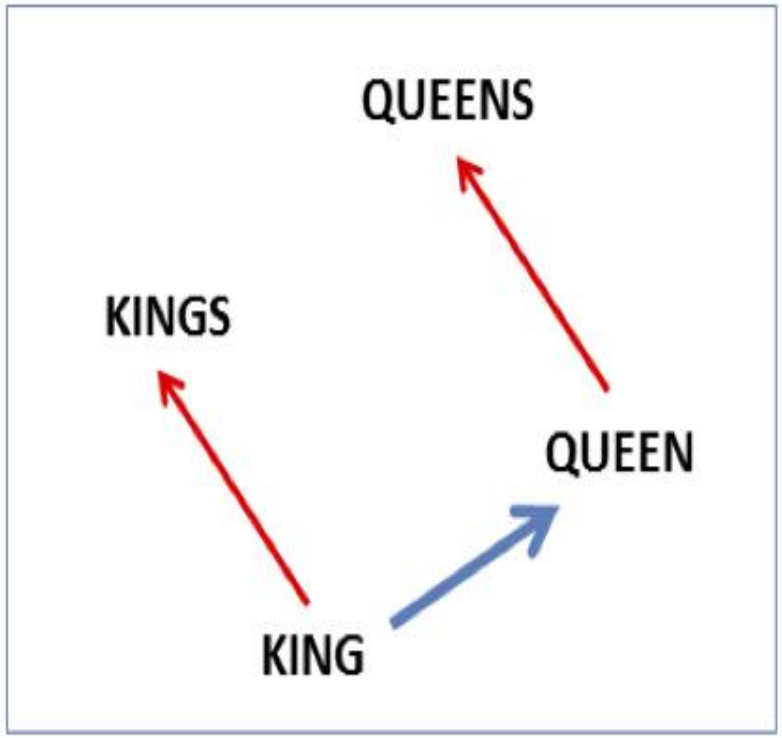
[-0.43337 0.10411 -0.49926 -0.06442 0.39005 0.65823 -1.3033 -0.39019 -0.41713 -0.038667 -0.37677 0.35303 -0.16425 -0.46671 -0.24358 0.10142 0.54664 0.065344 0.022733 -0.10422 -0.52839 0.20096 -0.23819 0.28856 0.098377 -0.12174 -0.035851 -0.92717 0.28408 0.22439 0.027598 0.47919 -0.51142 -0.062696 0.13002 0.20366 -0.050135 -0.4593 0.63779 -0.56294 -0.40989 -0.60342 0.8151 0.024929 -0.0096054 0.55748 0.35156 0.14476 -0.36327 -0.34856 0.1787 0.76398 0.27803 0.11999 0.42814 0.17844 0.092123 0.058811 -0.41114 -0.12101 -0.34001 -0.49946 -0.073549 0.15371 0.18034 0.34176 0.072738 -0.23442 -0.023682 -0.2499 -0.17334 -0.15008 -0.14599 0.51706 0.52797 -0.075328 0.13018 -0.069498 -0.15378 -0.31615 0.59434 -0.91396 -0.12803 0.32963 0.70337 -0.095882 -0.37066 0.16993 -0.62115 -0.76234 0.49005 -0.026823 -0.35171 -0.070227 0.19778 0.25563 -1.4504 -0.47122 -0.10107 -0.18279 -0.31553 0.090524 0.1975 0.073745 0.34809 -0.26728 -0.04808 -0.18467 0.18147 0.37255 0.26197 -0.0046708 0.51 -0.99408 -0.1942 -0.82518 0.59211 0.31112 0.3472 -0.066567 0.65975 -0.52254 -0.48302 0.30366 -0.35524 0.0022488 -0.89521 -0.096487 -0.36811 -0.1139 -0.039127 0.03701 0.18691 -0.28874 0.19926 -0.71229 0.21108 0.1768 0.27541 -0.72828 -0.74097 0.15007 -0.46696 0.52759 0.69806 0.28434 1.2781 0.033105 0.2153 -0.59069 -0.18089 -0.28775 -0.30792 -0.32764 -0.20838 -0.49774 0.2604 -0.26116 -0.29203 -0.18311 0.016024 0.26013 -0.4441 0.11857 0.61598 -0.25685 -0.49715 0.58277 -0.093157 -0.078187 -0.18587 -0.021307 0.52742 0.75704 0.091185 -0.41006 -0.26896 0.43715 0.13183 -0.49085 -0.84639 -0.22379 -0.094786 0.35858 -1.16 -0.019064 -0.29052 0.21588 0.026218 0.22063 -0.64061 0.89117 -0.14541 -0.47563 0.77044 -0.45668 0.49585 0.45303 -0.24904 0.24502 0.42608 0.0077214 -0.55742 0.17449 -0.2142 0.26996 -0.26239 0.18933 -0.66798 -0.004951 0.062785 0.45616 -0.77372 0.29266 -0.76515 0.2079 -0.52916 -0.13621 -0.60588 -0.049171 -0.21234 -0.071004 0.092045 0.87973 -1.0929 -0.028515 -0.28424 0.26105 0.2524 0.35996 -0.74328 -0.27066 -0.046789 0.081494 0.70996 0.18443 -0.10652 -0.32352 -0.026315 0.079707 0.14203 -0.21906 0.49254 -0.29718 0.25746 -1.108 -0.35566 -0.14188 -0.35727 -0.34243 -0.37111 1.0034 -0.13679 0.92309 -0.16716 -0.36536 0.019904 -0.33768 0.18646 0.4391 0.045023 -0.25726 -0.14073 0.77022 -0.18926 -0.27791 0.2978 -0.78997 0.052217 0.72518 0.0089245 0.1029 1.0213 -0.70057 0.31295 0.29313 -0.53556 0.75132 0.29351 -0.70482 -0.61882 -0.33732 0.60293 -0.33575 -0.12536 0.27659 -0.20361 0.12683 0.10469 -0.47956 0.187 0.38118 0.16238 -0.0484 0.43112 0.0089624 0.0051162 -0.67922 0.1709 -0.020472]

[Glove 300d vector]

Semantic Compositionality

- vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('*queen*')





[Mikolov 2013]

Demos

- https://playground.tensorflow.org/
- https://p.migdal.pl/interactive-machine-learning-list/
- https://demo.allennlp.org/reading-comprehension
- https://allenai.org/demos/
- http://text-processing.com/demo/
- https://explosion.ai/demos/
- https://huggingface.co/hmtl/
- https://corenlp.run/
- http://nlp.stanford.edu:8080/corenlp/
- https://talktotransformer.com/
- https://cs.stanford.edu/people/karpathy/convnetjs/

Epoch
000,000

Learning rate: 0.03
Activation: Sigmoid
Regularization: None
Regularization rate: 0
Problem type: Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 80%

Noise: 0

Batch size: 5

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1
X_2
X_1^2
X_2^2
X_1X_2
$\sin(X_1)$
$\sin(X_2)$

3 HIDDEN LAYERS

5 neurons | 3 neurons | 2 neurons

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

*This is the output from one **neuron**. Hover to see it larger.*

OUTPUT

Test loss 0.501
Training loss 0.500

Colors shows data, neuron and weight values.

-1 0 1

Show test data | Discretize output

Epoch
000,785

Learning rate: 0.03

Activation: Sigmoid

Regularization: None

Regularization rate: 0

Problem type: Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 80%

Noise: 0

Batch size: 5

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1

X_2

X_1^2

X_2^2

X_1X_2

$\sin(X_1)$

$\sin(X_2)$

3 HIDDEN LAYERS

5 neurons

3 neurons

2 neurons

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

*This is the output from one **neuron**. Hover to see it larger.*

OUTPUT

Test loss 0.272

Training loss 0.217

Colors shows data, neuron and weight values.

-1 0 1

Show test data

Discretize output

Epoch
023,073

Learning rate: 0.03

Activation: Sigmoid

Regularization: None

Regularization rate: 0

Problem type: Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 80%

Noise: 0

Batch size: 5

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1

X_2

X_1^2

X_2^2

X_1X_2

$\sin(X_1)$

$\sin(X_2)$

3 HIDDEN LAYERS

5 neurons

3 neurons

2 neurons

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

*This is the output from one **neuron**. Hover to see it larger.*

OUTPUT

Test loss 0.000

Training loss 0.003

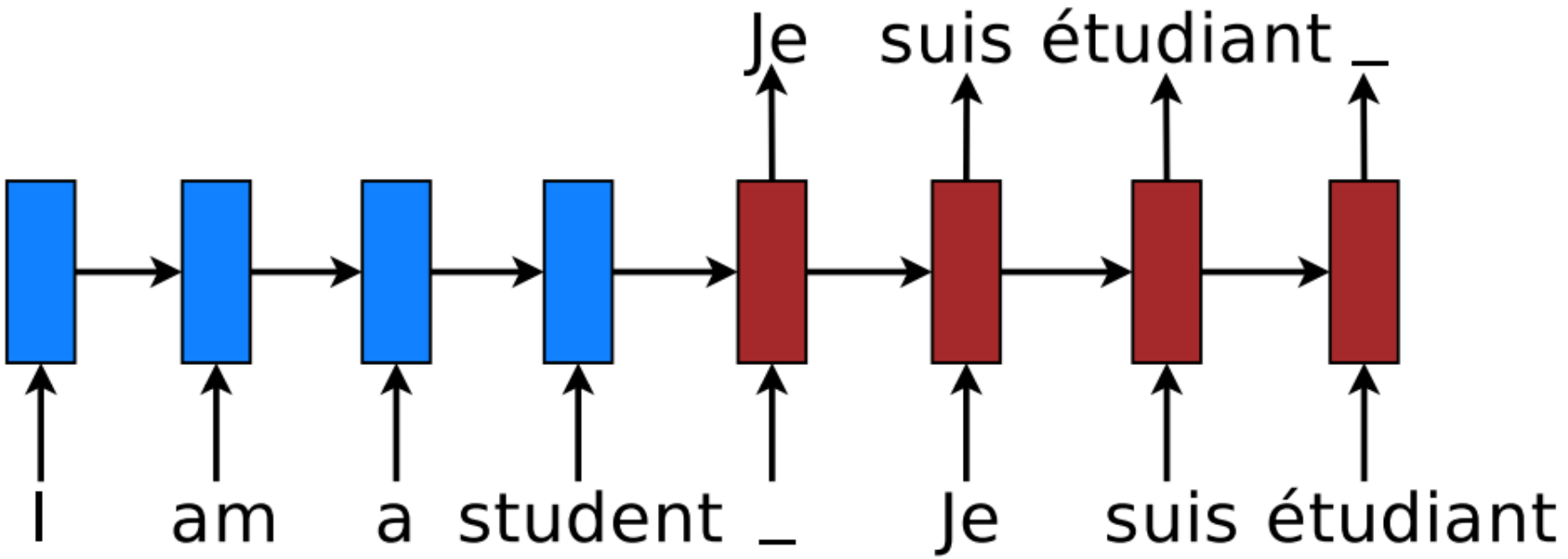
Colors shows data, neuron and weight values.

-1 0 1

Show test data

Discretize output

Sequence to Sequence



[Sutskever, Vinyals, and Le, 2014]

NLP

Introduction to NLP

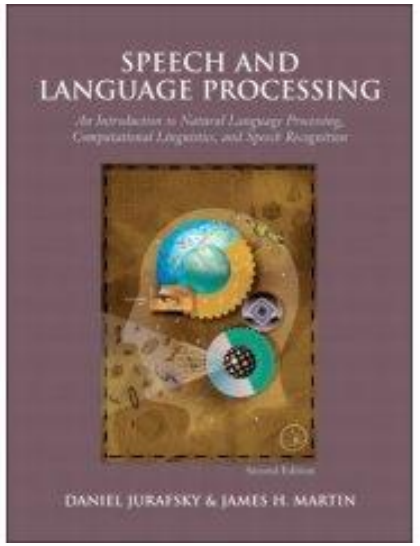
Class Logistics

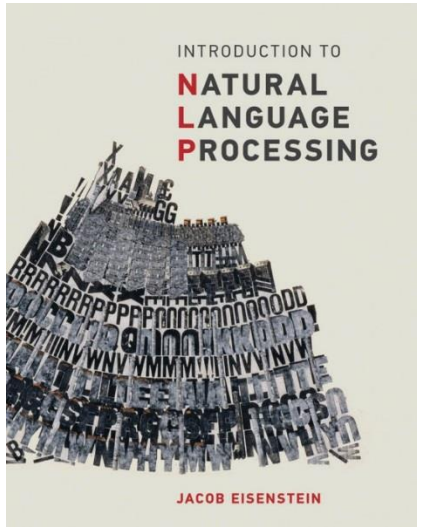
CPSC 477/577

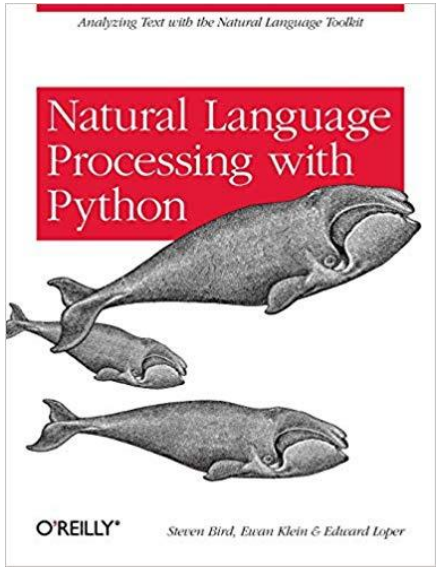
- Instructor:
  - Dragomir Radev
  - dragomir.radev@yale.edu
- Class times:
  - TTh 1-2:15
  - Location: Zoom
- Teaching staff:
  - Aditya Chander, Evan Cudone, Aarohi Srivastava, Michael Linden
- Office hours
  - Drago: F 1-2:15 (same Zoom link) or by appointment via email
  - Others: TBA

Course Readings

- Speech and Language Processing
  - Daniel Jurafsky and James Martin
  - Third edition, 2019
  - http://web.stanford.edu/~jurafsky/slp3/
- Introduction to Natural Language Processing
  - Jacob Eisenstein
  - First edition, 2019
  - https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- Additional readings:
  - Natural Language Processing using NLTK (Bird et al.) http://www.nltk.org
  - AAN http://aan.how







