Introduction to Statistical NLP

CSCI 3907/6907

Fall 2019

Lecture 1

Ayah Zirikly

ayah.zirikly@gmail.com

08/27/2019

What is NLP

Natural Language?

  - Languages spoken by people, e.g. English, Japanese, Arabic, as opposed to artificial languages, like C++, Java, …

- Computers using and processing natural language input (data) and producing useful information
- Software that can recognize, analyze and generate text and speech
- Typically NLP refers to processing unstructured data – text in free form

What is NLP

- Contrast to structured data
  - Information in “tables”

Employee	**Manager**	**Salary**
Smith, John	David, Richard	$80,000
Turner, Ian	**Smith, John**	**$59,000**
Huang, Chang	Smith, John	$69,000

- Typically allows numerical range and exact match (for text) queries

Salary < 60000 AND Manager = Smith, should return ***Turner, Ian***

What is NLP

- Contrast to structured data
  - Information in “tables”

Employee	Manager	Salary
Smith, John	David, Richard	$80,000
Turner, Ian	**Smith, John**	**$59,000**
Huang, Chang	Smith, John	$69,000

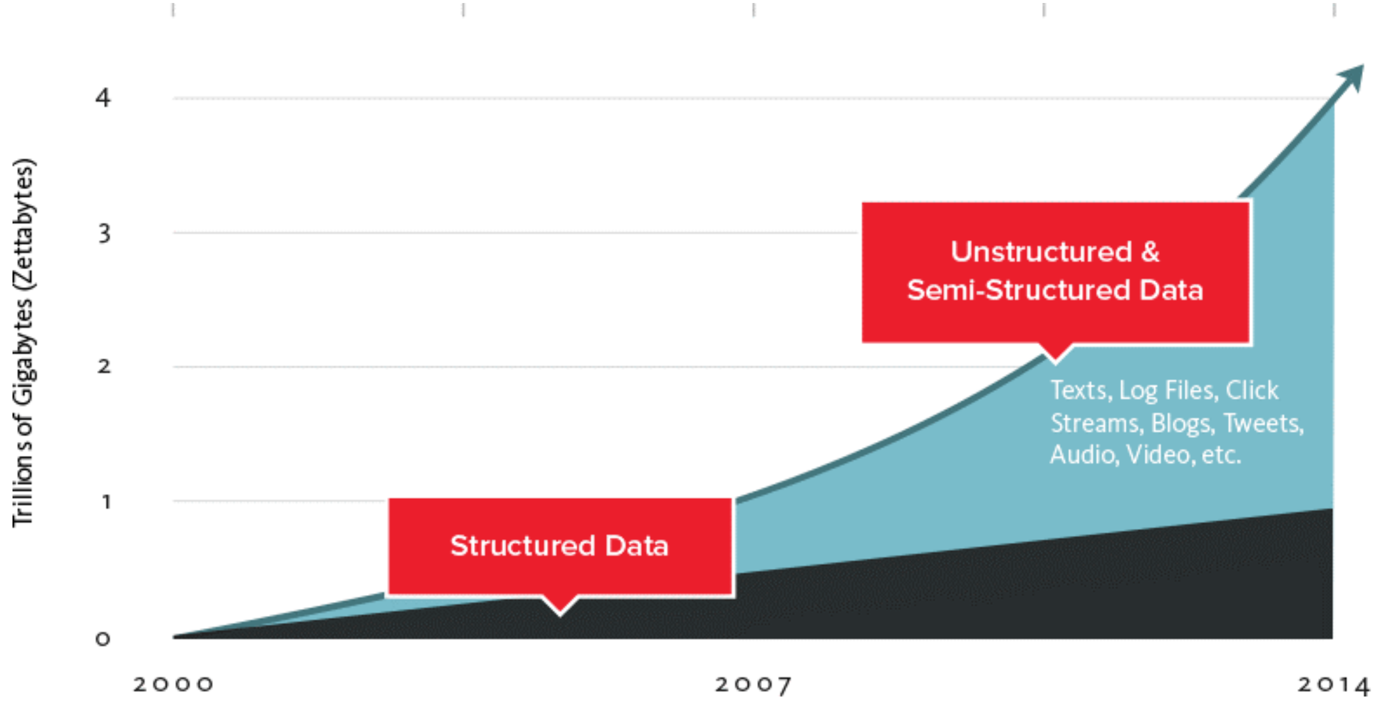
- Typically allows numerical range and exact match (for text) queries

What should we return?

Salary < 60000 AND Manager = Smith’, should return ***Turner, Ian***

From languages to information

- Unstructured → Structured data (Database, schemas, knowledge base)



https://www.couchbase.com

From languages to information

- For humans, going from the largely unstructured languages of the web to information is effortlessly easy
- But it's hard for computers
- Important for building the next generation of intelligent agents to make decisions on your behalf
    - Answering your routine email
    - Booking your next trip to Fiji

They need to be able to go from languages to information

Why should you care?

- An enormous amount of knowledge is now available in machine readable form as natural language text
- Conversational agents are becoming an important form of human-computer communication
- Much of human-human communication is now mediated by computers
- Very cool stuff! And with lots of commercial interest

Why NLP

kJfmmfj mmmvvv nnnffn333

Uj iheale eleee mnster vensi credur

Baboi oi cestnitze

Coovoel2^ ekk; ldsllk lkdf vnnjfj?

Fgmflmllk mlfm kfre xnnn!

Can you READ this? You, yes you!

