Statistical Learning for Text Data Analytics
Lecture 1: Introduction

Yangqiu Song

Hong Kong University of Science and Technology

yqsong@cse.ust.hk

Spring 2018

$*$Contents are based on materials created by Chris Manning, Percy Liang, Hongning Wang, and Haixun Wang

Reference Content

- Chris Manning. CS 224N/Ling 237. Natural Language Processing. https://web.stanford.edu/class/cs224n/
- Percy Liang. ICML tutorial on Natural Language Understanding: Foundations and State-of-the-Art https://icml.cc/2015/tutorials/icml2015-nlu-tutorial.pdf
- Hongning Wang. CS6501 Text Mining. http://www.cs.virginia.edu/~hw5x/Course/Text-Mining-2015-Spring/_site/

Overview

1. Logistics
2. Introduction to NLP
   - Why is NLP Important?
   - Machine Learning for NLP: Algorithms, Tasks, and Challenges

Logistics

- Instructor: Yangqiu Song
- Email: `yqsong@cse.ust.hk`
- Office: RM3518 (Lift25/26)
- Canvas (`https://canvas.ust.hk`)
- No class meeting on Feb. 14th
  - Will have a make up session in March
- For CSE students, this course does not apply for the requirement: "The 3 credits may be satisfied by courses from other Schools"

Course Information

- Weekly reading notes (40%): one paper per week, related to the lectures
- Mid-term project proposal: title and abstract (10%):
    - Could be a discussion paper for Math students or a project for CSE students
    - A particular research problem, e.g., structural output learning
    - A particular mathematical challenge, e.g., variance-reduced gradient descent, black-box variational inference for probabilistic topic model
    - An application: sequence tagging, sentiment analysis, etc.
    - Investigate an algorithm, e.g., deep learning (RNN)
- Project report (30%)
- Final project presentation(20%)

Topic Covered

- Text modeling: language models, distributed representations
- Document classification: supervised learning, semi-supervised learning
- Topic models: SVD, probabilistic models
- Word tagging: sequence models, constrained models, posterior regularization
- Deep learning

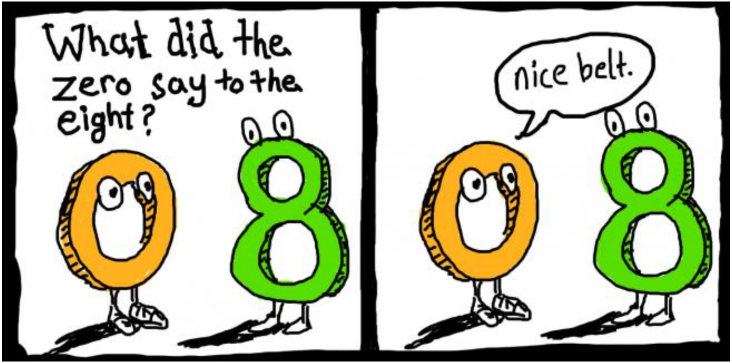
Overview

1. Logistics

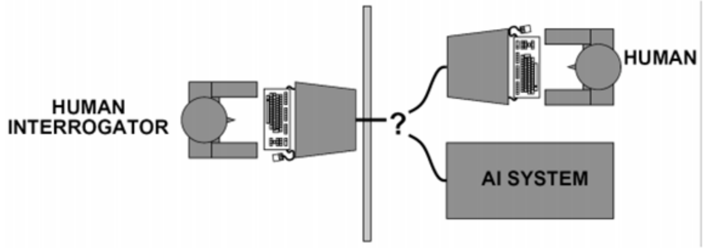
2. Introduction to NLP
   - Why is NLP Important?
   - Machine Learning for NLP: Algorithms, Tasks, and Challenges

Natural Language

- Understanding language is a very complex thing
- But something that humans are amazingly good at



Artificial Intelligence: Turing Test (1950)