Course Dates

Feb	**2 4**	**9 11**	**16 18**	**23 25**	
Mar	**2 4**	**÷ 11**	**16 18**	**23 25**	**30**
Apr	**1**	**6 .**	**13 15**	**20 22**	**27 29**
May	**4 6**				

- **No class** on March 9 (Tue) and April 8 (Thu)
- Official "middle of the term" date
  - Friday March 30
- Midterm
  - TBA (not necessarily on March 30)
- Final exam
  - TBA (during the week of May 14 – May 19)

https://yalecollege.yale.edu/academics/academic-calendar/2020-2021-academic-calendar

Structure of the Course

- Background
  - linguistic, mathematical, and computational
- Computational models
  - morphology, syntax, semantics, discourse, pragmatics
- Core NLP technology
  - parsing, part of speech tagging, text generation, semantic analysis
- Applications
  - text classification, sentiment analysis, text summarization, question answering, machine translation, etc.
- Neural Networks and Deep Learning
  - distributed semantics, sequence to sequence methods, attention, transformers

Major Goals of the Class

- **Learn** the basic principles and theoretical issues underlying natural language processing
- **Understand** how to view textual data from a linguistic and computational perspective
- **Appreciate** the complexity of language and the corresponding difficulty in building NLP systems
- **Learn** techniques and tools used to develop practical, robust systems that can understand text and communicate with users in one or more languages
- **Understand the limitations** of these techniques and tools
- Gain insight into some **open research problems** in natural language processing


Start of semester
End of semester

Draft Syllabus

Introduction	Question Answering
Language Modeling	Summarization
Part-of-Speech Tagging	Dialogue and Conversational Agents
Hidden Markov Models	Machine Translation
Formal Grammars of English	Sentiment Analysis
Syntactic Parsing	Vector Semantics
Statistical Parsing	Dimensionality Reduction
Features and Unification	Word Embeddings
Dependency Parsing	Neural Networks
The Representation of Meaning	Attention
Computational Semantics	Transformers
Lexical Semantics	Recent Developments

Linguistic Knowledge

- Constituents
  - Children eat pizza.
  - They eat pizza.
  - My cousin's neighbor's children eat pizza.
  - Eat pizza!
- Collocations
  - Strong beer but *powerful beer
  - Big sister but *large sister
  - Stocks rise but ?stocks ascend
- How to get this knowledge in the system
  - Manual rules (?)
  - Automatically acquired from large text collections (corpora)

Areas of Linguistics

- Phonetics and phonology
  - the study of sounds
- Morphology
  - the study of word components
- Syntax
  - the study of sentence and phrase structure
- Lexical semantics
  - the study of the meanings of words
- Compositional semantics
  - how to combine words
- Pragmatics
  - how to accomplish goals
- Discourse conventions
  - how to deal with units larger than utterances

Mathematical Background

- Linear algebra
  - vectors and matrices
- Probabilities
  - Bayes theorem
- Calculus
  - derivatives
- Optimization
- Numerical methods

Math Background Links

- Matrix multiplication
  - https://www.intmath.com/matrices-determinants/matrix-multiplication-examples.php
- Bayes theorem
  - https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/
- Derivative of the sigmoid function
  - https://beckernick.github.io/sigmoid-derivative-neural-network/

Theoretical Computer Science

- Automata
  - Deterministic and non-deterministic finite-state automata
  - Push-down automata
- Grammars
  - Regular grammars
  - Context-free grammars
  - Context-sensitive grammars
- Complexity
- Algorithms
  - Dynamic programming

Artificial Intelligence

- Logic
  - First-order logic
- Agents
  - Speech acts
- Search
  - Planning
  - Constraint satisfaction
- Machine learning
  - Neural Networks
  - Reinforcement Learning

Grading

- Assignments (50%)
  - HW0+HW1 = 2+8=10%
  - HW2 = 10%
  - HW3 = 10%
  - HW4 = 10%
  - HW5 = 10%
- Exams (45%)
  - midterm = 20%
  - final exam = 25%
- Class participation (5%)
  - In-class participation, asking questions on Piazza, answering questions, office hours

How to get the most out of the class?

- Attend the lectures and study the slides
  - Course syllabus + slides = road map
  - Some material may not be found in any of the readings
- Hands on experience
  - Implement what you've learned
- Ask questions in and after class

Questions?

- Use the right channel for communication
  - Piazza/Canvas
- In special cases (e.g., sickness, regrading), use email
  - Include [CPSC477] or [CPSC577] or [NLP Class] in the subject line
- Office Hours:
  - TBA

NLP Courses at Other Places

- Brick-and-Mortar
  - Stanford (Chris Manning, Dan Jurafsky, Richard Socher, Chris Potts)
  - Texas (Greg Durrett)
  - CMU (Graham Neubig)
  - Johns Hopkins (Jason Eisner)
  - UNC (Mohit Bansal)
  - Utah (Vivek Srikumar)
- Online
  - Manning/Jurafsky (2012, survey)
  - Michael Collins (2013, more advanced)
  - Radev (2015-2016, survey)
  - New Coursera

The Association for Computational Linguistics (ACL)

Association for
Computational Linguistics

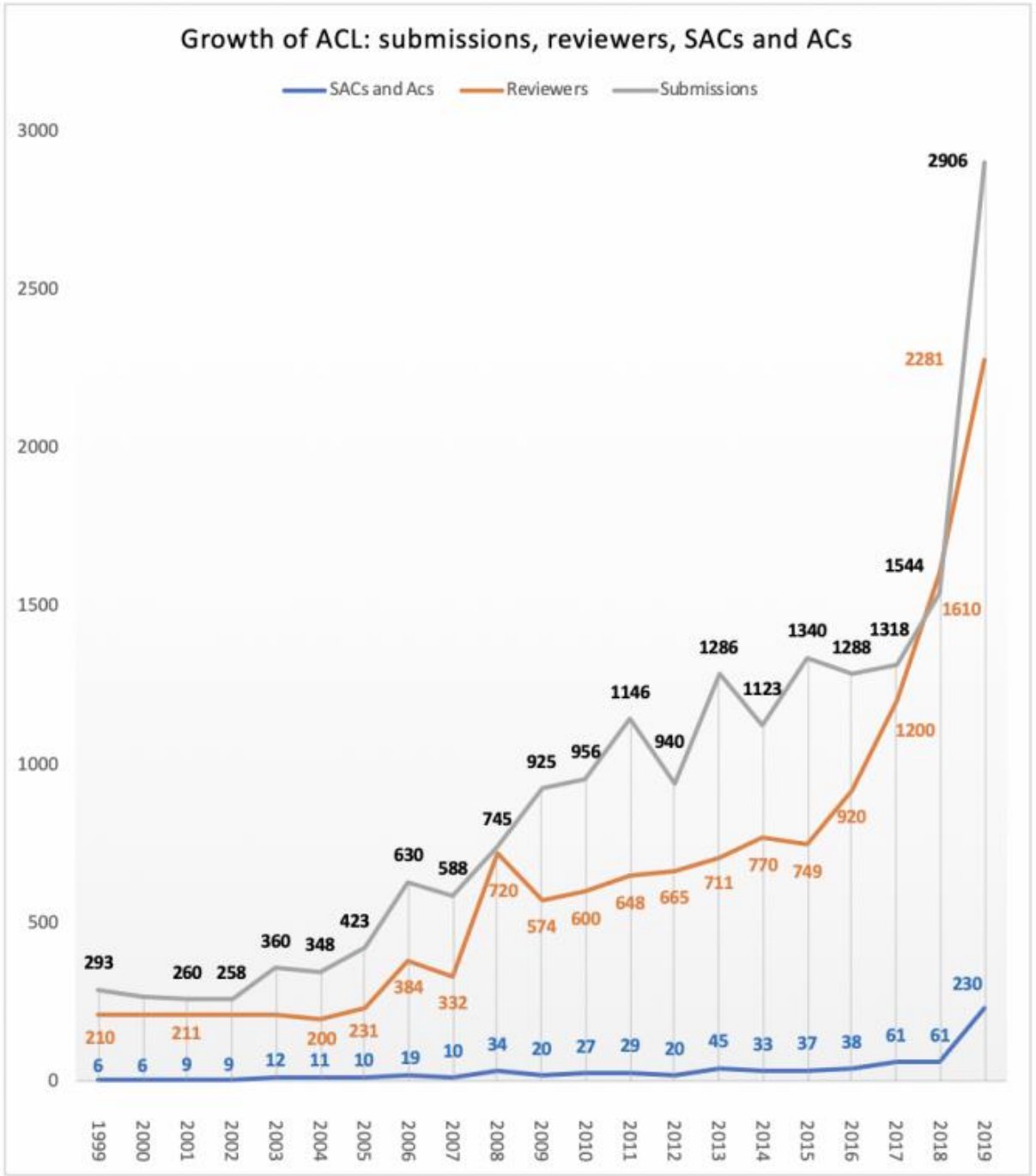
ACL 2020 Election Results

November 05, 2020 | BY webmaster

I am happy to announce the results of the elections for members of the ACL Executive Committee:

- Iryna Gurevych (Technical University (TU) of Darmstadt) has been elected as the VP-Elect.

- Yusuke Miyao (the University of Tokyo) has been elected as the Member at large.



https://aclweb.org/aclwiki/Conference_acceptance_rates

Research in NLP

- Conferences:
  - ACL, NAACL, EMNLP, SIGIR, AAAI/IJCAI, COLING, EACL, Interspeech, NeurIPS, ICLR, SIGDIAL
- Journals:
  - Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM Transactions on Information Systems, ACM TALIP, ACM TSLP
- University centers:
  - Stanford, Berkeley, Columbia, CMU, JHU, Brown, UMass, MIT, UPenn, Illinois, Michigan, Yale, Washington, Maryland, NYU, UNC, OSU, GA Tech, Princeton, etc.
  - Toronto, Edinburgh, Cambridge, Sheffield, Saarland, Trento, Prague, QCRI, NUS, and many others
- Industrial research sites:
  - Google, Facebook, MSR, IBM, SRI, BBN, MITRE, Baidu, Salesforce
- The ACL Anthology
  - http://www.aclweb.org/anthology
- The ACL Anthology Network (AAN)
  - http://aan.how

Students with Disabilities

- If you think you need an accommodation for a disability, please let me know at your earliest convenience.
- Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress.
- I will treat any information that you provide in as confidential a manner as possible.

Student Mental Health and Wellbeing

- Yale University is committed to advancing the mental health and wellbeing of its students.
- If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. Yale Counseling: **203-432-0290, 203-432-0123** (after hours)

Course Design Choices

- A wide-coverage survey course
- A mixture of traditional and neural techniques
- A non-trivial focus on linguistic issues
- A mixture of programming and written assignments
- Significant readings
- External links (tutorials)
- Fairly independent assignments

Programming environment

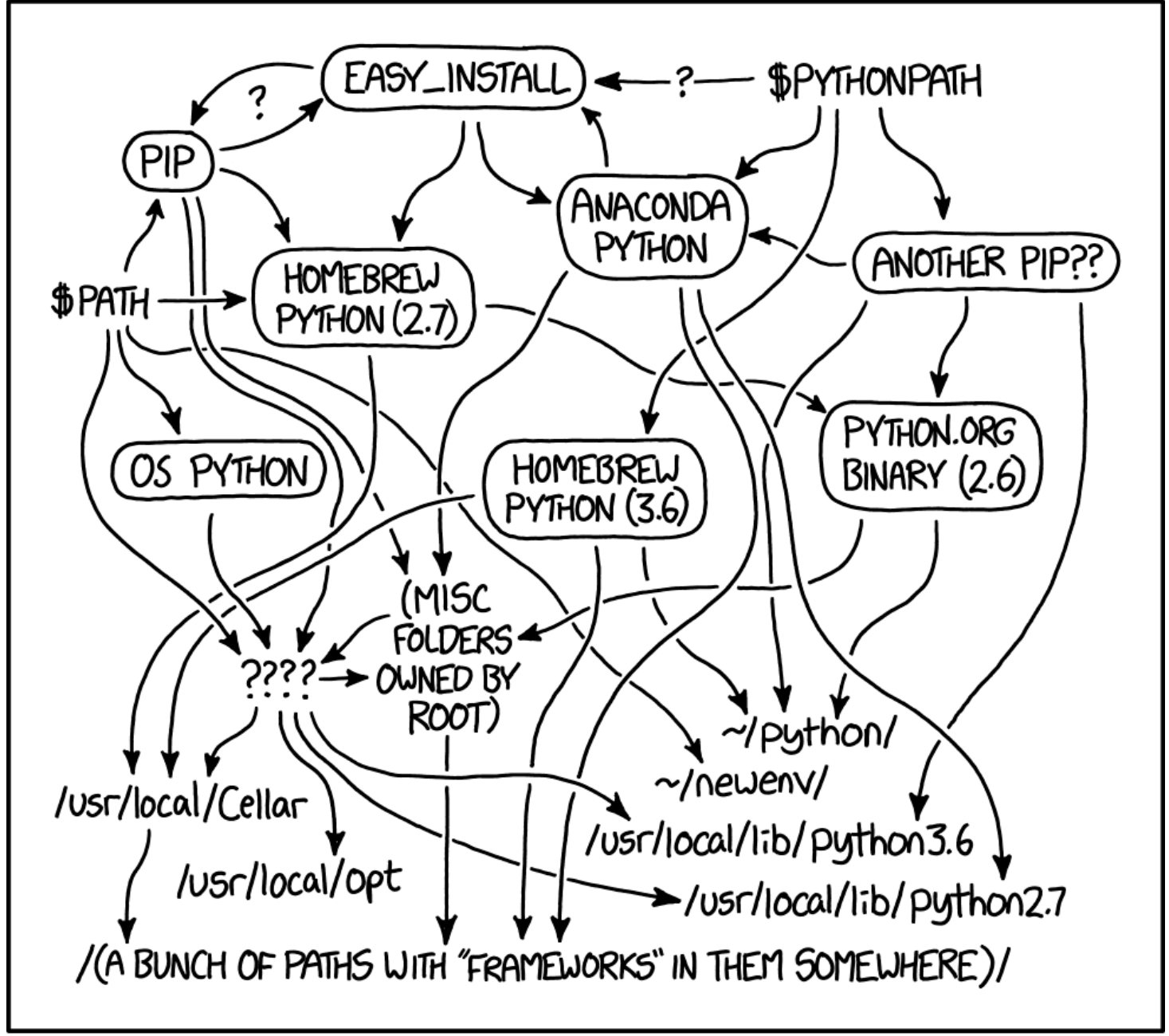
- python in a UNIX environment
- pytorch
- text manipulation
- scipy
- sklearn

Sample Programming Assignments

- Language Modeling and Part of Speech Tagging
- Dependency Parsing
- Vector Semantics and Word Sense Disambiguation
- Question Answering
- Deep Learning
- Machine Translation
- Sentiment Analysis
- Natural Language Interface to a Database
- Semantic Parsing

Programming Language

- The programming assignments will be in Python.
- You are expected to either know Python already or to learn it on your own.
- We will be using pytorch for most assignments.
- The code base will be installed on the **Zoo** machines.