Computers lack knowledge

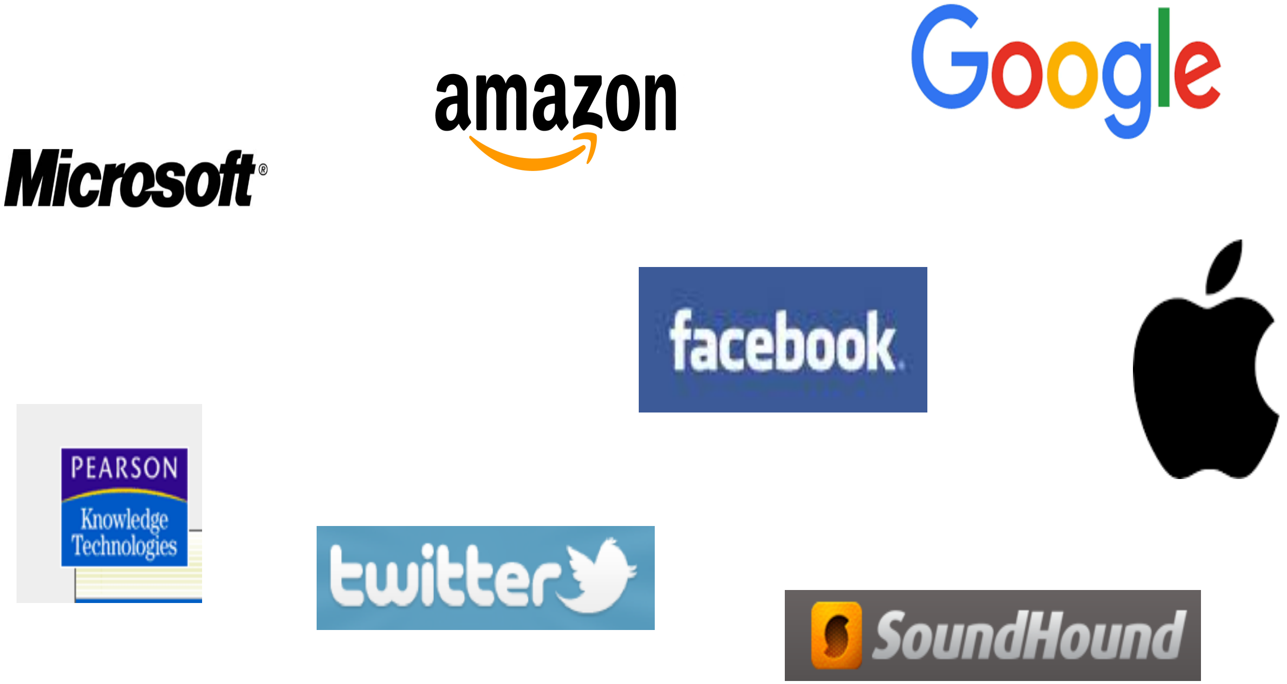
- Computers “see” text in English/Arabic/French the same way you saw the previous slide!
- People have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
- However, Computers have
  - No common sense knowledge
  - No reasoning capacity

  Unless we teach them!

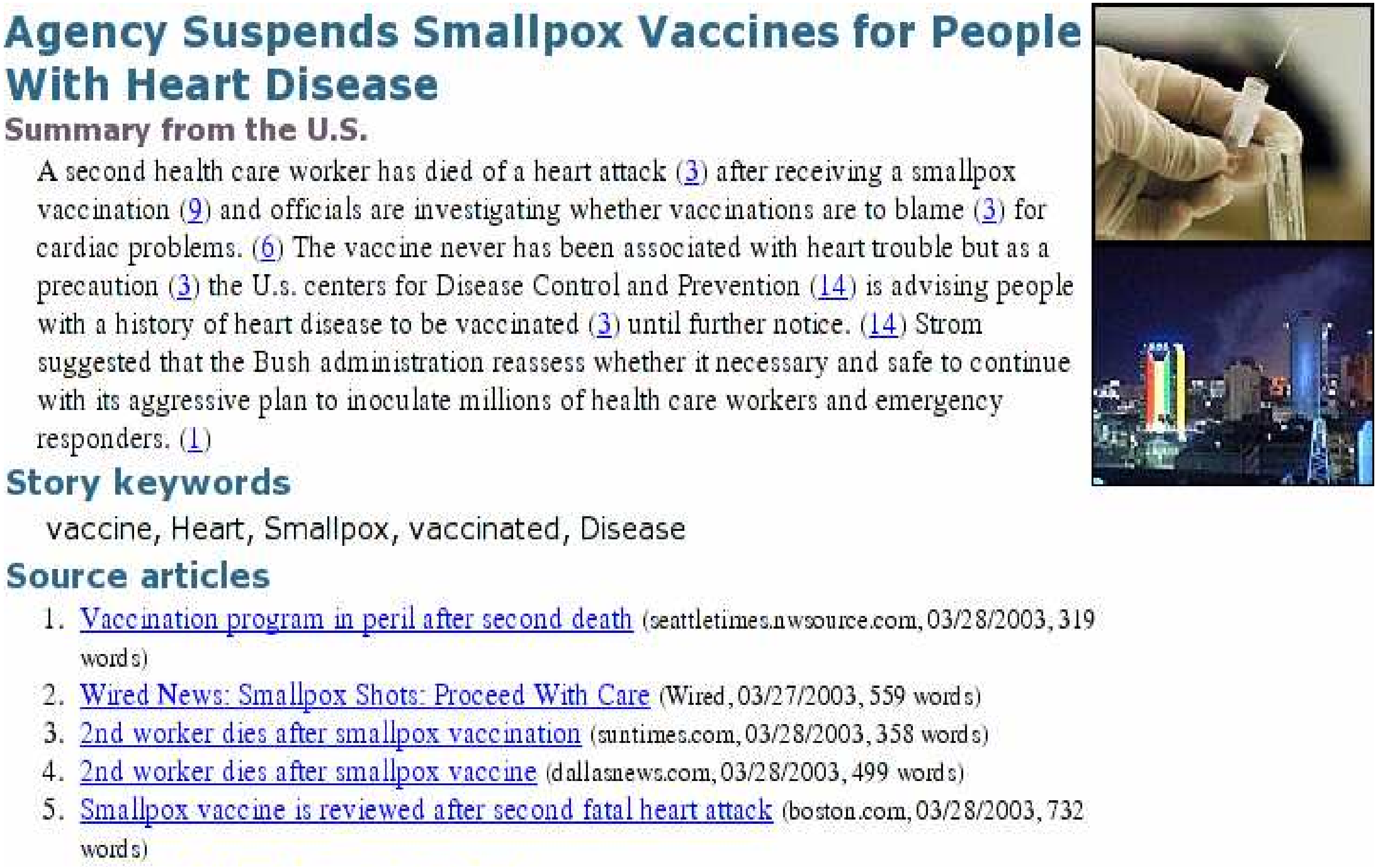
Applications of NLP

- Index and search large texts
- Automatic machine translation
- Automatic summarization
    - Condense 1 book into 1 page
- Question Answering
- Speech understanding
    - Understand phone conversations, personal assistants
- Text generation / dialogs
- Information extraction
    - Extract useful information from resumes
- Knowledge acquisition

Who uses NLP


Microsoft
amazon
Google
facebook
PEARSON
Knowledge Technologies
twitter
SoundHound

Text Summarization



Agency Suspends Smallpox Vaccines for People With Heart Disease

Summary from the U.S.

