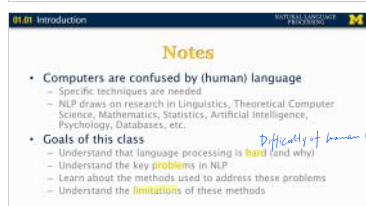
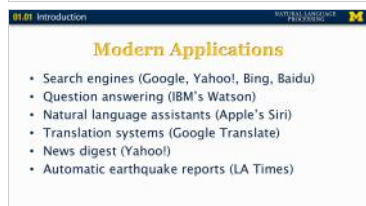
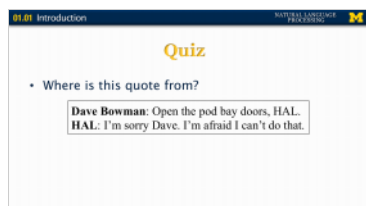
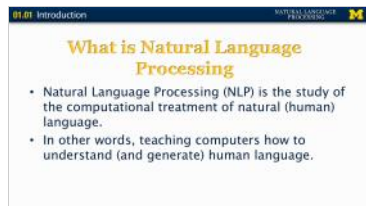
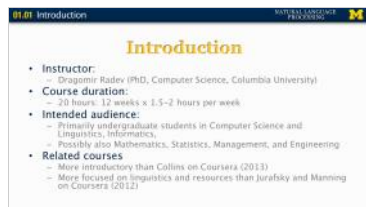


36608a8d  
3ab70b04















01.01 Introduction

NATURAL LANGUAGE PROCESSING

M

## 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), e.g. "he", "she", "this" refer to meaningful things only when given context

01.01 Introduction

NATURAL LANGUAGE PROCESSING

M


01.01 Introduction

NATURAL LANGUAGE PROCESSING

M

## Basic NLP Pipeline

- (U)nderstanding and (G)eneration



01.01 Introduction

NATURAL LANGUAGE PROCESSING

M

01.01 Introduction

NATURAL LANGUAGE PROCESSING

M

# NLP

01.01 Introduction

NATURAL LANGUAGE PROCESSING

M

Why NLP is challenging (with examples)

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

# NLP

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

## Introduction to NLP

### Examples of Text

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

## Understanding a News Story

**Brazil crowds attend funeral of late candidate Campos**

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday. They attended a funeral Mass and filed the streets of the city of Recife to follow the passage of his coffin. Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate. Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo. Investigators are still trying to establish the exact causes of the crash, which killed six other people. Mr Campos's private plane - a Cessna 441BQ - was travelling from Rio de Janeiro to the sea-side resort of Guarujá, near the city of Santos. **President Dilma Rousseff**, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

## Understanding a News Story

**Brazil crowds attend funeral of late candidate Campos**

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday. They attended a funeral Mass and filed the streets of the city of Recife to follow the passage of his coffin. Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate. Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo. Investigators are still trying to establish the exact causes of the crash, which killed six other people. Mr Campos's private plane - a Cessna 441BQ - was travelling from Rio de Janeiro to the sea-side resort of Guarujá, near the city of Santos. **President Dilma Rousseff**, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M

## Answer to the quiz

- Brazil crowds attend funeral of late candidate Campos
  - Current event
- Mr Campos's jet crashed in bad weather in Santos
  - Background event
- Mr Campos's Socialist Party is expected to appoint...
  - Speculation
- President Dilma Rousseff
  - Property
- They attended a funeral Mass
  - Pronominal reference to an entity in a previous sentence

01.02 Examples of Text

NATURAL LANGUAGE PROCESSING

M















































01.03 Funny Sentences

NATURAL LANGUAGE PROCESSING

M

# NLP

1.4

Administration of the class

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

# NLP

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## Introduction to NLP

### Administrative

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## Structure of the Course

- Four major parts:
  - Linguistic, mathematical and computational background
  - Computational models of morphology, syntax, semantics, discourse, pragmatics
  - Core NLP technology: parsing, part of speech tagging, text generation, etc.
  - Applications: text classification, machine translation, information retrieval, etc.
- Three major goals:
  - Learn the basic principles and theoretical issues underlying natural language processing
  - Learn techniques and tools used to develop practical, robust systems that can understand text and communicate with users in one or more languages
  - Gain insight into some open research problems in natural language

of mathematics of linguistics

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## Available Books

- Speech and Language Processing**
  - Daniel Jurafsky and James Martin
  - <http://www.cs.colorado.edu/~mactn/book.html>
- Foundations of Statistical Natural Language Processing**
  - Chris Manning and Henry Schütze
  - <http://nlp.stanford.edu/book/>
- Natural Language Understanding**
  - James Allen