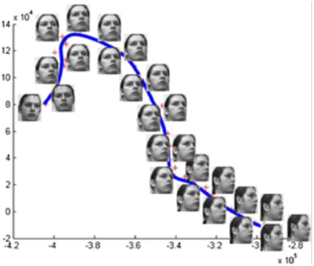


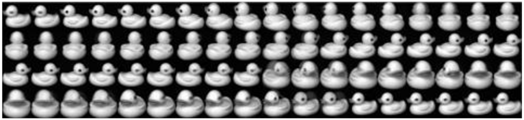
- Replacement of “Can machines think?”
  - “Can machines behave intelligently?”
  - Operational test for intelligent behavior: the Imitation Game (later dubbed the Turing test)
  - Suggested major components of AI: knowledge, reasoning, language understanding, learning

The AI Winter

- AI winter: 1974-80 and 1987-93
  - 1966: the failure of machine translation,
  - 1970: the abandonment of connectionism,
  - 1971-75: DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University,
  - 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill report,
  - 1973-74: DARPA's cutbacks to academic AI research in general,
  - 1987: the collapse of the Lisp machine market,
  - 1988: the cancellation of new spending on AI by the Strategic Computing Initiative,
  - 1993: expert systems slowly reaching the bottom, and
  - 1990s: the quiet disappearance of the fifth-generation computer project's original goals.

What's Special about Human Language?

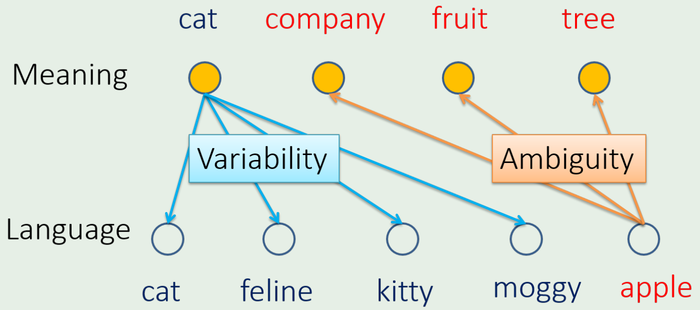




- A human language is a discrete/symbolic/categorical signaling system
- With very minor exceptions for expressive signaling ("I loooove it." "Whoomppaaa")
- Large vocabulary, symbolic encoding of words creates a problem for machine learning – sparsity!

Why is NLP Difficult?

Example (variability and ambiguity everywhere)



Why is NLP Difficult?

Example (“Get the cat with the gloves.”)

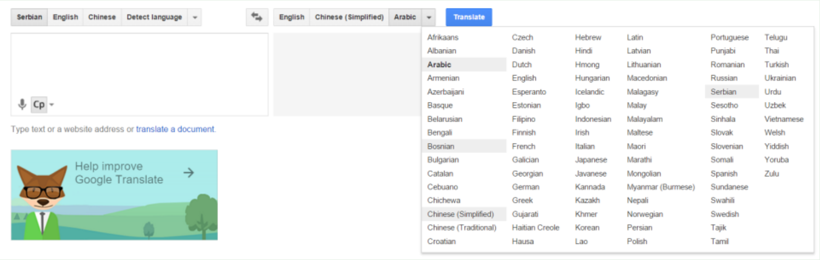
Texts in the Era of Big Data

- Huge in size
  - Google processes 5.13B queries/day (2013)
  - Twitter receives 340M tweets/day (2012)
  - Facebook has 2.5 PB of user data + 15 TB/day (4/2009) ($1PB=10^{15}$bytes=1000terabytes)
  - eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 80% data is unstructured (IBM, 2010)
  - Traditional NLP techniques (e.g., parsing) are too slow to handle them
  - Traditional NLP models are based on labeled data in specific domains (WSJ data)

NLP Enabled by Big Data

Example (Google Translate)

- 1966: the failure of machine translation
- Now: Google Translate can work with more than 100 languages



NLP Enabled by Big Data

Example (Facebook Translation)