A second health care worker has died of a heart attack (3) after receiving a smallpox vaccination (9) and officials are investigating whether vaccinations are to blame (3) for cardiac problems. (6) The vaccine never has been associated with heart trouble but as a precaution (3) the U.s. centers for Disease Control and Prevention (14) is advising people with a history of heart disease to be vaccinated (3) until further notice. (14) Strom suggested that the Bush administration reassess whether it necessary and safe to continue with its aggressive plan to inoculate millions of health care workers and emergency responders. (1)

Story keywords

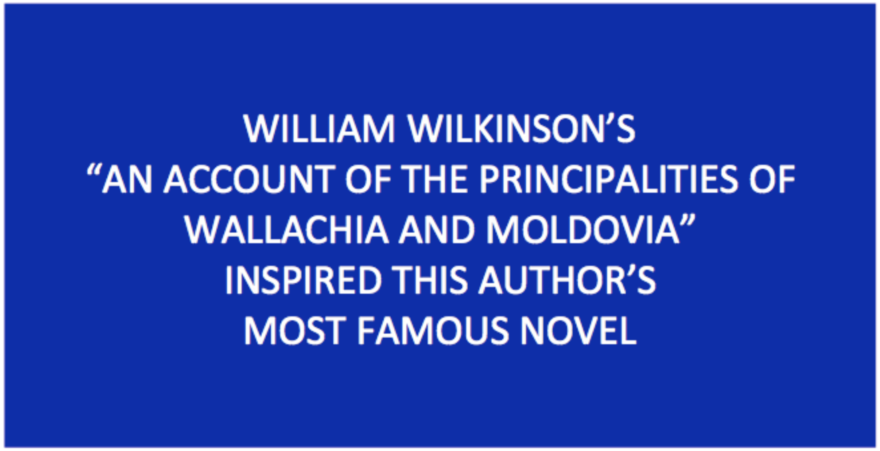
vaccine, Heart, Smallpox, vaccinated, Disease

Source articles

1. Vaccination program in peril after second death (seattletimes.nwsource.com, 03/28/2003, 319 words)
2. Wired News: Smallpox Shots: Proceed With Care (Wired, 03/27/2003, 559 words)
3. 2nd worker dies after smallpox vaccination (suntimes.com, 03/28/2003, 358 words)
4. 2nd worker dies after smallpox vaccine (dallasnews.com, 03/28/2003, 499 words)
5. Smallpox vaccine is reviewed after second fatal heart attack (boston.com, 03/28/2003, 732 words)

Question Answering

- IBM Watson won Jeopardy! in 2011



WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Applications

- Real-time interpreter

  Ambassador [0:47-2:03]

- Grammar and error checker

  Grammarly

- Personality analyzer
  - IBM personality insights

SENTIMENT ANALYSIS

Discovering people opinions, emotions and feelings about a product or service

http://paxcom.net/blog/why-is-customer-sentiment-analysis-important-for-your-brand/

Sentiment Analysis

Wow, great place!

Wow, 35 minutes to get a cup of coffee? Great job.

Not great but works as expected.

At first I hated it, but once the story hooked me, I found it difficult to put the book down

- Live demo

  http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

Blog Analytics

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

Descriptions of Languages

- Language = Words and Rules
  - Dictionary (vocabulary) + Grammar

Dictionary: set of words defined in the language; open (dynamic)

  - Traditional: paper based
  - Electronic: machine readable dictionaries

Grammar: set of rules which describe what is allowable in a language

  - Classical Grammars: meant for humans; mainly supported by examples; no (or almost no) formal description tools; cannot be programmed
  - Explicit Grammar: (CFG, Dependency Grammars, Link Grammars,...) formal description; can be programmed & tested on data (texts)

Typology of languages

A field of linguistics that studies and classifies languages according to their structural and functional features.

- Morphology
- Syntax
- Phonology

Typology of languages

A field of linguistics that studies and classifies languages according to their structural and functional features.

- Morphology
  - Meaningful morphological unit of a language that cannot be further divided

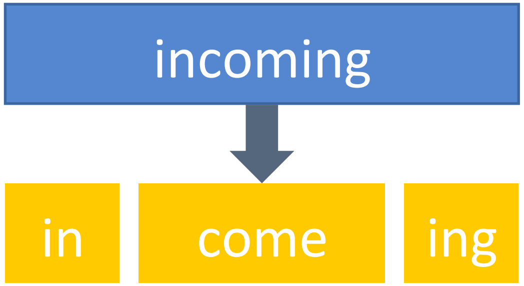
incoming

- Syntax
- Phonology

Typology of languages

A field of linguistics that studies and classifies languages according to their structural and functional features.

- Morphology
  - Meaningful morphological unit of a language that cannot be further divided



- Syntax
- Phonology

Typology of languages

- Morphological features
  - Analytical: using (function) words to express categories
    - There is little to no morphological change in words
    - English, classical Chinese, Vietnamese
      - I will eat an apple (convey future tense)
  - Inflective: using prefix/suffix/infix, combines several categories
    - Slavic: Czech, Russian, Arabic, German, French
    - كتب)Kataba >- كتبتُ)katabtu
  - Agglutinative:
    - words which may consist of more than one, and possibly many, morphemes
    - Examples: el-ler-imiz-in (Turkish)

Typology of languages

A field of linguistics that studies and classifies languages according to their structural and functional features.