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## Courses at Other Places

- Brick-and-Mortar**
  - Johns Hopkins University (Jason Eisner)
  - Cornell University (Lillian Lee)
  - Stanford University (Chris Manning)
  - U. Maryland (Hal Daumé)
  - Berkeley (Dan Klein)
  - U. Texas (Ray Mooney)
- Coursera**
  - Manning/Jurafsky (2012, survey)
  - Michael Collins (2013, more advanced)

slides available

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## The Association for Computational Linguistics (ACL)



There's a lot of overlap between ACL & NLP

01.04 Administrative

NATURAL LANGUAGE PROCESSING

M

## The Alphabet Soup

- NLP (Natural Language Processing)
- CL (Computational Linguistics)
- IR (Information Retrieval)
- SP (Speech Processing)
- HLT (Human Language Technology)
- NLE (Natural Language Engineering)
- ML (Machine Learning)

have formal/semi rep. but applications first life from disambig. computational statistical learning















01.04 Administrative

Research in NLP

- Conferences: *statistical / information techniques*
- Journals: *publish translations*
- University centers: *check out state-of-art study*
- Industrial research sites: *best place to find academic journals*
- The ACL Anthology
- The ACL Anthology Network (AAN)

01.04 Administrative

1.5

NLP

01.05 Why is NLP Hard?

NLP

01.05 Why is NLP Hard?

01.05 Why is NLP Hard?

Introduction to NLP

Why is NLP hard?

01.05 Why is NLP Hard?

01.05 Why is NLP Hard?

Example

Time flies like an arrow.

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

01.05 Why is NLP Hard?

01.05 Why is NLP Hard?

Quiz Answer

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

01.05 Why is NLP Hard?

01.05 Why is NLP Hard?

More Classic Examples

- Beverly Hills
- The box is in the pen *word order matters*
- The pen is in the box
- Mary and Sue are mothers *more examples of ambiguity*
- Mary and Sue are sisters *ambiguities*
- Every American has a mother
- We gave the monkeys the bananas because they were hungry *what they? break?*
- We gave the monkeys the bananas because they were over-ripe *what over-ripe? banana*

01.05 Why is NLP Hard?

01.05 Why is NLP Hard?

Syntax vs. Semantics

\* Little a has Mary lamb.

? Colorless green ideas sleep furiously.

*not problems with the semantics ideas don't sleep, don't sleep furiously ideas -> green ideas cannot have color*

*syntax really well from*

*very subtle statement -> hard to get the semantics*

01.05 Why is NLP Hard?















01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### Ambiguous Words

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

? each word has multiple pronunciation

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### 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)

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### Ambiguity

- Not in computer languages (by design)!
- Or Lojban
- Noun-noun phrases: (XY)Z vs. X(YZ)
  - science fiction writer
  - customer service representative
  - state chess tournament

Describe a noun with another noun  
7.10.14/2018

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### NACLO Problems

- One Two Tree, by Noah Smith, Kevin Gimbel, and Jason Eisner
  - <http://www.naclo.cs.cmu.edu/problems2012/N2012-R.pdf>
- Fakepapershellmaker, by Willie Costello
  - <http://www.naclo.cs.cmu.edu/problems2008/N2008-F.pdf>

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### NACLO Problem Solutions

- One Two Tree
  - <http://www.naclo.cs.cmu.edu/problems2012/N2012-RS.pdf>
- Fakepapershellmaker
  - <http://www.naclo.cs.cmu.edu/problems2008/N2008-FS.pdf>

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### Types of Ambiguity 1/2

- 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.
- Pp attachment:
  - Joe sat on the bench with a fork / with moonbeams / with Samantha / with pleasure.
- Sense:
  - Joe took the bus exam.
- Modality:
  - Joe may win the lottery.