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Submitting Assignments

- In the absence of a prior emailed authorization from the instructor, you should turn in your assignments electronically by 11:59:59 PM on the due date. For each day (or fraction of a day) that your submission is late, it will be penalized 10%, for a maximum of 30%. After three days, the assignment will be given a score of zero.
- You will need to hand in the source code for the project, relevant documentation, and a script of a test run of your program to show that it actually works on the Zoo machines.

Academic Honesty

- Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided.
- Any violation of the University's policy on Academic and Professional Integrity will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program.
- Violations of academic and professional integrity will be reported to Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed.

Integrity Policies

- Collaboration policy:
  - You may discuss the course material and the textbook with other students. You may also discuss the *requirements* of the assignments. However, you cannot get help with the assignments and exams themselves in oral or written form from anyone. If you are unsure about this policy, ask the instructors.
- Honesty policy:
  - We will be using high grade plagiarism detection code
  - Do not copy other people's code or misrepresent it as yours, period.

Specifics

- Coding and write up should be done independently
- Do not show your work to anyone
- Do not look at anyone’s work
- Do not use existing web code (e.g., github) that is specifically designed to solve the problem in the assignment. If you use other github code, attribute it properly in your submission

Grading Appeals

- If you have a question about your grade on a particular assignment (or exam), write a short email to the TA in charge of that assignment.
- Please submit any such requests within a week of receiving your grade.

NLP

Natural Language Processing

112

Examples of Text

What NLP is not about

- Romeo loves Juliet.
- ZZZZZ is a great stock to buy.

What it is about (1/2)

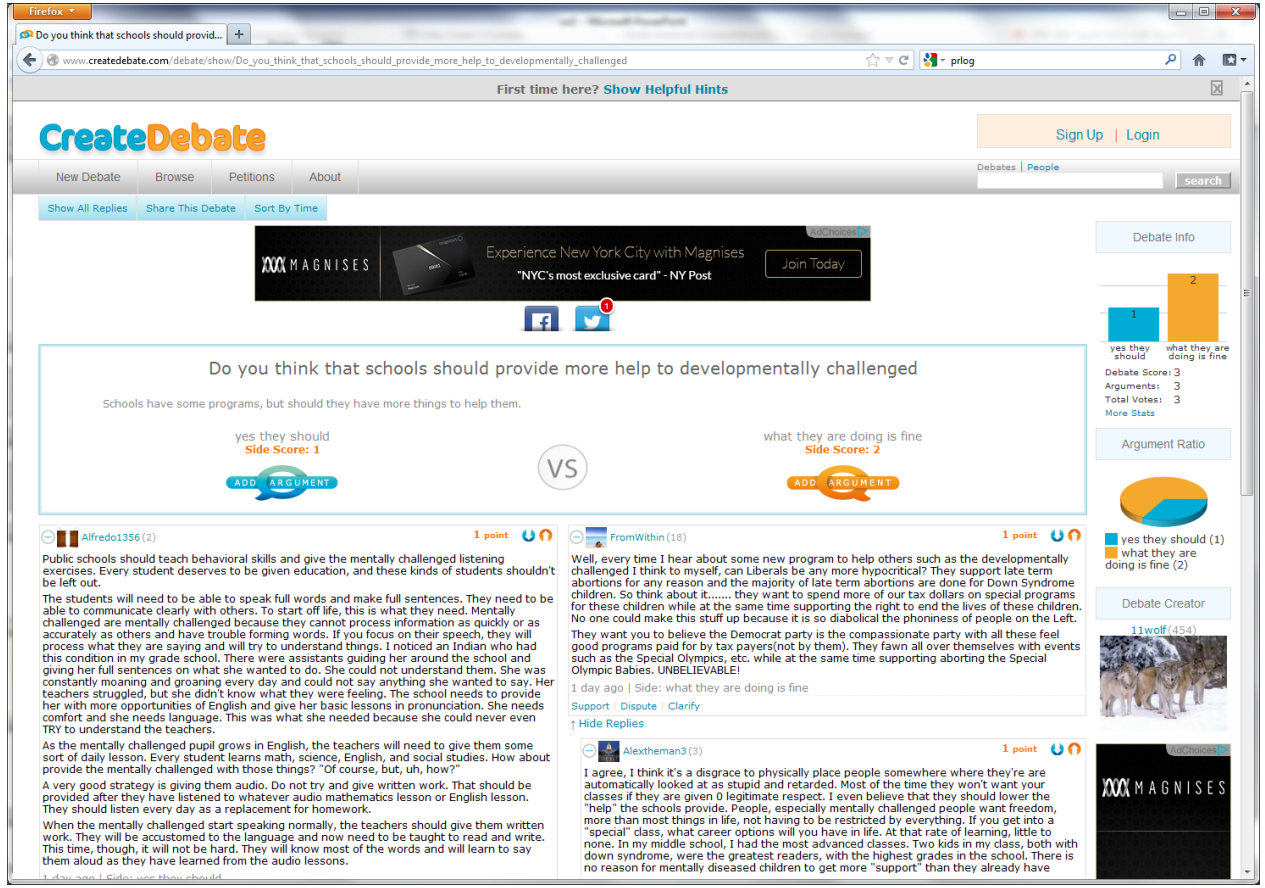
- After the ball, in what is now called the "balcony scene", Romeo sneaks into the Capulet orchard and overhears Juliet at her window vowing her love to him in spite of her family's hatred of the Montagues.

What it is about (2/2)

- **ZZZZZ Resources** (NYSE:ZZZZZ) in their third quarter financials present a picture of a company with a relatively high amount of debt versus shareholder equity, and versus revenues. The company had total liabilities in the third quarter of $4,416 million versus shareholders' equity of only $1,518 million. That is a very high 3 to 1 debt to equity ratio. The company had third quarter revenues of $306 million. On an annualized basis, revenues would come out to $1,224 million. The company's debt level is almost 3 times its annual revenues. And remember, third quarter revenue is from before oil prices dropped in half. It looks like ZZZZZ may have bitten off more than it can chew.
- **XXXXX Petroleum** (NYSE:XXXXX) is another company whose third quarter financials present a relatively high debt load. The company had total liabilities in the third quarter of $3,272 million versus shareholder equity of only $1,520 million. That represents a high 2 to 1 debt to equity ratio. The company had third quarter revenues of $350 million. On an annualized basis revenues would come out to $1,400 million. The company's debt is more than 2 times its annual revenue. While XXXXX is a very good operator, it looks like they have taken on the high debt strategy at the wrong time.
- **YYYYY Energy** (NYSE:YYYYY) has a relatively high debt load according to their third quarter financials. The company had total liabilities of $2,026 million versus shareholder equity of $1,079. That is almost a 2 to 1 debt to equity ratio. Their third quarter revenues were $207 million. When annualized, their third quarter revenues come out to $827 million. The company's debt is almost 2 1/2 times its annualized revenues, and that is before the collapse of oil prices in the fourth quarter. YYYYY has taken the Brigham model to heart and has been aggressively growing the company.

Genres of Text

- Blogs, emails, press releases, chats, debates, etc.
- Each presents different challenges to NLP



Re: essence documentation - Message (Plain Text)

File Edit View Insert Format Tools Actions Help

Reply | Reply to All | Forward

Extra line breaks in this message were removed.

From: radev@si.umich.edu on behalf of Ali Hakim [ahakim@mac.com] Sent: Wed 8/10/2005 1:29 PM
To: sam
Cc: radev@umich.edu; Rohit Laungani
Subject: Re: essence documentation

OK. I have added Essence.tex to the doc directory in the CLAIR project.

Can someone tell me how to set up my login shell so I don't have to export the CVS root every time or use the full path to CVS? Thanks.

Ali

On Aug 10, 2005, at 3:05 PM, sam wrote:

> In message <20050810185818.D11E0B8491@tangra.si.umich.edu>,
> radev@umich.edu writes:
>
>> sam wrote:
>>
>>>
>>> I will try to make a docs directory in the CVS repository.
>>> (I did this once. I think I just used cvs add with a directory just
>>> as I would have for a file....)
>>>
>>> Should it be parallel to the CLAIR directory or below it?
>>>
>>
>> Below it.
>>
>
>
> OK, it's there now.
>

CNN.com - Turmoil in Peru as PM quits - Aug 11, 2005 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.cnn.com/2005/WORLD/americas/08/11/peru.toledo.reut/index.html

Firefox prevented this site from opening a popup window. Click here for options...

CNN.com WITH FREE VIDEO | International Edition | Netscape | Member Services | MAKE CNN.com YOUR HOME PAGE

SEARCH THE WEB CNN.com SEARCH Powered by YAHOO! search

Home Page | World | U.S. | Weather | Business at CNNMoney | Sports at SI.com | Politics | Law | Technology | Science & Space | Health | Entertainment | Travel | Education | Special Reports | Video NOW FREE | Autos with Edmunds.com

University of Phoenix Online — To land the job you want.

SERVICES: E-mail Newsletters | Your E-mail Alerts | RSS | CNNtoGO | Contact Us

SEARCH: WEB CNN.com SEARCH Powered by YAHOO! search

WORLD

Turmoil in Peru as PM quits

Thursday, August 11, 2005; Posted: 8:01 p.m. EDT (00:01 GMT)

LIMA, Peru (Reuters) -- Peruvian President Alejandro Toledo said Thursday he had asked all his ministers to tender their resignations and would evaluate who would stay on in their jobs.

The government was plunged into crisis earlier when Prime Minister Carlos Ferrero quit, following the appointment of Toledo's controversial top ally, Fernando Olivera, as foreign minister.

Housing and Construction Minister Carlos Bruce also resigned immediately.

Bruce and Ferrero both publicly split with Olivera over legalizing some cultivation of the raw material for cocaine, and their resignations signal more turbulence ahead for the unpopular Toledo in his final months in office.

Under Peru's constitution, once a prime minister resigns all ministers must tender their resignations.

"Carlos Ferrero has tendered his irrevocable resignation as prime minister," a statement from his office said.

Ferrero, a veteran lawmaker who has been prime minister since December 2003, would have had to quit the Cabinet by early October if he wanted to run for a congressional seat, but the timing of his announcement -- minutes after Toledo's closest ally was sworn in as foreign minister -- was a surprise.

President Alejandro Toledo has a low public approval rating.

Search Jobs MORE OPTIONS — Enter Keywords — Enter City — ALL — SEARCH — careerbuilder.com

Want to know if a thief is using your credit card?

advertiser links — what's this?

Compare August Mortgage Rates
Get $150,000 loan for $720 per month. Refinance while rates are low.
www.lowermybills.com

MyCashNow - $100 - $1,500 Overnight
Payday Loan Cash goes in your account overnight. Very low fees. Fast decisions....
www.mycashnow.com

Comcast High-Speed Internet
Order today for a $19.99/mo. special, free modem, plus get $75 cash back when...
www.comcastoffers.com

Florist.com Save 10% - Fresh Flowers

Done

Crooked Timber - Mozilla Thunderbird

File Edit View Go Message Tools Help

Get Mail | Write | Address Book | Reply | Reply All | Forward | Delete | Junk | Print | Stop

Folders — View: All — Subject Or Sender

- News & Blogs
  - Trash
  - ann arbor is overrated
  - Crooked Timber
  - Dilbert
  - Ernie's 3D Pancakes
  - Geeking with Greg (3)
  - Hit and Run (160)
  - Interesting People (200)
  - Salon.com (324)
  - Slate Magazine (402)
  - The Long Tail (17)
  - The Technor...eblog (17)
- Local Folders
  - Unsent Messages
  - Drafts
  - Sent
  - Trash

Subject	Sender	Date
Declare, if thou hast understanding	Henry	8/3/2005 1:09 PM
"Critical literacy" to be scrapped in Queensland schools	John Quiggin	8/3/2005 10:36 PM
Spreading Statistics, cont.	Ted	8/4/2005 9:06 AM
Unpleasantly self-absorbed suicide bombers	Chris Bertram	8/5/2005 3:32 AM
APSA moves	Henry	8/5/2005 3:42 PM
Jimmy Doyle on human agency	Chris Bertram	8/6/2005 1:07 AM
Cheezeball	Chris Bertram	8/6/2005 5:32 AM

Subject: Spreading Statistics, cont.
From: Ted
Date: 8/4/2005 9:06 AM
Website: http://crookedtimber.org/2005/08/04/spreading-statistics-cont/

I noted a few days ago that Senator Rick Santorum made a claim in an online interview about federal taxation. Senator Santorum said that the federal tax rate for the average family has gone up from 2% (in 1950) to 27% today. Furthermore, he claimed that income from a second worker simply replaces the money that the family pays in increased federal taxes. They would enjoy the same net income if taxes went back to 1950 levels and the second worker stayed at home.

I'm really rather sure that this isn't true. I'm relying on the Tax Policy Center: They say that federal taxes on a family of four at the median income have gone up from about 7.4% to about 14.4%, and that the family would have saved $4436 if we could roll back tax rates. That doesn't correspond to the Senator's story.

I checked last night, and Santorum repeats this point in his book, *It Takes a Family*. It's on page 123 and 124, and there's no source. (There's a bibliography of sorts, but it just lists a series of sources used in each section. There's no way to connect any specific point to any source.) When I called his press office again to ask for a source, they referred me to the publisher, who couldn't help me. Nonetheless, he's repeated this claim at least two more times, on Hardball with Chris Matthews and on Fox News.

Shouldn't the Senator care whether what he's saying is right or wrong? Wouldn't it be nice if a journalist asked him about it?