- Morphology
- Syntax --Word order
  - VSO (Classical Arabic)
  - SVO (Arabic Dialects, English)
  - OSV (Turkish, Japanese)
- Phonology

Typology of languages

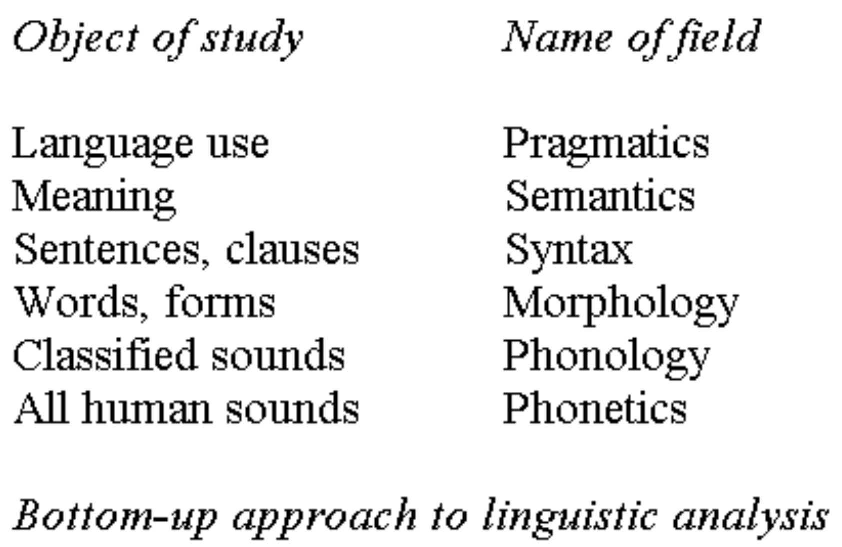
A field of linguistics that studies and classifies languages according to their structural and functional features.

- Morphology
- Syntax
- Phonology
    - Study of the patterns of sounds in a language and across languages
    - How speech sounds are organized in the mind and used to convey meaning
        - Different languages can use different phonemes, or different syllable structures (what sounds can go together to make sequences or words); phonology identifies these differences

Levels of Language Description

- 6 basic levels (more or less explicitly present in most theories):
  - and beyond (pragmatics/discourse/...)
  - Semantics : knowledge of meaning
  - Syntax : structural relationships between words
  - Morphology : meaningful components of words
  - Phonetics & Phonology
- Each level has an input and output representation
- output from one level is the input to the next (upper) level
- sometimes levels might be skipped (merged) or split

Levels of Language Description



Object of study	*Name of field*	*Size of unit*
Language use	Pragmatics	Largest
Meaning	Semantics	|
Sentences, clauses	Syntax	|
Words, forms	Morphology	|
Classified sounds	Phonology	|
All human sounds	Phonetics	Smallest ↑

Bottom-up approach to linguistic analysis

- Each level has an input and output representation
- output from one level is the input to the next (upper) level
- sometimes levels might be skipped (merged) or split

https://www.uni-due.de/SHE/REV_Levels_Chart.htm

Ambiguity

- All 6 levels of linguistic knowledge require resolving ambiguity

- Ambiguity results from the existence of multiple possibilities for each level

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Semantic ambiguity

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Semantic ambiguity

Lexical ambiguity

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Some Headlines…

- Iraqi Head Seeks Arms
- Teacher strikes idle kids
- Stolen painting found by tree
- Enraged Cow Injures Farmer With Ax
- Squad Helps Dog Bite Victim

Semantic ambiguity

Lexical ambiguity

Structural ambiguity

Ambiguity in Spoken Language

I made her duck

- I cooked waterfowl for her
- I cooked the waterfowl that belongs to her
- I created the ceramic duck she owns
- I caused her to quickly lower her head
- And more….

Dealing with Ambiguity

- Tightly coupled interaction among processing levels; knowledge from other levels can help decide at ambiguous levels.
- Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.
- Probabilistic approaches based on making the most likely choices.
- Don’t do anything, maybe it won’t matter.
  - We’ll leave when the duck is ready to eat.
  - The duck is ready to eat now.
    - Does the “duck” ambiguity matter with respect to whether we can leave?

Other difficulties

- Non-standard text
  - “ we’re soooo proud of u!”
- Idioms and metaphors
  - “dark horse” “cold feet” “lose face”
- Sarcasm
- Segmentation
  - “The New York-New Haven railroad”
- Named entities
  - “*Let It Be* sold millions”, “#elonmusk”
- …