NLP Enabled by Big Data

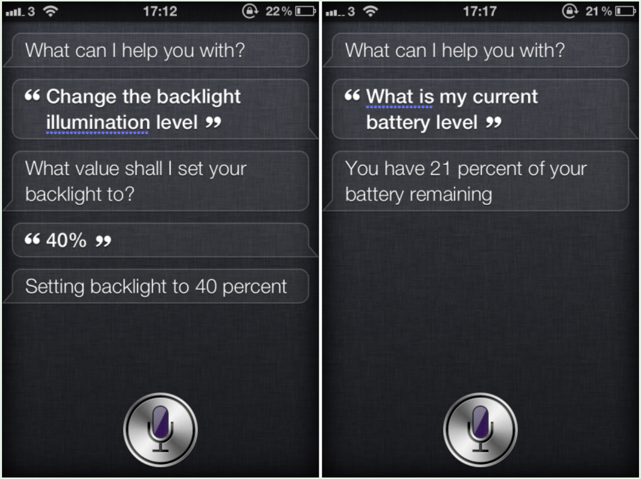
Example (IBM's Watson)

- 1971–75:DARPA's frustration with the Speech Understanding
- Now: "Watson is aquestion answering (QA) computing system that IBM built to apply advanced
  - natural language processing,
  - information retrieval,
  - knowledge representation,
  - automated reasoning, and
  - machine learning technologies
- to the field of open domain question answering."

In 2011, Watson competed on Jeopardy! against former winners Brad Rutter and Ken Jennings. Watson received the first place prize of $1 million.

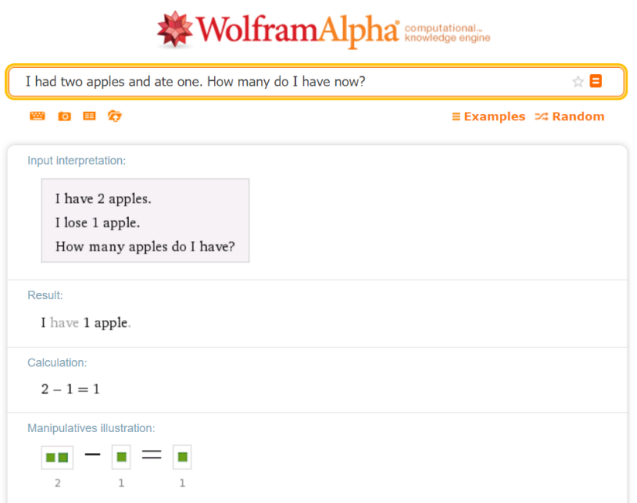
NLP Enabled by Big Data

Example (Apple's Siri)