Joe means negative or like former, not the latter  
Joe may hedge between two possible outcomes

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### Types of Ambiguity 2/2

- Subjectivity:
  - Joe believes that stocks will rise.
- Cc attachment:
  - Joe likes red apples and pears.
- Negation:
  - Joe likes his pizza with no cheese and tomatoes.
- Referential:
  - Joe arrived at Mike's and kissed the bride.
- Reflexive:
  - Joe thought Mike present.
- Ellipsis and parallelism:
  - Joe gave Mike a beer and Jeremy a glass of wine.
- Metonymy:
  - Boston called and left a message for Joe.

ripe apple and pear? ripe apple and pear  
Joe of Mike  
reflexive pronoun  
In person it  
Boston: the office of Boston

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### Other Sources of Difficulties

- Non-standard, slang, and novel words and usages
  - ASAP, FYI, etc. – 1-644-551-2222
  - "span" or "round" as verbs
  - yo-yo, yeller, chiller – not in any recognized dictionary words
  - www.urbandictionary.com – (Parental Warning)
- Inconsistencies
  - junior college, college junior
  - pet spray, pet flame
- Typoes and grammatical errors
  - Receipt, pond/rockies, should of have
- Parsing problems
  - Cup holder – person who holds the cup: fool in the car
  - Federal Reserve Board Chairman

01.06 Why is NLP Hard?

NATURAL LANGUAGE PROCESSING

M

### 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? 2:30 p.m. 1:15 (time)
- Language is hard even for humans (both L1 and L2)

if you were ... then  
sarcasm will make sense only with shared knowledge















01.06 Why is NLP Hard?
NATURAL LANGUAGE PROCESSING
M

### Synonyms and Paraphrases

The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72, its best close since June 12, 2001.

The Nasdaq gained 12.22, or 0.56 percent, to 2,188.44 for its best showing since June 8, 2001.

The DJIA rose 68.46, or 0.64 percent, to 10,705.55, its highest level since March 15.

01.06 Why is NLP Hard?
NATURAL LANGUAGE PROCESSING
M

### Synonyms and Paraphrases

The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72, its best close since June 12, 2001.

The Nasdaq gained 12.22, or 0.56 percent, to 2,188.44 for its best showing since June 8, 2001.

The DJIA rose 68.46, or 0.64 percent, to 10,705.55, its highest level since March 15.

01.06 Why is NLP Hard?
NATURAL LANGUAGE PROCESSING
M

# NLP

01.06 Background
NATURAL LANGUAGE PROCESSING
M

# NLP

01.06 Background
NATURAL LANGUAGE PROCESSING
M

## Introduction to NLP

### Background

01.06 Background
NATURAL LANGUAGE PROCESSING
M

### Linguistic Knowledge

- **Constituents** — *parts of a sentence that has a specific role*
  - Children eat pizza.
  - They eat pizza.
  - My sister's neighbor's children eat pizza.
  - Eat pizza. → *the subject not explicit, pronounced in the sentence "who"*
- **Collocations** — *groups of words that go together more likely than expected by chance*
  - Strong beer but "powerful" beer
  - Big sister but "large" sister
  - Stocks rise but "stocks" descend → *the collocation change over time*
  - in the past: 2.75,000 hits vs. 47 hits on Google, now 510,000 vs 57,000
- **How to get this knowledge in the system:**
  - Manual entry — *by hand*
  - Automatically acquired from large text collections (corpora)

01.06 Background
NATURAL LANGUAGE PROCESSING
M

### Linguistic Knowledge

- **Knowledge about language:**
  - 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 — *is possible*
  - Pragmatics – how to accomplish goals
  - Discourse conventions – how to deal with units larger than utterances → *multi sentence paragraphs, between (later sentences) in the same paragraph all refer to the first sentence*
- **Separate lecture**

01.06 Background
NATURAL LANGUAGE PROCESSING
M

### Finite-state Automata

is theory application in NLP

easy thing

multiple paths

perform post-tagging: cat, dog → is, sign → ?