NLP Categories

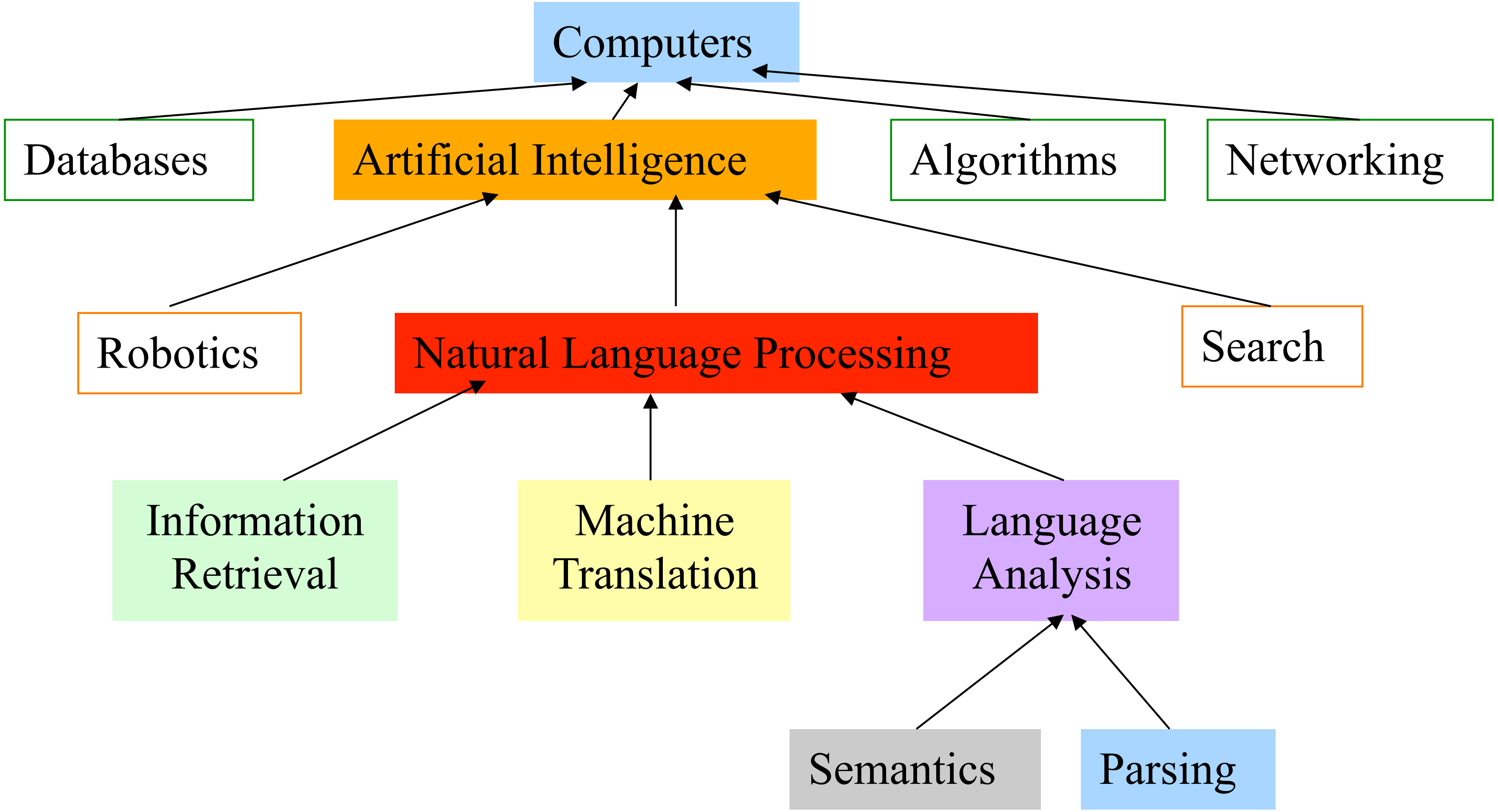
Applications

- Machine Translation (MT)
- Information Retrieval (IR) and Extraction (IE)
- Automatic Speech Recognition (ASR)
- Optical Character Recognition (OCR)
- Automatic Summarization, Speech Synthesis, etc.

Enabling Technologies

- Tokenization
- Part-of-Speech Tagging
- Syntactic Parsing
- Lemmatization
- Word Sense Disambiguation, etc…

NLP in CS taxonomy

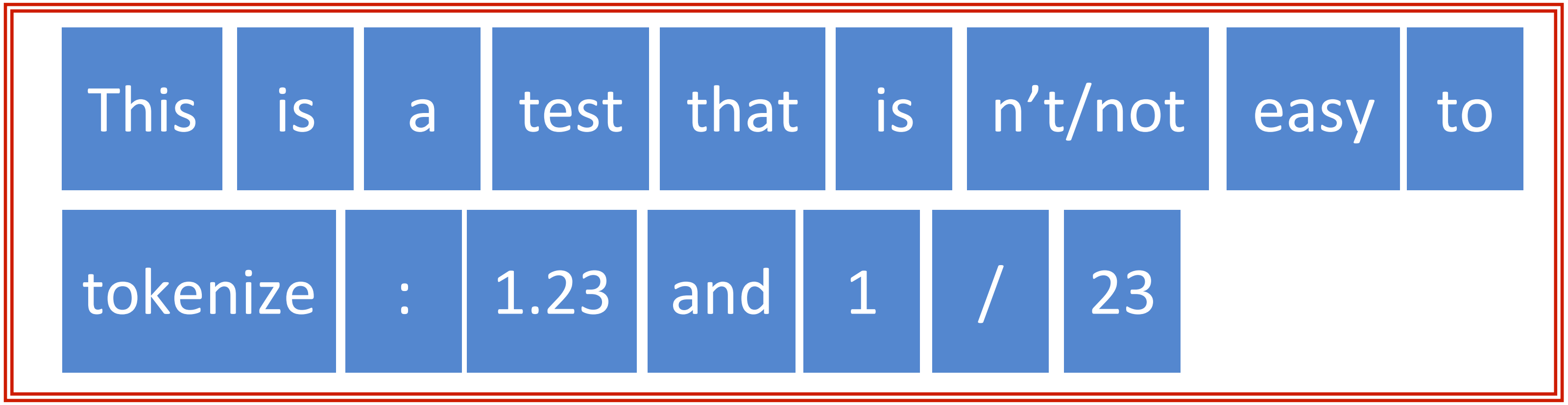


Tokenization

This is a test that isn’t easy to tokenize: 1.23 and 1/23

Tokenization

This is a test that isn’t easy to tokenize: 1.23 and 1/23

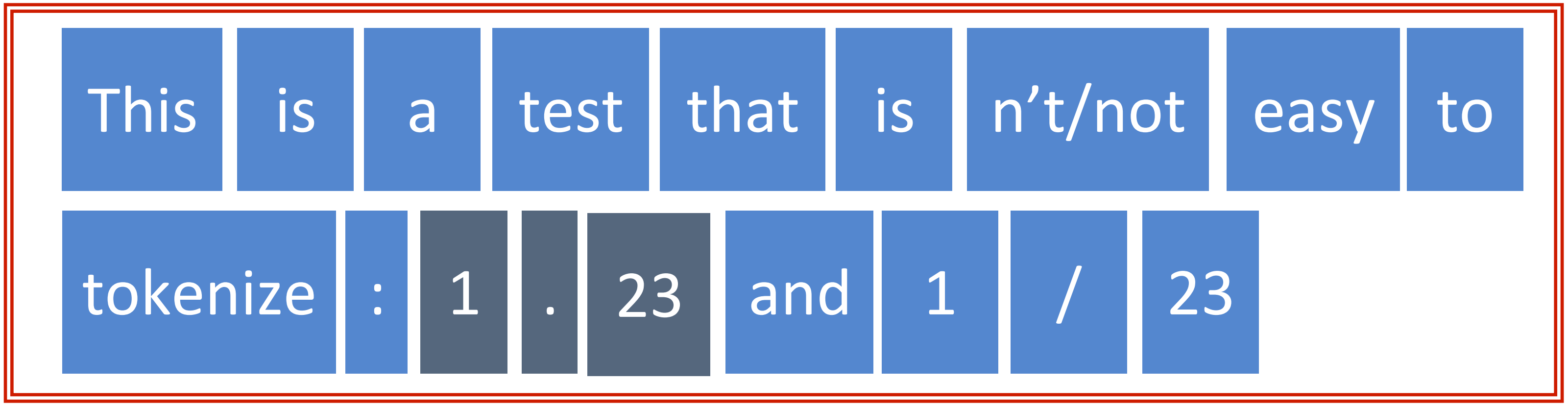


This is a test that is n’t/not easy to

tokenize : 1.23 and 1 / 23

Tokenization

This is a test that isn’t easy to tokenize: 1.23 and 1/23



This | is | a | test | that | is | n’t/not | easy | to

tokenize | : | 1 | . | 23 | and | 1 | / | 23

Applications

First, what makes an application a language processing application (as opposed to any other piece of software)?