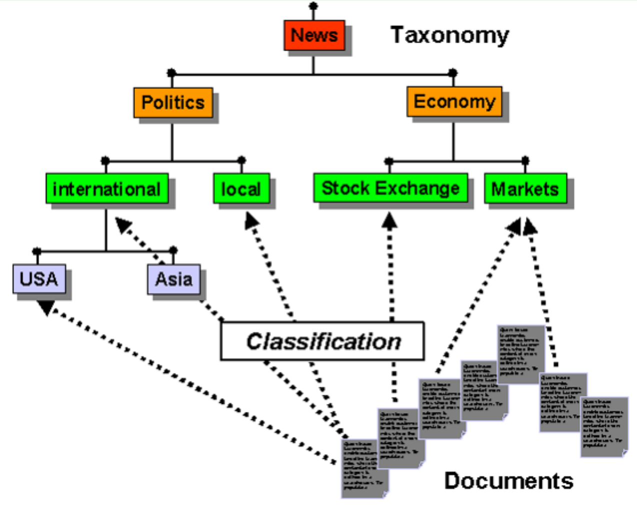


NLP Enabled by Big Data

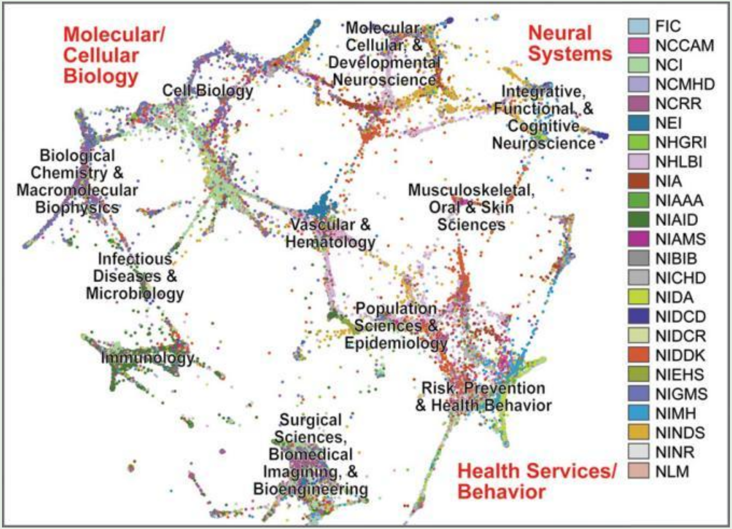
Example (WolframAlpha Knowledge Powered QA)



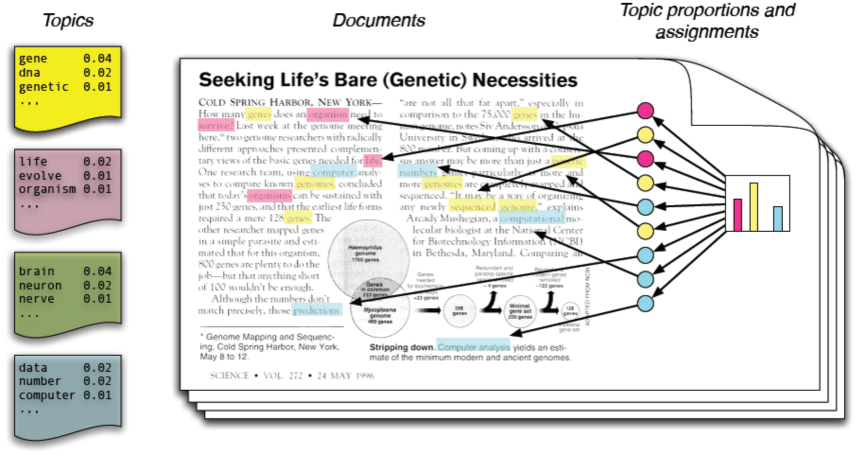
Example (Document categorization)



Example (Document categorization)

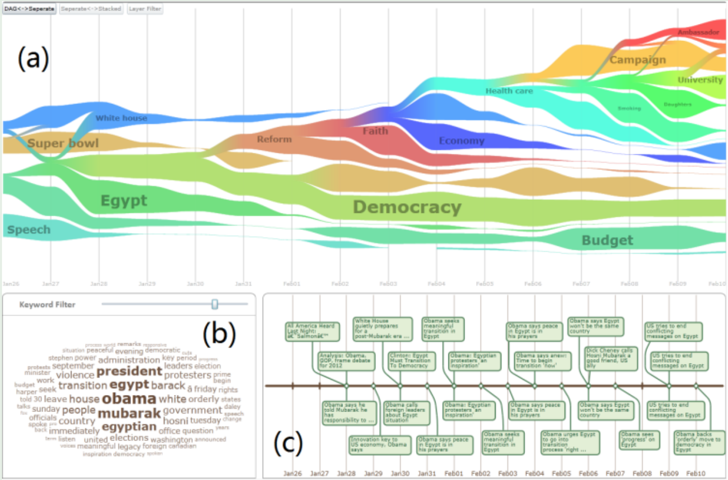


Example (Topic models)