(Incidentally, is there anything more depressing than the "Current Events" section of a modern-day bookstore? There are so many rows of hastily-written, 200-250 page books with giant print, huge margins, and a cover featuring a smug bastard under a title like "THEY'RE ALL AGAINST YOU: How Hollywood, the French, and the CIA Have Conspired to Pollute Your Precious Bodily Fluids and **What You Can Do To Stop Them**." Robert Bork's *Slouching Towards Gomorrah* looks like Winston Churchill in all that dreck.)

Unread: 0 Total: 118

Dell 2Q Profit Rises, Revenue Misses View - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://biz.yahoo.com/ap/050811/earns_dell.html?.v=7

Yahoo! My Yahoo! Mail — Search the Web — Search

YAHOO! FINANCE Sign In New User? Sign Up — Finance Home - Help — Associated Press

Welcome [Sign In] — To track stocks & more, Register

Financial News

Enter symbol(s) — Basic — Get — Symbol Lookup

AP

Dell 2Q Profit Rises, Revenue Misses View

Thursday August 11, 7:01 pm ET
By Matt Slagle, AP Technology Writer

Dell Second-Quarter Profit Rises 28 Percent, but Revenue Misses Wall Street Expectations

DALLAS (AP) -- Increased shipments boosted Dell Inc.'s second-quarter by 28 percent but revenue fell below Wall Street expectations as the company blamed overly aggressive pricing of its low-end desktop and notebook computers.

Revenue at the world's largest PC maker rose 15 percent to $13.43 billion from $11.71 billion last year.

The results reported Thursday matched analyst forecasts of 38 cents per share but fell below projected revenue of $13.71 billion, according to Thomson Financial.

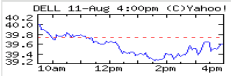
Dell shares, which fell 15 cents to close at $39.58 on the Nasdaq Stock Market, tumbled 7.7 percent, or $3.03, to $36.55 in late trading. Dell released the profit report after the market close.

Net income climbed to $1.02 billion, or 41 cents per share, from $799 million, or 31 cents per share, during the year-ago period.

The Round Rock, Texas-based company said its results included a 3-cent tax benefit, without which it would have earned 38 cents per share in the latest quarter.

Chief Executive Kevin Rollins blamed the revenue shortfall on lower prices on its low-end desktops and laptops.

Related Quote



DELL 11-Aug 4:00pm (C)Yahoo!

DELL 39.58 -0.15 News

View Detailed Quote
Delayed 20 mins
Providers - Disclaimer

Related News Stories

- Stocks to Watch: Dell, Red Robin Gourmet, Nvidia, Jamdat Mobile - at MarketWatch (8:09 pm)
- [$$] Dell's Net Rises on Overseas Strength - at The Wall Street Journal Online (8:03 pm)
- Disappointing revenues at Del - at FT.com (7:00 pm)
- Dell Takes Hit After Its Sales Lag Estimates - Investor's Business Daily (7:00 pm)

More...

- By industry: Computer hardware, Computers

Top Stories

- Dow Rises 91, Nasdaq Up 17 on Late Rally - AP (5:44 pm)
- Ex-WorldCom CFO Gets Five Years in Prison - AP (5:51 pm)
- July Retail Sales Climb on Auto Purchases - AP (5:09 pm)

Done

Google Press Center: Press Release - Mozilla Firefox

http://www.google.com/press/pressrel/univision.html

Google Press Center

Home
About Google
Press Center

Media Resources
Press Releases
Images and B-roll
Awards
Permissions

Products & Technology
Product Descriptions
Reviewer's Guides

Related Product Info
Google Labs
Software Principles

Related Corporate Info
History
Timeline
Executive Bios
Governance
Privacy Policy

Investor Relations

Find on this site:
Search

Univision.Com And Google Partner To Bring Search Services to Nation's Most Visited Spanish-Language Website

MOUNTAIN VIEW, Calif. and NEW YORK – July 28, 2005 – Univision Online, Inc., the interactive division of Univision Communications Inc. (NYSE: UVN), today announced that it has entered into multi-year partnership with Google Inc. (NASDAQ: GOOG). The partnership combines the power of the most visited Spanish-language website with an industry leader in search technology and advertising.

Through the partnership with Google, visitors to Univision.com will now receive more relevant, comprehensive, and timely Spanish-language search results optimized for quick delivery in a user-friendly environment. The relationship also connects Univision.com users with more products and services online through targeted advertisements on the search and content pages of the website. In addition, these ads increase the visibility of websites and advertisers serving Hispanics.

"As the most-visited Spanish-language website, Univision.com is excited to work with Google to bring the benefits of more targeted search capabilities to our users," said Javier Saralegui, President, Univision Online. "In addition, for those consumer companies not yet targeting Hispanics online, this partnership should provide additional impetus for them to develop websites in Spanish."

By combining the No.1 Spanish-language website with sophisticated and effective search technology, Univision.com and Google are taking the lead in ensuring that search becomes a more prominent functionality and effective targeting method in the Hispanic online marketplace.

"Google's agreement with the largest Spanish-language media company in the United States provides Univision's users with relevant information from search results and ads and gives our advertisers a new and targeted way to reach the Hispanic audience," said Tim Armstrong, Vice President, Advertising Sales, Google Inc.

About Univision Online
Univision Online, Inc. is the interactive division of Univision Communications Inc. (NYSE: UVN), the premier Spanish-language media company in the United States. Univision Online owns and operates the premier Spanish-language Internet destination in the U.S., located at www.univision.com, and is a wholly owned subsidiary of Univision Communications Inc. Its operations include Univision Network, the most-watched Spanish-language broadcast television network in the U.S. reaching 98% of U.S. Hispanic Households; TeleFutura Network, a general-interest Spanish-language broadcast television network, which was launched in 2002 and now reaches 85% of U.S. Hispanic Households; Univision Television Group, which owns and operates 27 Univision Network television stations and 1 non-Univision television station; TeleFutura Television Group, which owns and operates 33 TeleFutura Network television stations; Galavisión, the country's leading Spanish-language cable network; Univision Radio, the leading Spanish-language radio group which owns and/or operates 66 radio stations in 16 of the top 25 U.S. Hispanic markets and 4 stations in Puerto Rico; Univision Music Group, which includes Univision Records, Fonovisa Records, and a 50% interest in Mexico-based Disa Records labels as well as Fonomusic and America Musical Publishing companies. Univision Communications also has a 50% interest in TuTv, a joint venture formed to broadcast Televisa's pay television channels in the U.S. , and a non-voting 27% interest in Entravision Communications Corporation, a public Spanish-language media company. Univision Communications is headquartered in Los Angeles with television network operations in Miami and television and radio stations and sales offices in major cities throughout the United States.

For more information, please visit www.univision.net.

About Google Inc.
Google's innovative search technologies connect millions of people around the world with information every day. Founded in 1998 by Stanford Ph.D. students Larry Page and Sergey Brin, Google today is a top web property in all major global markets. Google's targeted advertising program provides businesses of all sizes with measurable results, while enhancing the overall web experience for users.

Done

2004_ibm_annual.pdf (application/pdf Object) - Mozilla Firefox

1 of 100

PTEK HOLDINGS INC (Form: 8-K, Received: 09/07/2004 17:03:57) - Mozilla Firefox

http://www.shareholder.com/common/edgar/880804/1193125-04-152871/04-00.html

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, DC 20549

FORM 8-K

CURRENT REPORT PURSUANT
TO SECTION 13 OR 15(D) OF THE
SECURITIES EXCHANGE ACT OF 1934

Date of report (Date of earliest event reported) September 7, 2004

PTEK HOLDINGS, INC.

(Exact Name of Registrant as Specified in Its Charter)

GEORGIA
(State or Other Jurisdiction of Incorporation)

000-27778	59-3074176
(Commission File Number)	(IRS Employer Identification No.)

3399 Peachtree Road, NE, Suite 700, Atlanta, Georgia 30326
(Address of Principal Executive Offices) (Zip Code)

404-262-8400
(Registrant's Telephone Number, Including Area Code)

(Former Name or Former Address, if Changed Since Last Report)

Check the appropriate box below if the Form 8-K filing is intended to simultaneously satisfy the filing obligation of the registrant under any of the following provisions (*see* General Instruction A.2. below):

☐ Written communications pursuant to Rule 425 under the Securities Act (17 CFR 230.425)

☐ Soliciting material pursuant to Rule 14a-12 under the Exchange Act (17 CFR 240.14a-12)

☐ Pre-commencement communications pursuant to Rule 14d-2(b) under the Exchange Act (17 CFR 240.14d-2(b))

☐ Pre-commencement communications pursuant to Rule 13e-4(c) under the Exchange Act (17 CFR 240.13e-4(c))

Done

Tony Blair - Wikipedia, the free encyclopedia - Mozilla Firefox

http://en.wikipedia.org/wiki/Tony_Blair

article | discussion | edit this page | history

Create account / log in

Wikimedia needs your help in its 21-day fund drive. See our fundraising page.
Over US$185,000 has been donated since the drive began on 19 August. Thank you for your generosity!

Tony Blair

From Wikipedia, the free encyclopedia.

The Right Honourable **Anthony Charles Lynton Blair** (born May 6, 1953 in Edinburgh, Scotland) is the current Prime Minister of the United Kingdom. He has led the Labour Party since July 1994, (following the death of John Smith in May of that year) and brought Labour into power with a landslide victory in the 1997 general election, replacing John Major as Prime Minister and ending 18 years of Conservative government. He is now the Labour Party's longest-serving Prime Minister, and the only person to have led the party to three consecutive general election victories, just as Margaret Thatcher was the only Conservative Prime Minister to win three consecutive general elections.

Blair moved the Labour Party towards the centre of British politics, using the term "New Labour" to distinguish what he identifies as "modern social democracy" and his party's refusal to reverse privatisation and support for a market economy from its past belief in nationalisation and Fabian socialism. However, critics on the left feel that he has compromised the principles of the founders of the Labour party, and that the Blair government has moved too far to the right, placing insufficient emphasis on traditional Labour priorities such as the redistribution of wealth.

Since the advent of the War on Terror, much of the Prime Minister's political agenda has been dominated by foreign affairs, particularly those concerning Iraq, and he has supported many aspects of the foreign policy of United States president George W. Bush, sending British forces to participate in the 2003 invasion of Iraq and the subsequent occupation and conflict. Blair's Labour party still managed to win an unprecedented third term in the 2005 general election (unprecedented for the Labour Party; the Conservatives having won three consecutive terms twice since the Second World War). Although Labour's majority in the House of Commons was reduced considerably to 67 MPs, this remains a substantial working majority and a measure of the relative political weakness and poor credibility of the main opposition party, the Conservatives. Some journalists immediately noticed that New Labour no longer has a majority independently of Old Labour, although this may discourage some MPs from mounting damaging rebellions against the leadership.

While Blair is in no danger of losing a potential vote of no confidence, the fall in the Labour vote (from 41% to 35%) has renewed speculation amongst commentators as to how long his leadership can continue. It is widely predicted that he will be succeeded by his ambitious Chancellor of the Exchequer Gordon Brown before the next General Election (which will occur at the latest in 2010).

Contents [hide]
1 Early and private life
2 Begins political career
3 In opposition
3.1 Leader of the Labour Party
4 First term 1997 to 2001
4.1 Establishment of the Third Way
4.2 Control over House of Commons
4.3 Domestic policies
5 Second term 2001 to 2005
5.1 Iraq war
5.2 Domestic politics
5.3 Attempted impeachment
5.4 Health problems
6 Third term 2005 to present
6.1 G8 and EU presidencies
6.2 2012 Summer Olympics
6.3 2005 London bombings
6.4 Departure
7 Caricature and satire of Blair
8 See also
9 References
10 Further reading

The Rt Hon. Tony Blair

Appointed PM	2 May 1997
PM predecessor	John Major
Date of birth	6 May 1953
Place of birth	Edinburgh, Scotland
Political Party	Labour
Constituency	Sedgefield

Early and private life [edit]

Blair was born in Edinburgh, Scotland. His father, Leo Blair, was a barrister and later a law lecturer who, having been a communist in his youth, was active in the Conservative Party. Leo Blair had had ambitions to stand for Parliament in Durham, which were thwarted when he had a stroke when his son was 11, an event which affected

navigation: Main Page, Community portal, Current events, Recent changes, Random article, Help, Contact us, Donations
search: Go, Search
toolbox: What links here, Related changes, Upload file, Special pages, Printable version, Permanent link
in other languages: Bosanski, Català, Česky, Cymraeg, Dansk, Deutsch, Español, Esperanto, Français, Gaeilge, Galego, 한국어, हिन्दी, Ido, Bahasa Indonesia, Italiano

Done

Induction of Influenza-Specific Mucosal Immunity by an Attenuated Recombinant Sendai Virus

Thuc-vy L. Le[1], Elena Mironova[2], Dominique Garcin[2], Richard W. Compans[1]*

1 Department of Microbiology and Immunology and Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia, United States of America,
2 Department of Microbiology and Molecular Medicine, University of Geneva School of Medicine, Geneva, Switzerland

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the *Paramyxoviridae* family, belongs in the genus respirovirus and shares 60–80% sequence homology to human parainfluenza virus type 1 (HPIV-1) [9,10]. The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPIV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16]. These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