01.06 Background
NATURAL LANGUAGE PROCESSING
M

### Theoretical Computer Science

- **Automata**
  - Deterministic and non-deterministic finite-state automata
  - Push-down automata
- **Grammars**
  - Regular grammars
  - Context-free grammars
  - Context-sensitive grammars
- **Complexity** — *how long it takes...*
- **Algorithms**
  - Dynamic programming





























these languages have a common ancestor

- \* Cognates
  - English: nut (French), Nacht (German), nacht (Dutch), nag (Afrikaans), night (Scottish Gaelic), Nočevnik (Slovene), nait (Danish), nait (Faroese), nait (Icelandic), nait (Czech, Slovak, Polish), nox (nocturnal) (Latin), nocht (Macedonian), nosch (Bulgarian), nos, nish (Romanian), nose (Serbian), noc (Croatian), nox (Ancient Greek), vytra/nychta in Modern Greek), nox/noe (Latin), nok (Sanskrit), nach (Albanian), noche (Spanish), nos (Welsh), nosche (Austrian), note (Portuguese and Galician), notte (Italian), no (Arabic), nachts (Dutch), Noapte (Romanian), nakts (Lithuanian) and nakits (Dhivehi), all meaning "night" and derived from the Proto-Indo-European PIE \*nokʷs, "night".

From wikipedia



- **Altaic**
  - Turkish
- **Uralic (Finno-Ugric)**
  - Finnish
  - Hungarian
- **Semitic**
  - Arabic
  - Hebrew
- **Uto-Aztec**

[illegible]

Chinese

Etymology (708 languages)

- **Grimm's Law**
  - Voiceless stops turn into voiceless fricatives
  - Voiced stops become voiceless stops
  - Voiced aspirated stops change to voiced stops or fricatives
- **Example 1**
  - Ancient Greek: *πούς*; Latin: *pēs*; Sanskrit: *pāda*
  - English: *foot*; German: *Fuß*; Swedish: *för*
- **Example 2**
  - Ancient Greek: *κύων*; Latin: *canis*; Welsh: *ci*
  - English: *hound*; Dutch: *hond*; German: *Hund*

Change of language over time

voice  $\rightarrow$  voiceless

- **All in the Family**
  - <http://www.naclo.cs.cmu.edu/problems2012/N2012-D.pdf>

- All in the Family
  - <http://www.naclo.cs.cmu.edu/problems2012/N2012-DS.pdf>

- Can you guess the source, language, and period of this text?

















- Beowulf
- Epic poem
- 8<sup>th</sup>-11<sup>th</sup> Century
- Old English



large difference / modern English  
borrow a lot of words from other languages

Hwæt! We Gardena in gear-dagum,  
 þeod-cyninga, þrym gefrunon,  
 hu 8a aþelingas ellen fremedon.  
 Oft Scyld Scefing sceapena þreatas,  
 monnum magnum, meodoseta  
 oðrað,  
 egode eorlas. Syððan aereð wearð  
 feaceast farden, he þus trofre gebad,  
 weox under wolcnum, weorðmyndum  
 þah.  
 oððer him aeghwec þara weðisittend

Lo! the Spear-Danes' glory through splendid  
achievements  
The folk-kings' former fame we have heard of,  
How princes displayed then their prowess-in-battle  
Oft Scyld the Scefing from scathens in numbers  
From many a people their mead-benches tore.  
Since first he found him friendless and wretched,  
The earl had had terror: comfort he got for it,  
Waxed 'neath the welkin, world-honor gained,  
Till all his neighbors o'er sea were compelled to—

erst (as in *erscheine*) = first  
<http://lit.griwies.com/> [http://www5.psu.edu/~ehs/department/medieval\\_library/object/texts/4.1.html](http://www5.psu.edu/~ehs/department/medieval_library/object/texts/4.1.html)  
<http://www.gutenberg.org/files/16328/16328-h/16328-h.htm>