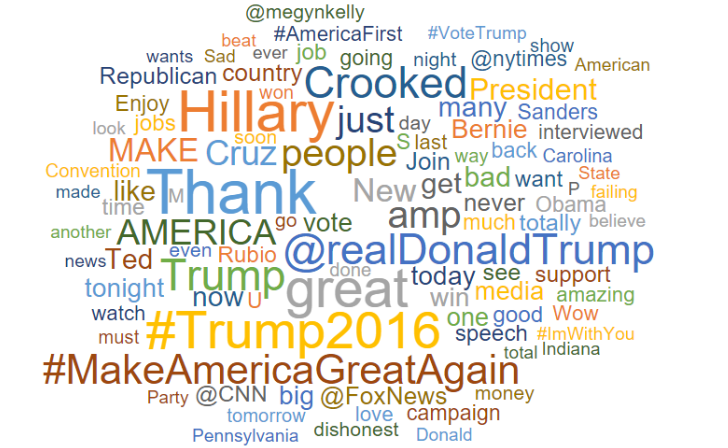


Text Mining in the Era of Big Data

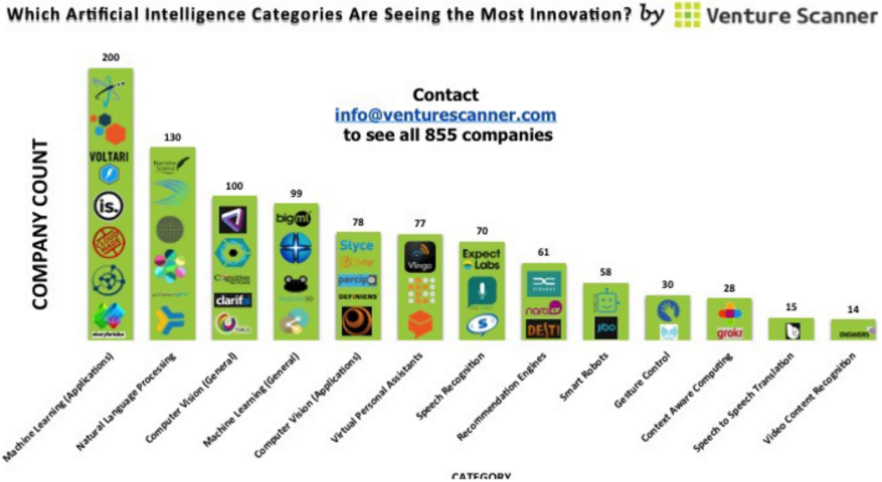
Example (Time line analysis)



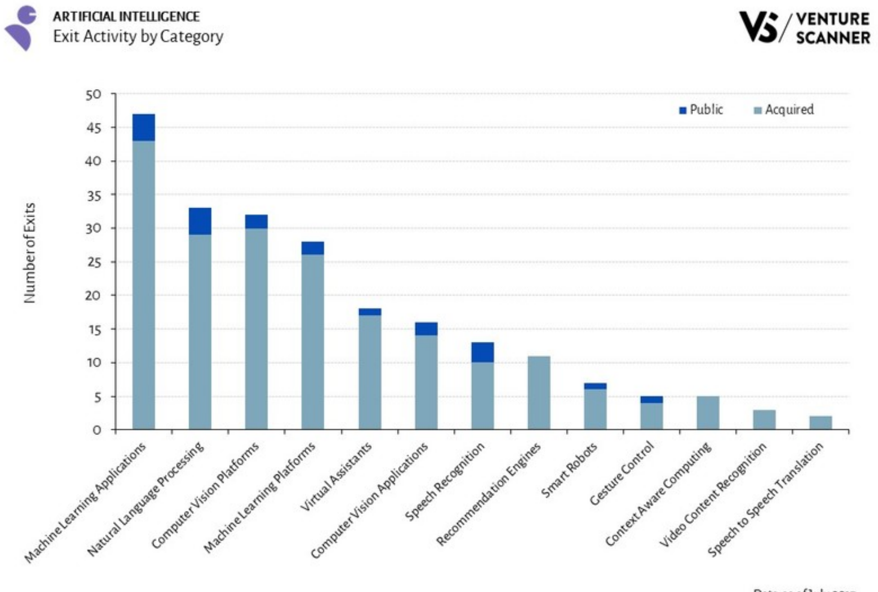
Example (Sentiment analysis)