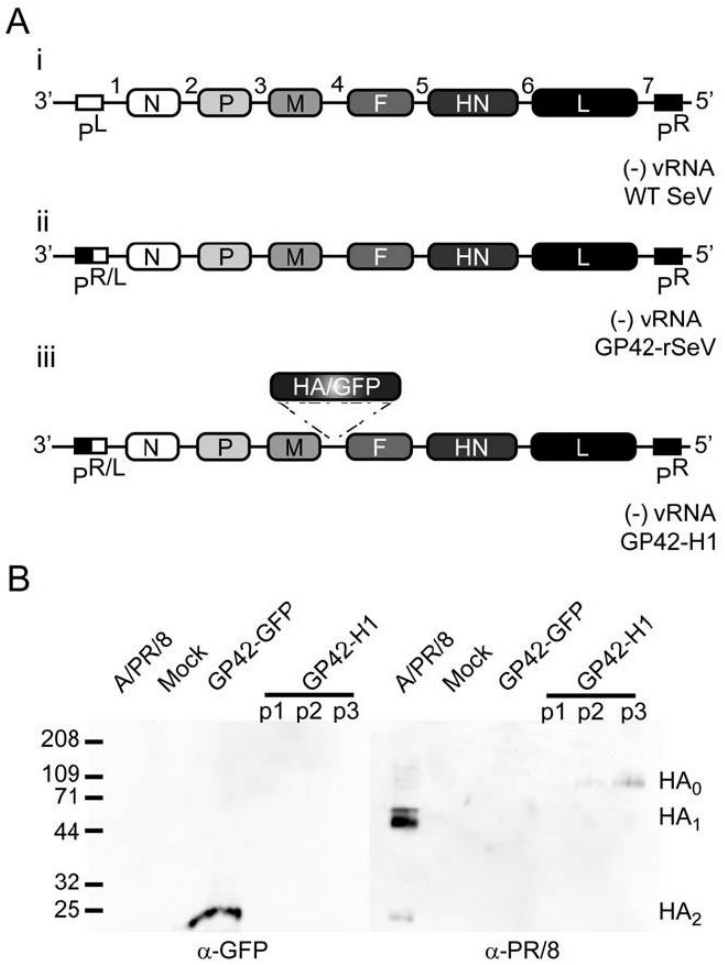


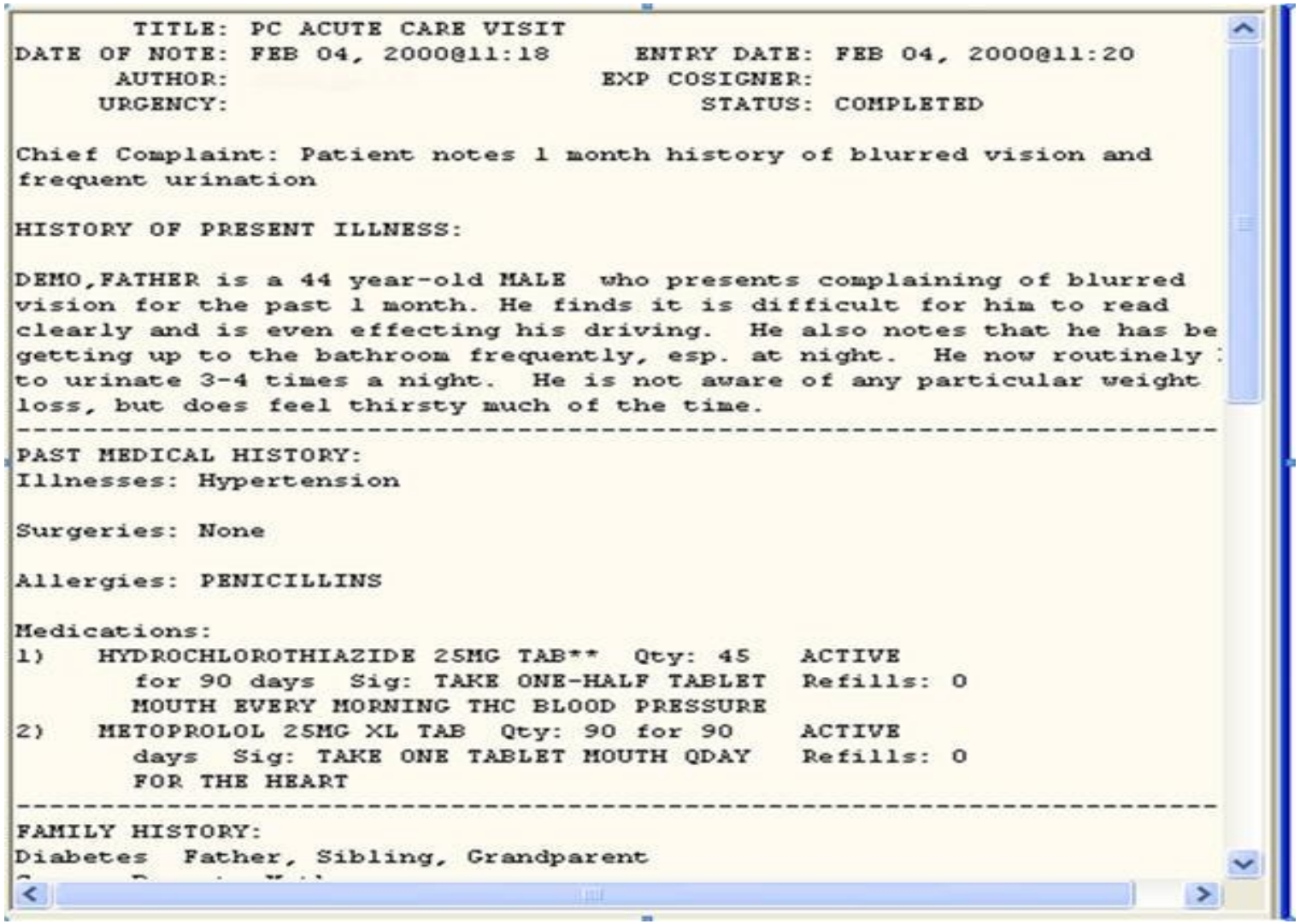
Figure 1. Generation of the GP42-H1 vector. A) Representation of wild-type Sendai virus gene construct (i) illustrating the major viral genes N, P, M, F, HN and L. The left promoter (P^L) and right promoter (P^R) function as the genomic promoter and anti-genomic promoter respectively. Seven gene boundaries that encode the conserved regulatory transcription start and transcription stop sequence are represented numerically. The mutant GP42-SeV (ii) genomic RNA is identical to WT Sendai virus with the exception of 3′ P^L in which 42 nucleotides of the P^L were replaced with the corresponding sequence from P^R ($P^{R/L}$). Additional transcription start, stop, poly-adenylation sequences and a unique Mlu I restriction site were introduced into the intergenic region between the Sendai M and F genes. Using the unique Mlu I restriction site, the GFP or HA gene was inserted (respecting the rule of six) generating the recombinant Sendai GP42-GFP [19] or GP42-H1 vectors (iii). B) Recombinant GP42-H1 virus was cultured in BSR-T7 cells for three passages. Cell free supernatant containing virus suspensions were collected and used to infect CV-1 cells (refer to Materials and Methods). Proteins from GP42-GFP or GP42-H1 infected cell extracts were resolved on SDS-PAGE and screened for GFP (left) or HA (right) expression by western analysis. Mock infected cells and allantoic fluid from PR/8 infected eggs are also shown.
doi:10.1371/journal.pone.0018780.g001

Plos ONE
DOI: 10.1371/journal.pone.0018780

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the Paramyxoviridae family, belongs in the genus respirovirus and shares 60–80% sequence homology to human parainfluenza virus type 1 (HPIV-1) [9,10].
The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPIV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses **have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16]**.
These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

Named entities + variants (human parainfluenza virus type, HPIV-1)
Speculation (reported, suggesting)
Species (human)
Cell types (**nasal epithelial cells**)
Facts
References

Electronic Health Records



```
        TITLE: PC ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18      ENTRY DATE: FEB 04, 2000@11:20
      AUTHOR:                      EXP COSIGNER:
     URGENCY:                           STATUS: COMPLETED

Chief Complaint: Patient notes 1 month history of blurred vision and
frequent urination

HISTORY OF PRESENT ILLNESS:

DEMO,FATHER is a 44 year-old MALE  who presents complaining of blurred
vision for the past 1 month. He finds it is difficult for him to read
clearly and is even effecting his driving.  He also notes that he has be
getting up to the bathroom frequently, esp. at night.  He now routinely
to urinate 3-4 times a night.  He is not aware of any particular weight
loss, but does feel thirsty much of the time.
-------------------------------------------------------------------------
PAST MEDICAL HISTORY:
Illnesses: Hypertension

Surgeries: None

Allergies: PENICILLINS

Medications:
1)    HYDROCHLOROTHIAZIDE 25MG TAB**  Qty: 45    ACTIVE
        for 90 days  Sig: TAKE ONE-HALF TABLET  Refills: 0
        MOUTH EVERY MORNING THC BLOOD PRESSURE
2)    METOPROLOL 25MG XL TAB  Qty: 90 for 90     ACTIVE
        days  Sig: TAKE ONE TABLET MOUTH QDAY    Refills: 0
        FOR THE HEART
-------------------------------------------------------------------------
FAMILY HISTORY:
Diabetes  Father, Sibling, Grandparent
```

http://www.upassoc.org/upa_publications/jus/2009february/smelcer5.html

Literary Texts

- Project Gutenberg (http://www.gutenberg.org/browse/scores/top)
- A team of horses passed from Finglas with toiling plodding tread, dragging through the funereal silence a creaking waggon on which lay a granite block. The waggoner marching at their head saluted.
  - Ulysses - http://www.gutenberg.org/files/4300/4300-h/4300-h.htm
- There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question.
  - Jane Eyre - http://www.gutenberg.org/files/1260/1260-h/1260-h.htm
- Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. Their house was small, for the lumber to build it had to be carried by wagon many miles. There were four walls, a floor and a roof, which made one room; and this room contained a rusty looking cookstove, a cupboard for the dishes, a table, three or four chairs, and the beds. Uncle Henry and Aunt Em had a big bed in one corner, and Dorothy a little bed in another corner. There was no garret at all, and no cellar--except a small hole dug in the ground, called a cyclone cellar, where the family could go in case one of those great whirlwinds arose, mighty enough to crush any building in its path. It was reached by a trap door in the middle of the floor, from which a ladder led down into the small, dark hole.
  - The Wizard of Oz - http://www.gutenberg.org/files/55/55-h/55-h.htm

A Really Long Literary Sentence

- Try parsing this
  - “Bloat is one of the co-tenants of the place, a maisonette erected last century, not far from the Chelsea Embankment, by Corydon Throsp, an acquaintance of the Rossettis' who wore hair smocks and liked to cultivate pharmaceutical plants up on the roof (a tradition young Osbie Feel has lately revived), a few of them hardy enough to survive fogs and frosts, but most returning, as fragments of peculiar alkaloids, to rooftop earth, along with manure from a trio of prize Wessex Saddleback sows quartered there by Throsp's successor, and dead leaves off many decorative trees transplanted to the roof by later tenants, and the odd unstomachable meal thrown or vomited there by this or that sensitive epicurean-all got scumbled together, eventually, by the knives of the seasons, to an impasto, feet thick, of unbelievable black topsoil in which anything could grow, not the least being bananas.“
- Do you know the source?

Quiz Answer

- “Gravity’s Rainbow” (by Thomas Pynchon), known for its use of very arcane words and complicated sentence (and plot) structure.
- Another such work is “Finnegans Wake” by James Joyce.
- Poetry is even more difficult.

Natural Language Processing

Research Projects and Evaluations

Demos

- https://playground.tensorflow.org/
- https://p.migdal.pl/interactive-machine-learning-list/
- https://demo.allennlp.org/reading-comprehension
- https://allenai.org/demos/
- http://aristo-demo.allenai.org/
- http://text-processing.com/demo/
- https://explosion.ai/demos/
- https://huggingface.co/hmtl/
- https://corenlp.run/
- http://nlp.stanford.edu:8080/corenlp/
- https://talktotransformer.com/
- https://cs.stanford.edu/people/karpathy/convnetjs/

State of the Art NLP Challenges

https://gluebenchmark.com/leaderboard
https://leaderboard.allenai.org/arc/submissions/public
https://yale-lily.github.io/spider
https://stanfordnlp.github.io/coqa/
https://rajpurkar.github.io/SQuAD-explorer/
http://www.msmarco.org/leaders.aspx
https://github.com/salesforce/decaNLP
http://ruder.io/tracking-progress-nlp/
http://nlpprogress.com/english/question_answering.html
https://quac.ai/
https://www.tau-nlp.org/csqa-leaderboard
https://nlp.stanford.edu/software/sempre/
http://nlp.cs.washington.edu/triviaqa/
http://lic.nlp.cornell.edu/nlvr/
https://decanlp.com/
https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a/
http://fever.ai/task.html
http://www.fakenewschallenge.org/
https://www.nyu.edu/projects/bowman/multinli/
http://convai.io/
http://sharedtask.duolingo.com/

YOU ASKED ARISTO:

Fossil fuels were formed from

(A) volcanoes
(B) the remains of living things
(C) gases in the atmosphere
(D) water trapped inside rocks

ARISTO ANSWERED:

Question: Fossil fuels were formed from

Aristo's Answer: (B) the remains of living things

Correct Answer: B

Confidence: 84.13%

as computed from these reasoners:

Information Retrieval: 50.50% MORE INFO

Justification Sentence: Fossil fuels were formed from the fossillized remains of dead plants and animals that once lived millions of years ago.

Table Reasoning: 34.70% MORE INFO

Knowledge Used: [living organisms | some form of respiration]

Topic Matching: 19.37% MORE INFO

Topic: fossil

Tuple Reasoning: 45.87% MORE INFO

Knowledge Used: [Fuels | come | from the remains of living things] [Fossil fuel | is formed | by the remains of plants and animals] [Fossil fuels | are formed | from the remains of once living organisms]

AristoRoBERTa: 86.98% MORE INFO

Ethnicity	Islam	Christianity	Judaism	Buddhism	Other	None / Atheism	n/a
TOTALS	70.20%	26.32%	0.03%	0.09%	0.02%	2.82%	0.51%
Kazakh	98.34%	0.39%	0.02%	0.01%	0.02%	0.98%	0.26%
Russian	1.43%	91.64%	0.04%	0.02%	0.03%	6.09%	0.75%
Uzbek	99.05%	0.39%	0.01%	0.01%	0.02%	0.37%	0.16%
Ukrainian	0.94%	90.74%	0.03%	0.01%	0.02%	7.31%	0.94%
Uyghur	98.35%	0.51%	0.02%	0.01%	0.03%	0.61%	0.47%
Tatar	79.57%	10.24%	0.02%	0.03%	0.06%	8.11%	1.97%
German	1.58%	81.59%	0.05%	0.04%	0.11%	13.96%	2.68%
Korean	5.24%	49.35%	0.21%	11.40%	0.14%	28.51%	5.16%
Turkish	99.13%	0.30%	0.01%	0.01%	0.02%	0.33%	0.21%
Azeri	94.81%	2.51%	0.02%	0.02%	0.03%	1.86%	0.76%
Belorussian	0.79%	90.16%	0.04%	0.01%	0.03%	7.82%	1.15%
Dungan	98.93%	0.37%	0.01%	0.03%	0.04%	0.34%	0.28%
Kurdish	98.28%	0.53%	0.03%	0.02%	0.02%	0.74%	0.38%
Tajik	97.78%	0.91%	0.01%	0.02%	0.08%	0.85%	0.35%
Polish	0.69%	90.07%	0.04%	0.01%	0.13%	7.30%	1.76%
Chechen	93.69%	2.99%	0.02%	0.01%	0.05%	2.08%	1.16%
Kyrgyz	96.67%	0.89%	0.03%	0.03%	0.02%	1.51%	0.86%
Others	34.69%	52.32%	0.82%	0.91%	0.13%	8.44%	2.69%

URL http://en.wikipedia.org/wiki?action=render&curid=16642&oldid=602424190
Title Kazakhstan
Table # 0

nt-83

which ethnicity is previous from dungan

Belorussian

nt-1833

which ethnicity has more followers of islam: tatar or tajik?