- An application that requires the use of knowledge about human languages
  - Is Unix wc (word count) an example of a language processing application?

Applications

Word count?

- When it counts words: Yes
  - To count words you need to know what a word is
  - That's knowledge of language

- When it counts lines and bytes: No
  - Lines and bytes are computer artifacts, not linguistic entries

Adapted from Speech and Language Processing - Jurafsky and Martin

Information Extraction

As a task: **Filling slots in a database from sub-segments of text**

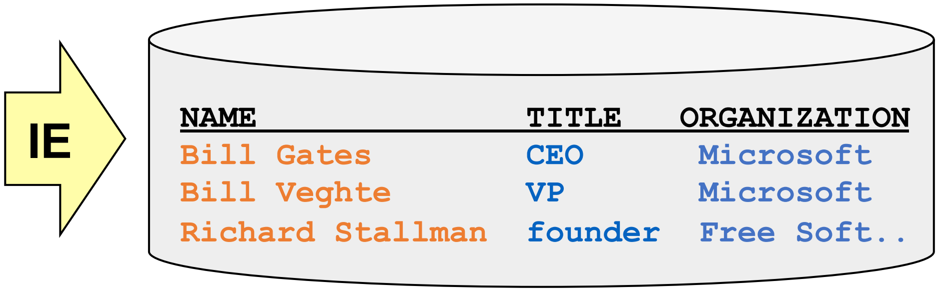
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Slide from Jason Eisner, Chris Brew, adapted from slide by William Cohen

Phrase types to identify for IE

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

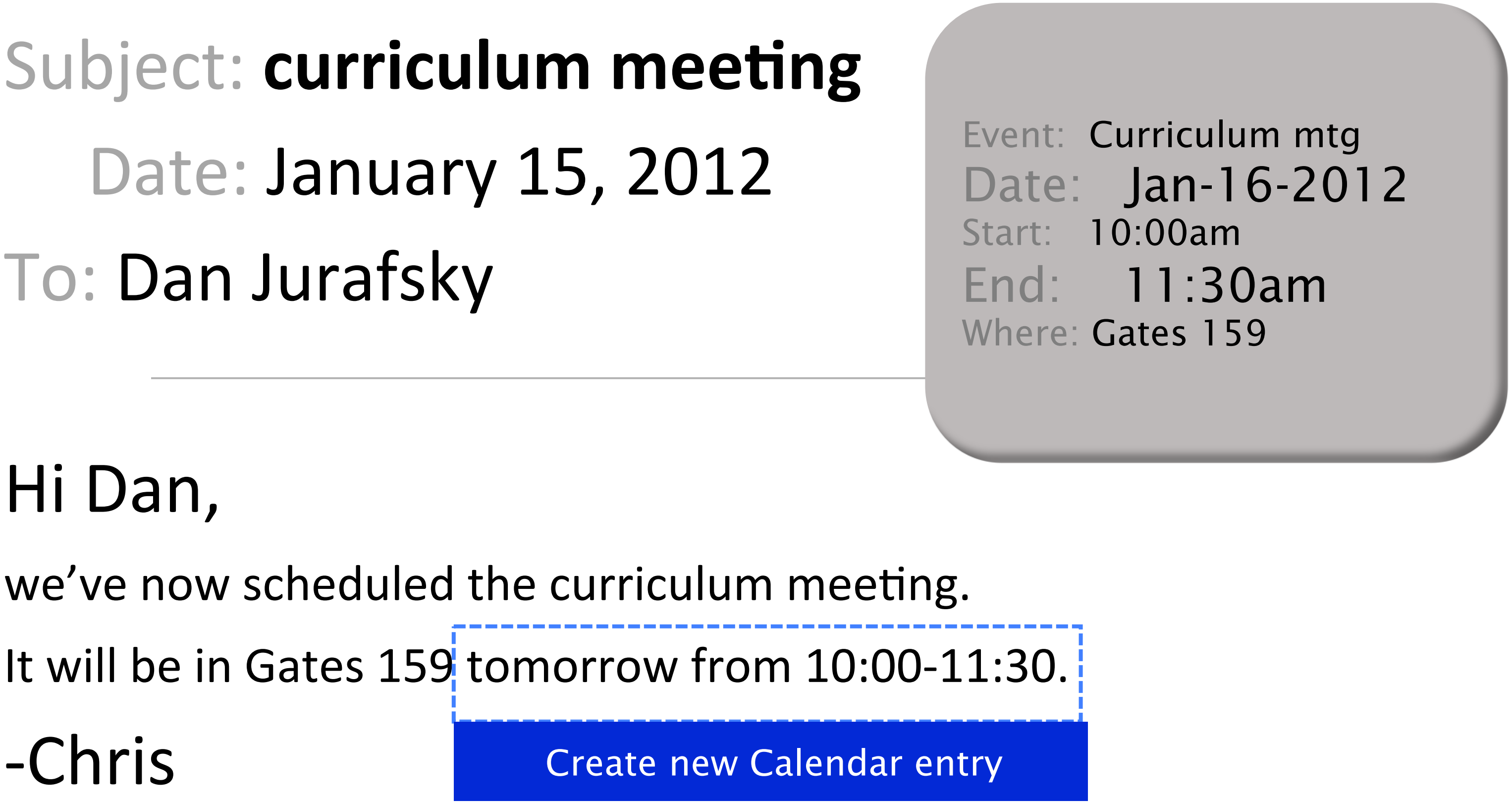
Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Slide from Jason Eisner, Chris Brew, adapted from slide by William Cohen