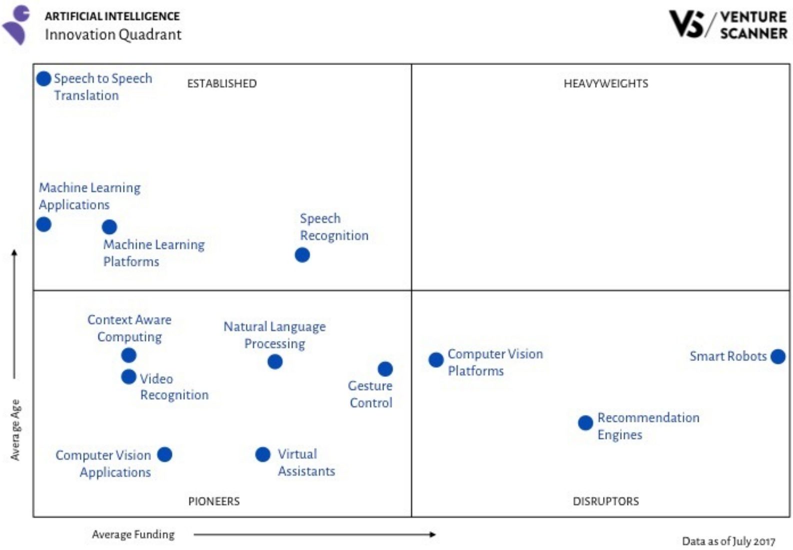
Startup Companies (2015)



Number of Exits (Acquisitions and IPOs, 2017)



Funding Size vs. Company Age (2017)



Overview

Statistical Machine Learning

- Natural Language Processing
  - Natural Language Understanding (NLU)
  - Natural Language Generation (NLG)
- Machine learning has been widely used in both NLU and NLG
  - given that we have a lot of data now

Popular Statistical Machine Learning Algorithms for NLP

- Mid-1970s: Hidden Markov Models (HMMs) for speech recognition $\rightarrow$ probabilistic models
- Early 2000s: Conditional Random Fields (CRFs) for part-of-speech tagging $\rightarrow$ structured prediction
- Early 2000s: Latent Dirichlet Allocation (LDA) for modeling text documents $\rightarrow$ topic modeling
- Mid 2010s: sequence-to-sequence models for machine translation $\rightarrow$ Deep Learning neural networks with memory/state
- Now: ??? for natural language understanding/generation
    - Reinforcement learning?

NLP Tasks: Levels of Linguistic Analysis

Morphology: basic unit of words	⇦ naming your world
Syntax: what is grammatical?	⇦ no compiler errors
Semantics: what does it mean?	⇦ no implementation bugs
Pragmatics: what does it do?	⇦ implemented the right algorithm

Analogy with Programming Languages

- Syntax: no compiler errors
- Semantics: no implementation bugs
- Pragmatics: implemented the right algorithm

- Different syntax, same semantics (5):

$$2 + 3 \Leftrightarrow 3 + 2$$

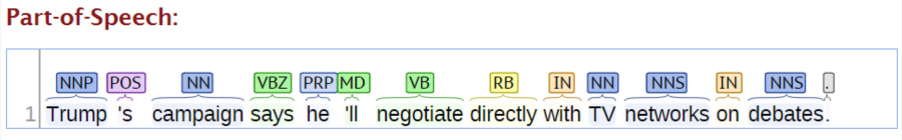
- Same syntax, different semantics (1 and 1.5):

$$3 \ / \ 2 \text{ (Python 2.7)} \not\Leftrightarrow 3 \ / \ 2 \text{ (Python 3)}$$

- Good semantics, bad pragmatics:

correct implementation of deep neural network
for estimating coin flip prob.