Tajik

nt-2689

are there more christian russians or ukrainians in kazakhstan?

Russian

nt-3253

how many times is each religion listed?

once

nt-4590

which ethnicity is above german

Tatar

nt-5037

what is the difference in percentage between korean buddists and german buddists?

11.36%

nt-6191

which ethnicity has the most buddhists in kazakhstan?

Korean

nt-6739

which ethnicity is first on the chart

Kazakh

nt-10453

what ethnicity is next under belorussian on the chart?

Dungan

Spider 1.0

Yale Semantic Parsing and Text-to-SQL Challenge

What is Spider?

Spider is a large-scale *complex and cross-domain* semantic parsing and text-to-SQL dataset annotated by 11 Yale students. The goal of the Spider challenge is to develop natural language interfaces to cross-domain databases. It consists of 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables covering 138 different domains. In Spider 1.0, different complex SQL queries and databases appear in train and test sets. To do well on it, systems must *generalize well to not only new SQL queries but also new database schemas*.

Why we call it "Spider"? It is because our dataset is complex and cross-domain like a spider crawling across mutiple complex(with many foreign keys) nests(databases).

Spider Paper (EMNLP'18)

Spider Post

Related challenges: multi-turn ***SParC*** and conversational ***CoSQL*** text-to-SQL tasks.

SParC Challenge (ACL'19)

CoSQL Challenge (EMNLP'19)

Leaderboard - Exact Set Match without Values

For exact matching evaluation, instead of simply conducting string comparison between the predicted and gold SQL queries, we decompose each SQL into several clauses, and conduct set comparison in each SQL clause. Please refer to the paper and the Github page for more details.

Rank	Model	Dev	Test
1 Dec 13, 2019	RATSQL v2 + BERT (DB content used) *Anonymous*	65.8	**61.9**
2 Dec 18, 2019	IRNet++ + XLNet (DB content used) *Anonymous*	65.5	60.1
3 Nov 12, 2019	RYANSQL + BERT *Anonymous*	66.6	58.2
4 Dec 13, 2019	RATSQL v2 (DB content used) *Anonymous*	62.7	57.2
5 Dec 13, 2019	RASQL + BERT *Anonymous*	60.8	55.7
6 Dec 13, 2019	EditSQL+LSL + BERT *Anonymous*	57.9	55.2
7 June 24, 2019	IRNet v2 + BERT *Microsoft Research Asia*	63.9	55.0
8 Sep 20, 2019	GIRN + BERT *Anonymous*	60.2	54.8

Not secure | nlp.stanford.edu:8080/corenlp/process

Stanford CoreNLP

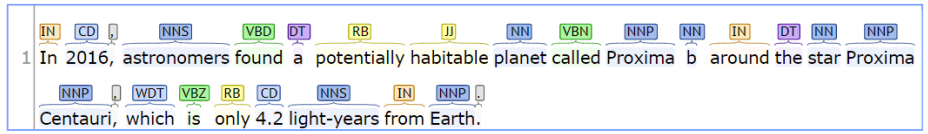
Output format: Visualise

Please enter your text here:

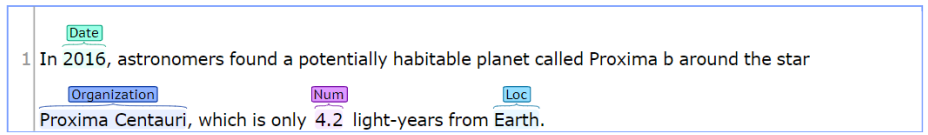
```
In 2016, astronomers found a potentially habitable
planet called Proxima b around the star Proxima
Centauri, which is only 4.2 light-years from Earth.
```

Submit Clear

Part-of-Speech:



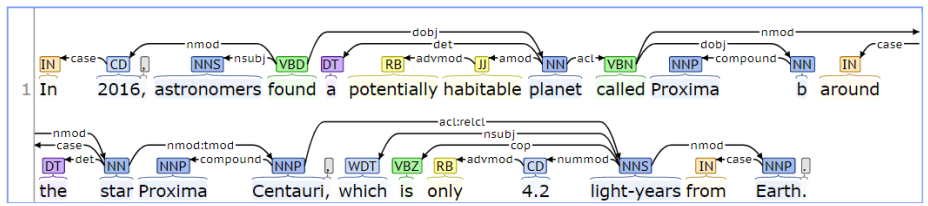
Named Entity Recognition:



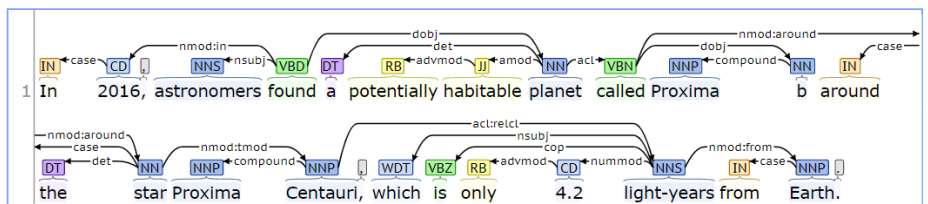
Coreference:

1 In 2016, astronomers found a potentially habitable planet called Proxima b around the star Proxima Centauri, which is only 4.2 light-years from Earth.

Basic Dependencies:



Enhanced Dependencies:



Visualisation provided using the brat visualisation/annotation software.

The LILY Lab

- http://lily.yale.edu
- http://aan.how
- http://www.nacloweb.org

LILY Projects

- Text summarization
- Crosslingual information retrieval
- Analysis of electronic health records
- Survey generation
- Text to SQL
- Dialogue systems
- Multilingual computing
- Text generation

- **Ask me for details**

NLP research in LILY

- **Multilingual information retrieval** -- We collaborate with researchers from Columbia University, the University of Maryland, the University of Edinburgh, and the University of Cambridge to build search engines for English users to query documents written in other languages. This cross-lingual information retrieval system improves our capability of understanding and processing different low-resource languages and it offers users a reliable access to foreign documents. Currently working on Tagalog, Swahili, Somali, Lithuanian, Bulgarian, Pashto.
- **Resources for learning NLP and AI** -- We aim to make dynamic research topics more accessible to the public by generating surveys of topics, discovering prerequisite relations among topics and recommending appropriate resources based on a given individual's education background and needs. We host a search engine, AAN (All About NLP) and tool which is available at http://aan.how.
- **Medical NLP –** We work on NLP on electronic health records. The tasks include abbreviation disambiguation, patient digital pairs, patient history record summarization, medical text segmentation and translation,. We are interested in how to transfer general knowledge to the medical domain and how to take good advantage of the limited high-quality data. We are also developing an NLP toolkit for the medical domain to perform tasks like named entity recognition and relation extraction.
- **Natural language database interfaces and dialogue systems** -- The goal of this project is to allow users with no SQL background to talk to relational databases directly using human language. This way, anyone can easily query and analyze a vast amount of data. We also aim to build conversational interfaces for even more natural information access, where the users participate in a conversation and the system takes the responsibility of choosing data sources and developing queries.
- **Text summarization and text generation –** single and multi-document, monolingual and multilingual, extractive and abstractive summarization, opinion summarization, etc.

NLP

Natural Language Processing

113

Brief History of NLP

The Turing Test

- Alan Turing: the *Turing test*
  - language as test for intelligence
- Three participants
  - a computer and two humans (one is an interrogator)
- Interrogator's goal
  - to tell the machine and human apart
- Machine's goal
  - to fool the interrogator into believing that a person is responding
- Other human's goal
  - to help the interrogator reach his goal

Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: 105621 (after a pause)

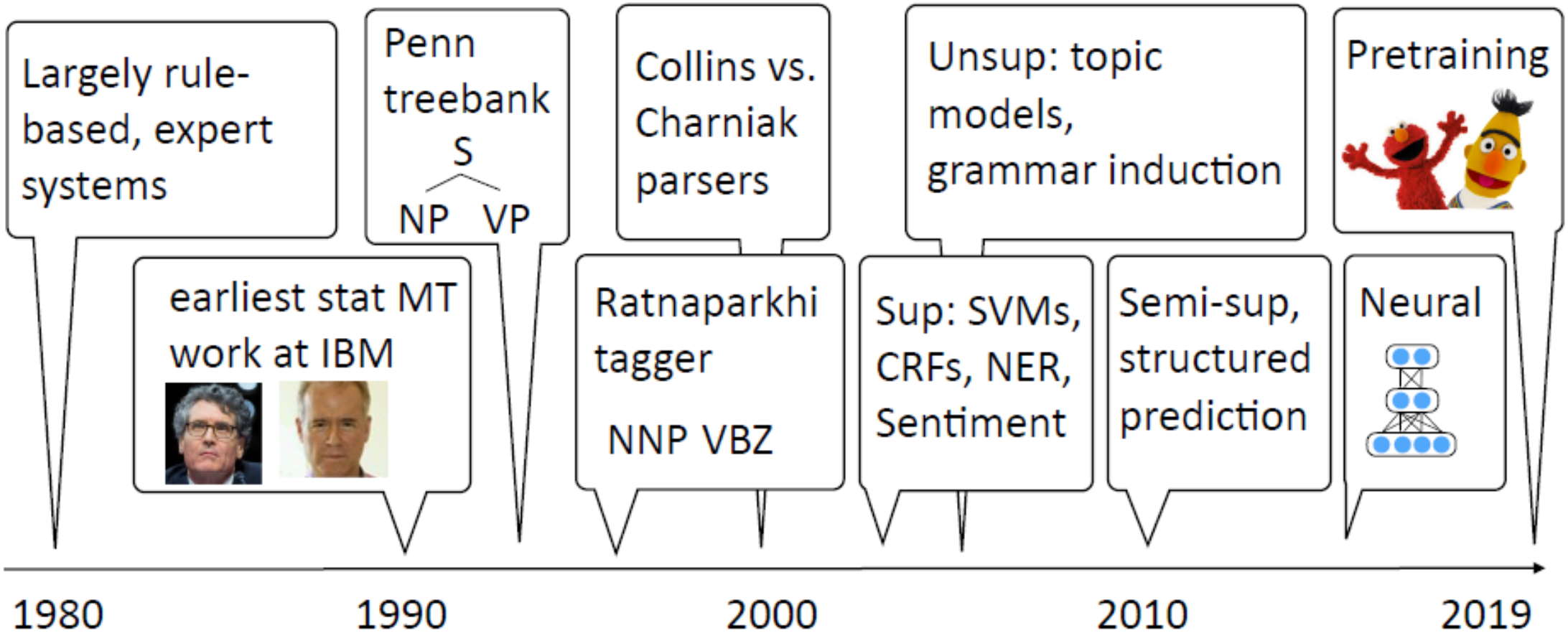
Eliza

$User_1$:	You are like my father in some ways.
$ELIZA_1$:	WHAT RESEMBLANCE DO YOU SEE
$User_2$:	You are not very aggressive but I think you don't want me to notice that.
$ELIZA_2$:	WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
$User_3$:	You don't argue with me.
$ELIZA_3$:	WHY DO YOU THINK I DON'T ARGUE WITH YOU
$User_4$:	You are afraid of me.
$ELIZA_4$:	DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

Longer example here:

http://thoughtcatalog.com/oliver-miller/2012/08/a-conversation-with-eliza

A Brief History of NLP



[Slide from Greg Durrett]

Some Brief History

- Foundational insights (1940's and 1950's)
  - automaton (Turing)
  - probabilities
  - information theory (Shannon)
  - formal languages (Backus and Naur)
  - noisy channel and decoding (Shannon)
  - first systems (Davis et al., Bell Labs)
- Two camps (1957-1970): symbolic and stochastic
  - Transformation Grammar (Harris, Chomsky)
  - Artificial Intelligence (Minsky, McCarthy, Shannon, Rochester)
  - automated theorem proving and problem solving (Newell and Simon)
  - Bayesian reasoning (Mosteller and Wallace)
  - Corpus work (Kučera and Francis)

Some Brief History

- Four paradigms (1970-1983)
  - stochastic (IBM)
  - logic-based (Colmerauer, Pereira and Warren, Kay, Bresnan)
  - nlu (Winograd, Schank, Fillmore)
  - discourse modelling (Grosz and Sidner)
- Empiricism and finite-state models redux (83-93)
  - Kaplan and Kay (phonology and morphology)
  - Church (syntax)
- Late years (1994-2010)
  - integration of different techniques
  - different areas (including speech and IR)
  - probabilistic models
  - machine learning
  - structured prediction
  - topic models

The Most Recent Years

- Machine learning methods
  - SVM, logistic regression (maxent), CRFs
- Shared tasks
  - TREC, DUC, TAC, *SEM
- Semantic tasks
  - RTE, SRL
- Semi-supervised and unsupervised methods
  - Zero-shot learning, transfer learning, self training
- Deep Learning
  - Embeddings, LSTM, CNN, Attention, GANs, Transformers, BERT

Natural Language Processing

113

NACLO

NACLO and IOL

- The North American Computational Linguistics Open Competition (formerly Olympiad)
  - Competition held since 2007 in the USA and Canada
  - http://www.nacloweb.org
- Best individual US performers so far:
  - Adam Hesterberg (2007)
  - Rebecca Jacobs (2007-2009) – 3 team golds + 2 individual medals
  - Ben Sklaroff (2010)
  - Morris Alper (2011)
  - Alex Wade (2012, 2013) – 2 team golds + 2 individual golds + 1 individual silver
  - Tom McCoy (2013) – Yale grad
  - Darryl Wu (2012, 2014)
  - James Wedgwood (2015, 2016) – Yale senior
  - Brian Xiao (2017), Swapnil Garg (2018)
- Other strong countries:
  - Russia, UK, Netherlands, Poland, Bulgaria, South Korea, Canada, China, Sweden, Slovenia, Brazil, Taiwan
- IOL – the International contest
  - Since 2003
  - 2014 in China, 2015 in Bulgaria, 2016 in India, 2017 in Ireland, 2018 in Czechia, 2019 in Korea, 2021 in Latvia
  - http://www.ioling.org

Consider these phrases in Ancient Greek (in a Roman-based transcription) and their unordered English translations:

(A) ***ho tōn hyiōn dulos***	
(B) ***hoi tōn dulōn cyrioi***	(1) the donkey of the master
(C) ***hoi tu emporu adelphoi***	(2) the brothers of the merchant
(D) ***hoi tōn onōn emporoi***	(3) the merchants of the donkeys
(E) ***ho tu cyriu onos***	(4) the sons of the masters
(F) ***ho tu oicu cyrios***	(5) the slave of the sons
(G) ***ho tōn adelphōn oicos***	(6) the masters of the slaves
(H) ***hoi tōn cyriōn hyioi***	(7) the house of the brothers
	(8) the master of the house

C1. Place the number of the correct English translation in the space following each Greek sentence. Explain your answers!