Information Extraction



Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

Hi Dan,

we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

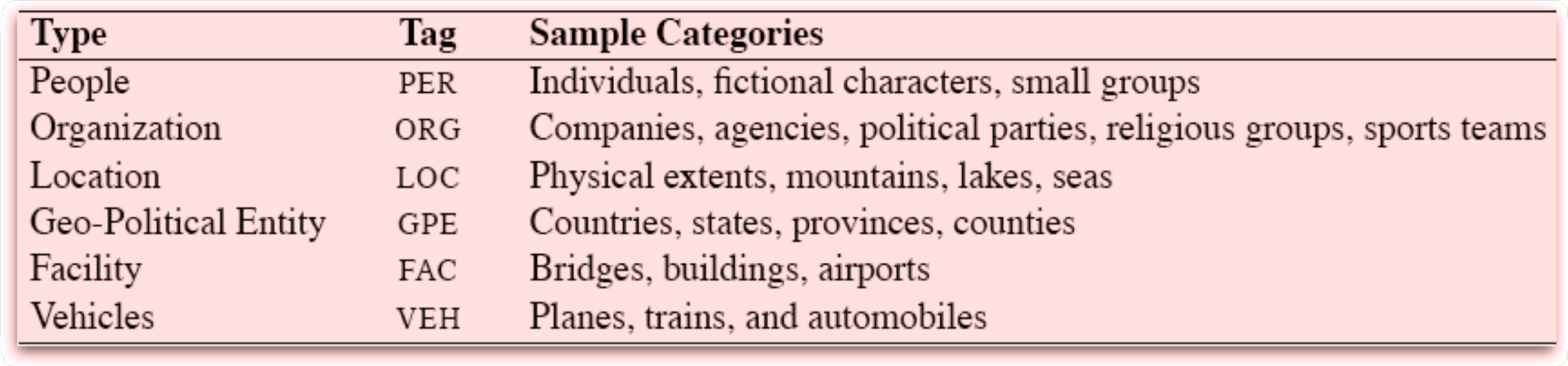
Event:	Curriculum mtg
Date:	Jan-16-2012
Start:	10:00am
End:	11:30am
Where:	Gates 159

Named Entity Recognition (NER)

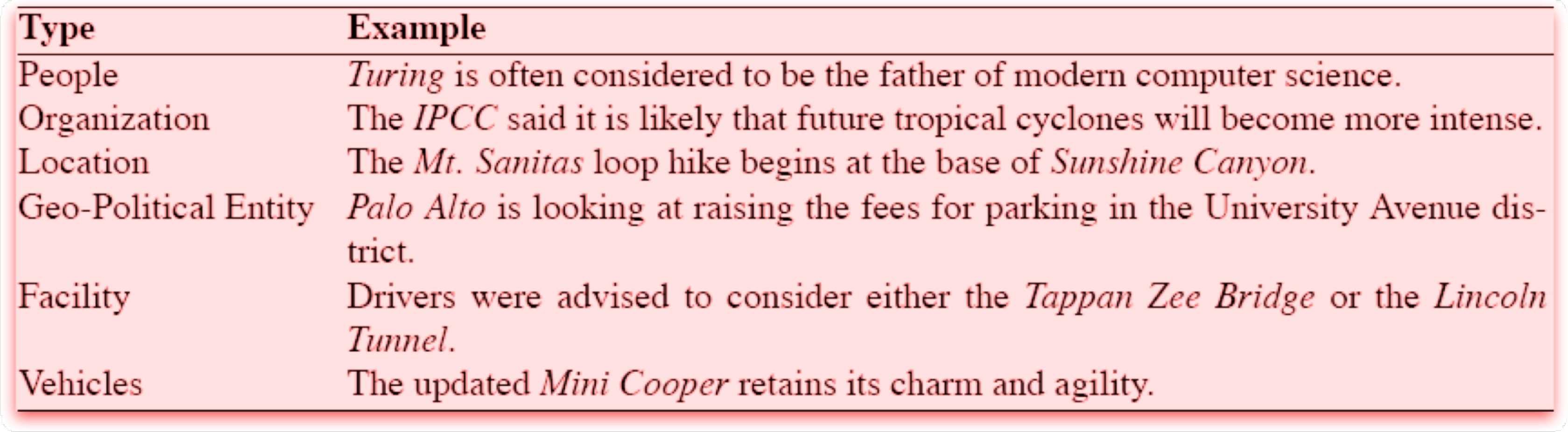
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

What is a Named Entity?

NE Types



Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles



Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

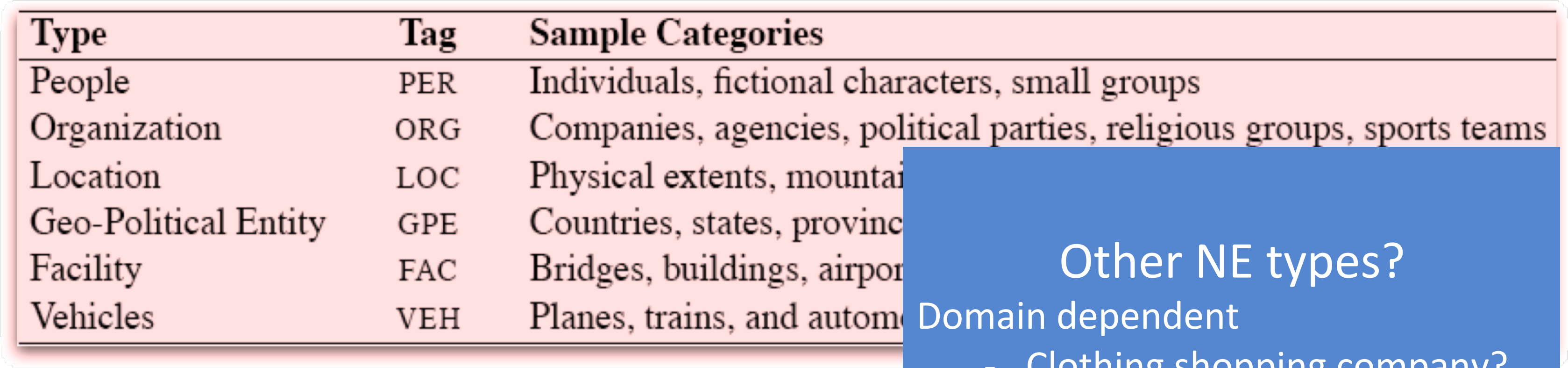
Slide adapted from Jason Eisner and Jim Martin