<http://www.gutenberg.org/files/16378/16378-h/16378-h.htm>  
<http://www.mcc.edu/home/poddk/Transition/beyond3.htm>

- Articles *the, a*
- Cases (e.g., in Latin)
  - Purely positional system
- Sound systems
  - Glottal stop (middle and end of "uh-oh") - pro
  - Velar fricatives - articulated with the back of the tongue at the soft palate
  - Voicelless /x/ - used e.g., in Russian
  - Voiced /r/ - used e.g., in Modern Greek
- Social status (e.g., in Japanese)
  - *otousan*, *otōsan* = someone else's father
  - *chichi*, *chi* = one's own father
- Kinship systems (e.g., in Warlpiri) - see next slide

words indicate which is subject  
for. The boy <sup>pro</sup> <sub>of the tongue at the soft</sub> <sup>not pronounce</sup> <sub>languages have 200</sub> ways to refer to the relationship

- Warlpiri Kinship - by Alan Chang *Relationship of different languages.*  
- <http://www.naclo.cs.cmu.edu/pdf-split/N2013-0.pdf>

Relationship of different languages.

- Warlpiri Kinship
  - <http://www.naclo.cs.cmu.edu/pdf-split/N2013-05.pdf>

- Warlpiri Kinship
  - <http://www.naclo.cs.cmu.edu/pdf-split/N2013-05.pdf>

- Two types
  - unconditional
  - conditional
- Examples
  - All languages have verbs and nouns.
  - All spoken languages have consonants and vowels.
  - [Greenberg 19] "In declarative sentences with a nominal subject and object, the dominant order is almost always one in which the subject precedes the object."
  - [Greenberg 29] "If a language inflects, it always has derivation."

All languages have verbs and nouns.

Inflection: change in position  
change in meaning  
change in position

have verbs and nouns

languages have consonants and vowels

In declarative sentences with Nominial subject dominant order is almost always one in which cedes the object.

Derivation: divide - derivable

"If a language has inflection, it always has the changing women derivation"

can change a word by changing morphology

- <http://wals.info>
- **Feature 83A: Order of Object and Verb**
  - by Matthew S. Dryer
  - OV (713 languages), VO (705), no dominant order (101)
  - <http://wals.info/feature/83A#2/18.0/152.9>
- **Other features:**
  - 18A Absence of common consonants (by Ian Maddieson): no bilabials (5 languages), no fricatives (49), no nasals (12)
  - 67A **Inflectional feature** tense (by Osten Dahl, Viveka Velupillai): yes (110), no (112)

*I will ...*  
*but some Latin language use inflection of verbs. Not French.*

- <http://wals.info>
- **Feature 83A: Order of Object and Verb**
  - by Matthew S. Dryer
  - OV (713 languages), VO (705), no dominant order (101)
  - <http://wals.info/feature/83A#2/18.0/152.9>
- **Other features:**
  - 18A Absence of common consonants (by Ian Maddieson): no bilabials (5 languages), no fricatives (49), no nasals (12)
  - 67A **Inflectional feature** tense (by Osten Dahl, Viveka Velupillai): yes (110), no (112)

*I will ...*  
*but some Latin language use inflection of verbs. Not French.*

- **Ethnologue** *cover lots of languages*  
- <http://www.ethnologue.com/>
- **Number words in many languages**  
- <http://www.zompist.com/numbers.shtml>
- **Endangered languages** *language disappearing*  
- <http://www.endangeredlanguages.com/>
- **Google fights to save 3,054 dying languages** *etc.*  
- <http://www.cnn.com/2012/06/21/tech/web/google-fights-save-languages-mashable/index.html>

- **Ethnologue** *cover lots of languages*  
- <http://www.ethnologue.com/>
- **Number words in many languages**  
- <http://www.zompist.com/numbers.shtml>
- **Endangered languages** *language disappearing*  
- <http://www.endangeredlanguages.com/>
- **Google fights to save 3,054 dying languages** *etc.*  
- <http://www.cnn.com/2012/06/21/tech/web/google-fights-save-languages-mashable/index.html>













18