C2. Translate into Ancient Greek:

the houses of the merchants;
the donkeys of the slave

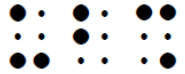
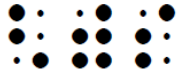
Explain your answers!

[A Donkey in Every House, by Todor Tchervenkov, NACLO 2007]

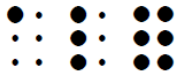
To the right is a Japanese word written in the ***tenji*** ("dot characters") writing system. The large dots represent the raised bumps; the tiny dots represent empty positions.

karaoke 

A1. The following *tenji* words represent ***atari***, ***haiku***, ***katana***, ***kimono***, ***koi***, and ***sake***. Which is which? You don't need to know either Japanese or Braille to figure it out; you'll find that the system is highly logical.

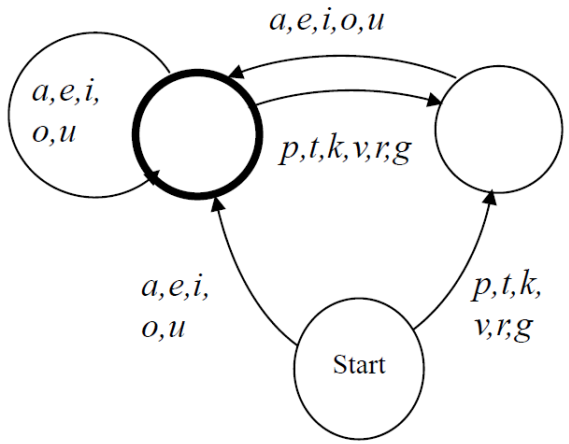
a. _____ 	b. _____ 
c. _____ 	d. _____ 
e. _____ 	f. _____ 

A2. What are the following words?

g. _____ 	h. _____ 

[Tenji Karaoke, by Patrick Littell, NACLO 2009]

Here is a finite state automaton that can distinguish between possible and impossible words in Rotokas, a language spoken on the island of Bougainville off the coast of New Guinea. Rotokas has a very simple system of sounds and allows us to create a very small FSA.



An FSA works like a board game. Choose a word, and place your pencil on the space marked "Start". Going through the letters of the word one at a time, move your pencil along the path marked with that letter. If the word ends and you're at a space marked with a thicker circle, the word succeeds: it's a possible Rotokas word! If the word ends and you're not at a thicker circle, or you're midway through the word and there's no path corresponding to the next letter, the word fails: it's *not* a possible Rotokas word!

Try it out with these possible and impossible words; the automaton should accept all the possible words and reject the impossible ones.

Possible Rotokas words	
tauo	*kareveiepa*
puraveva	*ovokirovuia*
avaopa	*ouragaveva*

Impossible Rotokas words	
grio	*ouag*
ovgi	*vonoka*
gataap	*oappa*

I1. Now, using the automaton above, put a check mark next to each possible Rotokas word:

___ *iu*	___ *uente*	___ *voav*
___ *idau*	___ *urioo*	___ *uaia*
___ *oire*	___ *raorao*	___ *oratreopaveiepa*

[aw-TOM-uh-tuh, by Patrick Littell, NACLO 2008]

On her visit to Armenia, Millie has gotten lost in Yerevan, the nation's capital. She is now at the Metropoliten (subway) station named Shengavit, but her friends are waiting for her at the station named Barekamutyun. Can you help Millie meet up with her friends?

1. Assuming Millie takes a train in the right direction, which will be the first stop after Shengavit?

Note that all names of stations listed below appear on the map.

a. Gortsaranayin
b. Zoravar Andranik
c. Charbakh
d. Garegin Njdehi Hraparak
e. none of the above

2. After boarding at Shengavit, how many stops will it take Millie to get to Barekamutyun (don't include Shengavit itself in the number of stops)?

[Lost in Yerevan, by Dragomir Radev, NACLO 2010]

Some computational problems

http://www.nacloweb.org/resources/problems/2016/N2016-B.pdf
http://www.nacloweb.org/resources/problems/2016/N2016-H.pdf
http://www.nacloweb.org/resources/problems/2016/N2016-K.pdf
http://www.nacloweb.org/resources/problems/2016/N2016-P.pdf
http://www.nacloweb.org/resources/problems/2015/N2015-E.pdf
http://www.nacloweb.org/resources/problems/2015/N2015-K.pdf
http://www.nacloweb.org/resources/problems/2015/N2015-M.pdf
http://www.nacloweb.org/resources/problems/2015/N2015-P.pdf
http://www.nacloweb.org/resources/problems/2015/N2015-G.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-O.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-P.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-C.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-J.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-H.pdf
http://www.nacloweb.org/resources/problems/2014/N2014-L.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-C.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-F.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-H.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-L.pdf
http://www.nacloweb.org/resources/problems/2012/N2012-C.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-N.pdf
http://www.nacloweb.org/resources/problems/2013/N2013-Q.pdf
http://www.nacloweb.org/resources/problems/2012/N2012-K.pdf
http://www.nacloweb.org/resources/problems/2012/N2012-O.pdf
http://www.nacloweb.org/resources/problems/2012/N2012-R.pdf
http://www.nacloweb.org/resources/problems/2011/F.pdf
http://www.nacloweb.org/resources/problems/2011/M.pdf
http://www.nacloweb.org/resources/problems/2010/D.pdf
http://www.nacloweb.org/resources/problems/2010/E.pdf
http://www.nacloweb.org/resources/problems/2010/I.pdf
http://www.nacloweb.org/resources/problems/2010/K.pdf
http://www.nacloweb.org/resources/problems/2009/N2009-E.pdf
http://www.nacloweb.org/resources/problems/2009/N2009-G.pdf
http://www.nacloweb.org/resources/problems/2009/N2009-J.pdf
http://www.nacloweb.org/resources/problems/2009/N2009-M.pdf
http://www.nacloweb.org/resources/problems/2008/N2008-F.pdf
http://www.nacloweb.org/resources/problems/2008/N2008-H.pdf
http://www.nacloweb.org/resources/problems/2008/N2008-I.pdf
http://www.nacloweb.org/resources/problems/2008/N2008-L.pdf
http://www.nacloweb.org/resources/problems/2007/N2007-A.pdf
http://www.nacloweb.org/resources/problems/2007/N2007-H.pdf

http://www.nacloweb.org/resources.php

NLP

Natural Language Processing

114

Why is NLP hard?

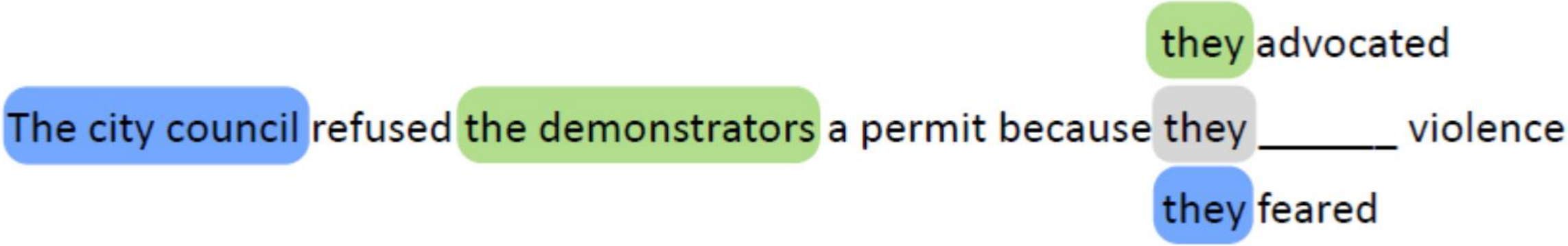
Silly Sentences

- Children make delicious snacks
- Stolen painting found by tree
- I saw the Rockies flying to San Francisco
- Court to try shooting defendant
- Ban on nude dancing on Governor’s desk
- Red tape holds up new bridges
- Government head seeks arms
- Cameron wins on budget, more lies ahead
- Local high school dropouts cut in half
- Hospitals are sued by seven foot doctors
- Dead expected to rise
- Miners refuse to work after death
- Patient at death's door - doctors pull him through
- In America a woman has a baby every 15 minutes. How does she do that?

The Winograd Schema Challenge

The city council refused the demonstrators a permit because they _____ violence

The Winograd Schema Challenge



More Classic Examples

- Beverly Hills
- Beverly Sills
- The box is in the pen
- The pen is in the box
- Mary and Sue are mothers
- Mary and Sue are sisters
- Every American has a mother
- Every American has a president


SADANDUSELESS.COM
Coffee or tea?
Coffee.
Wrong. It's tea.

Ambiguous Words

- ball, board, plant
  - meaning
- fly, rent, tape
  - part of speech
- address, resent, entrance, number, unionized
  - pronunciation – give it a try

Answer to the quiz

- address
  - The stress can be on either syllable. Compare with transport, effect, outline
- resent
  - As a verb infinitive or as “re-sent” a letter
- entrance
  - As a noun or as a verb meaning to put someone in a trance
- number
  - As a noun but also as the comparative of the adjective “numb”

Syntax vs. Semantics

** Little a has Mary lamb.*
? Colorless green ideas sleep furiously.

[Chomsky 1957]

Syntactic Amgbiguity

- How many different interpretations does the above sentence have?
- How many of them are reasonable/grammatical?

Time flies like an arrow.

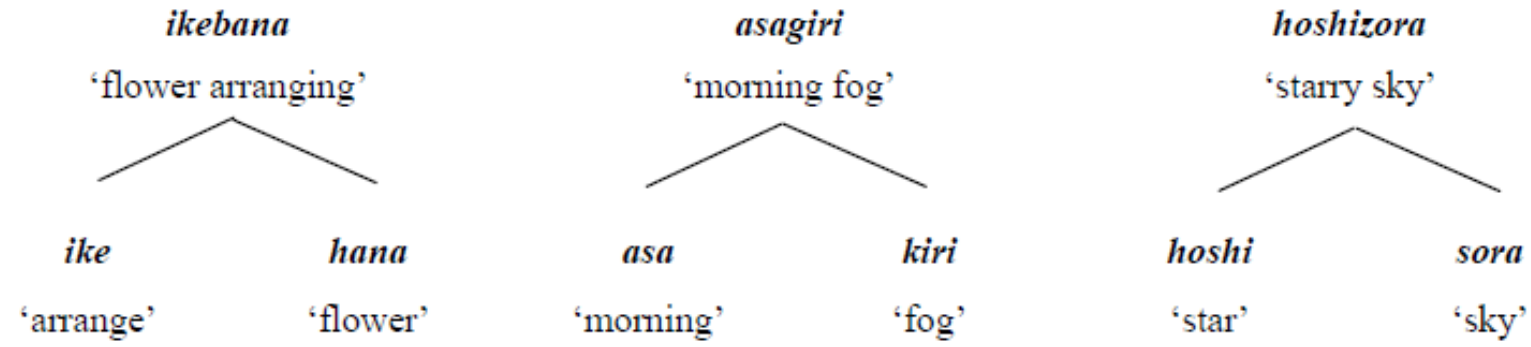
Quiz Answer

- The most obvious meaning is
  - time flies very fast; as fast as an arrow.
- This is a metaphorical interpretation.
  - Computers are not really good at metaphors.
- Other interpretations:
  - Flies like honey -> flies like an arrow -> fruit flies like an arrow
  - Take a stopwatch and time the race -> time the flies

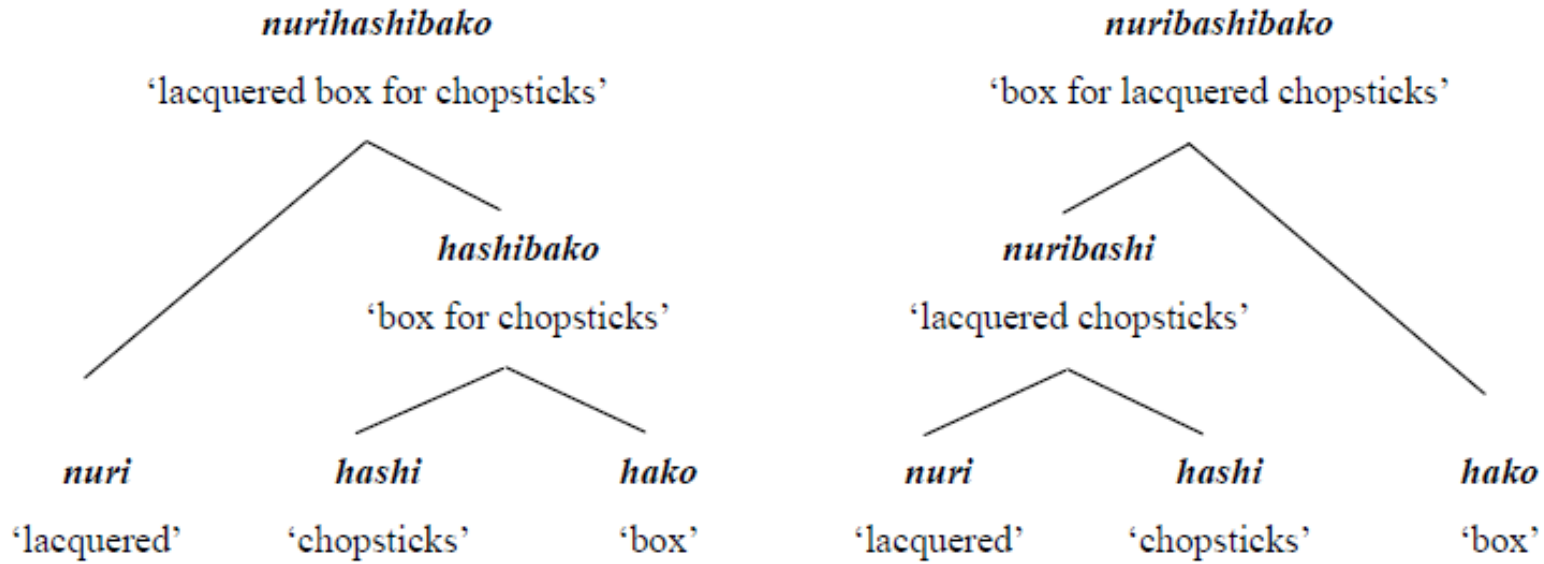
NACLO Problem

- Fakepapershelfmaker, by Willie Costello
  - http://www.nacloweb.org/resources/problems/2008/N2008-F.pdf

In English, we can combine two nouns to get a compound noun, such as in ‘mailbox’ or ‘sandcastle’. We can do this in Japanese as well, but just sticking the two words together isn’t enough. Instead, the words themselves undergo predictable changes:



Compound words can then be compounded again, creating compounds with three or more members. Study the diagrams below carefully. You’ll notice that the order in which the compound is built affects both the meaning and the final form of the word.