Named Entity Recognition

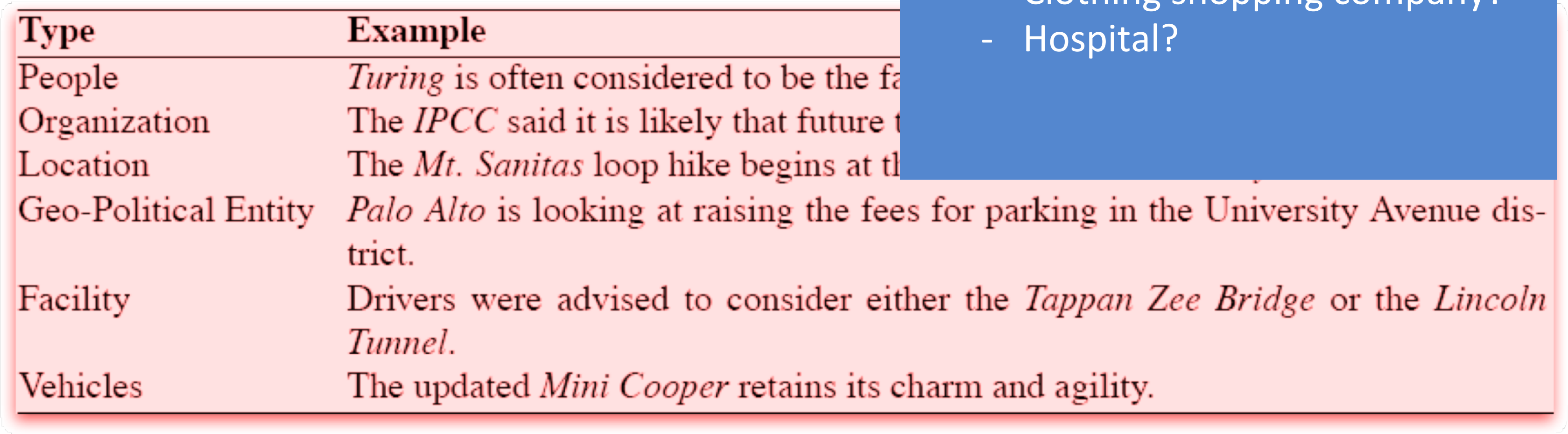
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Slide adapted from Jason Eisner and Jim Martin

NE Types



Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountai
Geo-Political Entity	GPE	Countries, states, provinc
Facility	FAC	Bridges, buildings, airpor
Vehicles	VEH	Planes, trains, and autom



Type	Example
People	*Turing* is often considered to be the fa
Organization	The *IPCC* said it is likely that future t
Location	The *Mt. Sanitas* loop hike begins at th
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Other NE types?

Domain dependent

- Clothing shopping company?
- Hospital?

NLP Toolkit

- Knowledge of Linguistics
    - NLPers call them features
- Rule-based systems
- Machine learning methods
    - Clustering
    - Classification (binary, multi-label)
    - Deep learning models
- Evaluation metrics
    - Precision, recall, F1, BLEU

NLP Approaches

- Rule-based/Symbolic Approaches
    - Linguists write rules that are applied by the machines
    - Works well on templates that have free text
- Corpus-based/Statistical Approaches
    - Supervised
        - Annotated data used for training
            - Parallel Corpora: translated text collections
            - Product reviews labeled for sentiment analysis
            - Speech Corpora with transcripts
    - Unsupervised – Unannotated data
    - Semi-supervised methods

BioNLP

- Information Extraction
  - Named Entity Recognition
  - Concept extraction and linking to existing knowledge bases
- Event extraction and temporal even ordering
- Relation extraction
  - Adverse Drug Event ADE: interaction between drugs and medication entities to prevent unwanted effects of drug

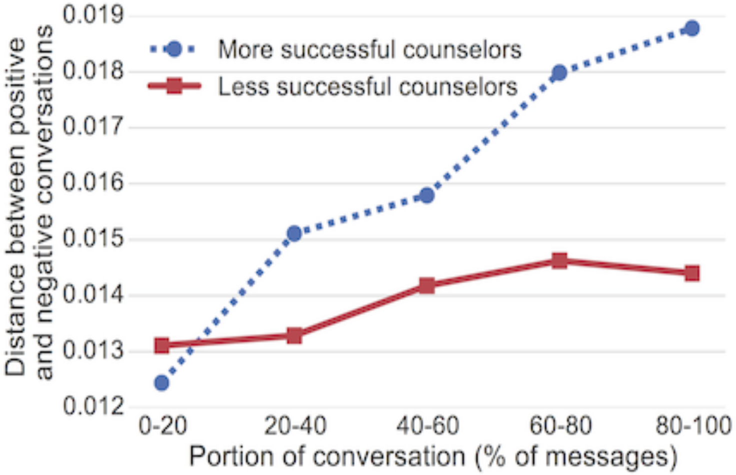
Example

- CLAMP: Clinical NLP toolkit @The university of Texas Health Science

NLP in Mental Health

How to help someone feel better?

	More successful counselors	Less successful counselors
Counselor message length (in words)	15.8	11.8
Counselor responds with check question	12.6%	4.1%
Counselor responds with suicide check	13.5%	10.3%
Counselor responds with thanks	6.3%	2.4%
Counselor responds with hedges	41.4%	36.8%



Adopted from: https://nlp.stanford.edu/blog/how-to-help-someone-feel-better-nlp-for-mental-health/

What are we going to study?

- How to get computers to perform useful and interesting tasks involving human languages
- Potential insights from CL into how humans process language in the mind (?)

What will we learn about in this course?

- **Morphology:** the way words are formed
- **Syntax:** the way words are grouped together into larger constituents and phrases and the way these phrases can be ordered
- **Semantics:** the context-independent ‘meaning’ of utterances
- **Pragmatics:** the context-dependent ‘meaning’ of utterances

And much more!!!

Goal: What is a speaker/writer meaning to convey?

Skills you will need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Basic machine learning
- Java or Python programming knowledge

What should you expect to get from this course?

- For the instructor to tell you what ☺
- But…

You will become an awesome NLPer

Contributions to the course material & slides

- Slides are sometimes adapted (with permission) from other great slide sets, namely from:
  - Mona Diab, Chris Manning, Dan Jurafsky, Jason Eisner, Jim Martin, Yassine Benajiba, Hanan Aldarmaki