F1. The following is a list of several Japanese words with their English meanings. Use this word bank to write definitions of the Japanese compounds (a)-(f). Be very specific with how you phrase your definition. If your definition is ambiguous (has two meanings), it will not be counted.

sakura	cherry blossom	***kami***	paper	***nise***	fake
shiru	soup	***tana***	shelf	***tsukuri***	maker
iro	color(ed)	***tanuki***	raccoon	***hako***	box

(a) ***nisetanukijiru***	
(b) ***nisedanukijiru***	
(c) ***irogamibako***	
(d) ***irokamibako***	
(e) ***nisezakuradana***	
(f) ***nisesakuradana***	

F2. Match the following four-member Japanese compound words on the left with their English meanings on the right. (Some will require you to stretch your imagination a bit!) One of the Japanese words will correspond to two possible English meanings.

____ (1) a fake (fraudulent) shelf-maker made of paper	(A) ***nisegamidanadzukuri***
____ (2) a maker of fake shelves for paper	(B) ***nisekamitanadzukuri***
____ (3) a fake (fraudulent) maker of shelves for paper	(C) ***nisegamitanadzukuri***
____ (4) a shelf-maker made of fake paper	(D) ***nisekamidanadzukuri***
____ (5) a maker of shelves for fake paper	

F3. Explain your answers to F1 and F2 in the space provided below.

Solution

F1. The following is a list of several Japanese words with their English meanings; use them to write definitions of the Japanese compounds.

sakura	cherry blossom	***kami***	paper	***nise***	fake
shiru	soup	***tana***	shelf	***tsukuri***	maker
iro	color(ed)	***tanuki***	raccoon	***hako***	box

(a) ***nisetanukijiru***	fake soup made out of raccoons
(b) ***nisedanukijiru***	soup made out of fake raccoons
(c) ***irogamibako***	box for colored paper
(d) ***irokamibako***	colored box for paper
(e) ***nisezakuradana***	shelf for fake cherry blossoms
(f) ***nisesakuradana***	fake shelf for cherry blossoms

F2. Match the following four-member Japanese compound words with their English meanings; one of the Japanese words has two possible meanings.

(1) a fake shelf-maker made of paper	B: ***nisekamitanadzukuri***
(2) a maker of fake shelves for paper	D: ***nisekamidanadzukuri***
(3) a fake maker of shelves for paper	D: ***nisekamidanadzukuri***
(4) a shelf-maker made of fake paper	C: ***nisegamitanadzukuri***
(5) a maker of shelves for fake paper	A: ***nisegamidanadzukuri***

F3. Explain your answers.

When we compound two Japanese words, the first word modifies/describes the second. For example, adding *hashi* before *hako* makes a word meaning a box (*hako*) for chopsticks (*hashi*). As another example, adding *nuri* before *hashi* makes a word meaning chopsticks (*hashi*) that are lacquered (*nuri*).

Every simple (noncompound) word has two forms: the basic form, used when it occurs alone, and the variant form, sometimes used in compound words.

Basic	Variant
hako	*bako*
hana	*bana*
hashi	*bashi*
kami	*gami*
kiri	*giri*
sakura	*zakura*

Basic	Variant
shiru	*jiru*
sora	*zora*
tana	*dana*
tanuki	*danuki*
tsukuri	*dzukuri*

The variant form has a different first letter, which depends on the first letter in the basic form. Specifically, we replace the initial *h* with *b*, initial *k* with *g*, initial *s* with *z*, initial *sh* with *j*, initial *t* with *d*, and initial *ts* with *dz*. As a side note, some letters do not require replacement, but they do not occur in the problem.

NACLO Problem Solutions

- One Two Tree
  - http://www.nacloweb.org/resources/problems/2012/N2012-RS.pdf
- Fakepapershelfmaker
  - http://www.nacloweb.org/resources/problems/2008/N2008-FS.pdf

Types of Ambiguity

- Morphological:
  - Joe is quite impossible. Joe is quite important.
- Phonetic:
  - Joe’s finger got number.
- Part of speech:
  - Joe won the first round.
- Syntactic:
  - Call Joe a taxi.
- Prepositional phrase attachment:
  - Joe ate pizza with a fork / with meatballs / with Samantha / with pleasure.
- Sense:
  - Joe took the bar exam.

Other Sources of Difficulty

- Subjectivity:
  - Joe believes that stocks will rise.
- Cc attachment:
  - Joe likes ripe apples and pears.
- Negation:
  - Joe likes his pizza with no cheese and tomatoes.
- Referential:
  - Joe yelled at Mike. He had broken the bike.
  - Joe yelled at Mike. He was angry at him.
- Reflexive:
  - John bought him a present.
  - John bought himself a present.
- Ellipsis and parallelism:
  - Joe gave Mike a beer and Jeremy a glass of wine.
- Metonymy:
  - Boston called and left a message for Joe.

Other Sources of Difficulties

- Non-standard, slang, and novel words and usages
  - A360, 7342.67, +1-646-555-2223
  - “spam” or “friend” as verbs
  - yolo, selfie, chillax – recently recognized as dictionary words
  - www.urbandictionary.com – (Parental Warning!)
- Inconsistencies
  - junior college, college junior
  - pet spray, pet llama
- Typoes and gramattical erorz ☺
  - reciept, John Hopkins, should of
- Parsing problems
  - Selbständigkeit (self-reliance)
  - cup holder
  - Federal Reserve Board Chairman

Other Sources of Difficulties

- Complex sentences
- Counterfactual sentences
- Humor and sarcasm
- Implicature/inference/world knowledge:
  - I was late because my car broke down.
  - Implies I have a car, I use the car to get to places, the car has wheels, etc.
  - What is not explicitly mentioned, what is world knowledge?
- Semantics vs. pragmatics
  - Do you know the time?
- Language is hard even for humans
  - Both first language and second language

What do we need in order to understand language?

- World knowledge: have access to information beyond the training data

DOJ *greenlights* *Disney* - *Fox* *merger*

Department of Justice

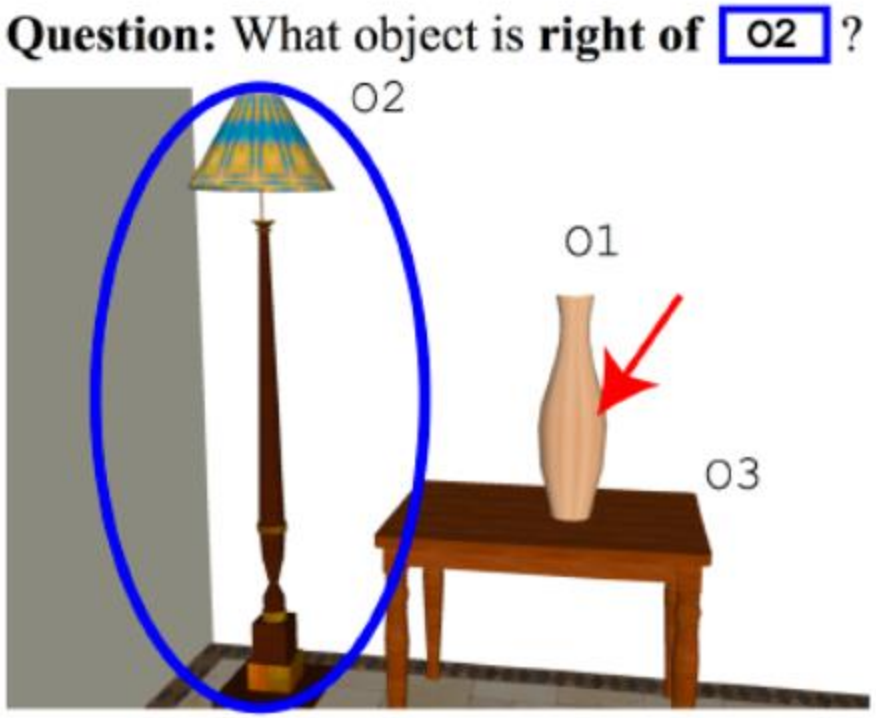
metaphor;
"approves"

- What is a green light? How do we understand what "green lighting" does?
- Need commonsense knowledge

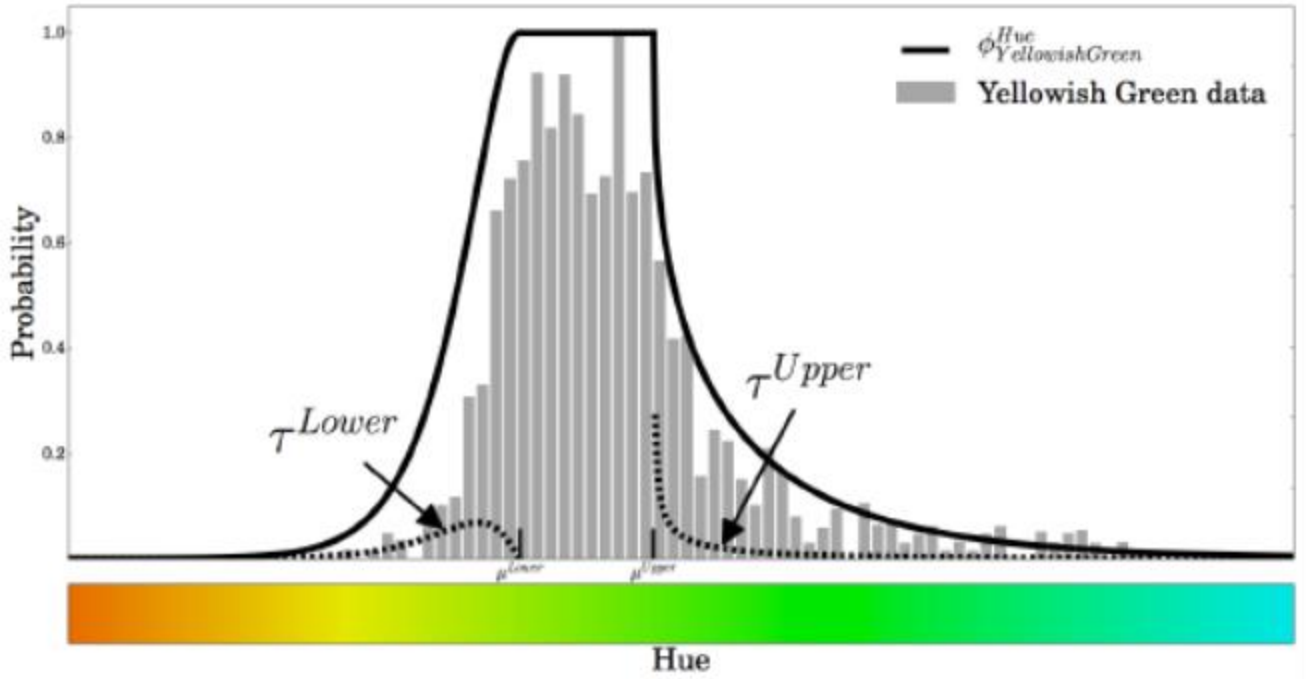
[Slide from Greg Durrett]

What do we need in order to understand language?

- Grounding: learn what fundamental concepts actually mean in a data-driven way



Golland et al. (2010)



McMahan and Stone (2015)

[Slide from Greg Durrett]

What do we need in order to understand language?

- Linguistic structure
- …but computers probably won't understand language the same way humans do
- However, linguistics tells us what phenomena we need to be able to deal with and gives us hints about how language works

a. John has been having a lot of trouble arranging his vacation.

b. He cannot find anyone to take over his responsibilities. (he = John)
C_b = John; C_f = {John}

c. He called up Mike yesterday to work out a plan. (he = John)
C_b = John; C_f = {John, Mike} (CONTINUE)

d. Mike has annoyed him a lot recently.
C_b = John; C_f = {Mike, John} (RETAIN)

e. He called John at 5 AM on Friday last week. (he = Mike)
C_b = Mike; C_f = {Mike, John} (SHIFT)

Centering Theory
Grosz et al. (1995)

[Slide from Greg Durrett]

What is needed to build a robust NLP system

- Lots of data
- Linguistic intuition
- Appropriate representation
- Robust algorithms
- World knowledge
- Grounding

SOURCE	Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.
HUMAN	That would be an interim solution which would make it possible to work towards a binding charter in the long term .
1x DATA	[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]
10x DATA	[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]
100x DATA	[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]
1000x DATA	[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]

[Example from Dan Klein]

NLP