Example (Part of speech)



Tags:

- NN: common noun
- NNP: proper noun
- VB: verb, base form
- VBZ: verb, 3rd person singular
- ...

Syntax (1): Part of Speech

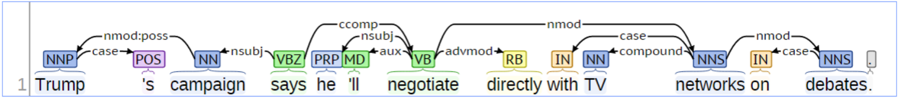
Penn Treebank part-of-speech tags (including punctuation).

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*'s*	”	Right quote	(’ or ”)
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Syntax (2): Dependency Parse Tree

Example (Dependency parse)

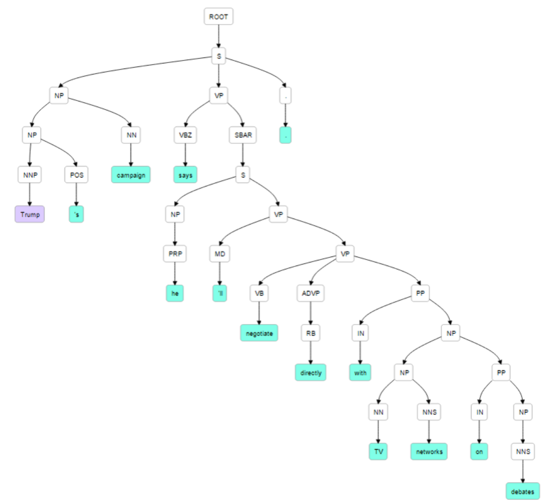
Basic Dependencies:



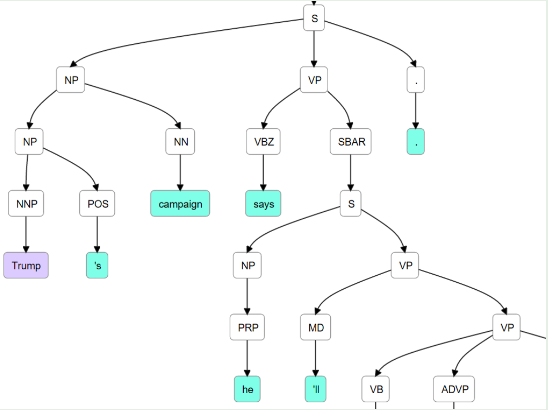
Dependency relations:

- nsubj: subject (nominal)
- advmod: adverbial modifier
- ...

Example (Constituency parsing)



Example (Constituency parsing)



- POS: possessive ending
- PRP: personal pronoun
- MD: modal; can, should

Semantics

- Syntax: no compiler errors
- Semantics: no implementation bugs
- Pragmatics: implemented the right algorithm

- Semantics: meanings
  - Lexical semantics: what words mean
  - Compositional semantics: how meaning gets combined

Semantics (1): What's a Word?

Example

Words

light

Multi-word expressions: meaning unit beyond a word

light bulb

Morphology: meaning unit within a word

light lighten lightening relight

Polysemy: one word has multiple meanings (word senses)

- The light was filtered through a soft glass window.
- He stepped into the light.
- This lamp lights up the room.
- The load is not light.

Semantics (1): Synonymy

Example (Synonymy)

Words:

confusing unclear perplexing mystifying

Sentences:

- I have fond memories of my childhood.
- I reflect on my childhood with a certain fondness.
- I enjoy thinking back to when I was a kid.

Beware: no true equivalence due to subtle differences in meaning; think distance metric

But there's more to meaning than similarity...

Other Lexical Relations

Hyponymy (is-a):

a cat is a mammal

Meronomy (has-a):

a cat has a tail

Useful for entailment:

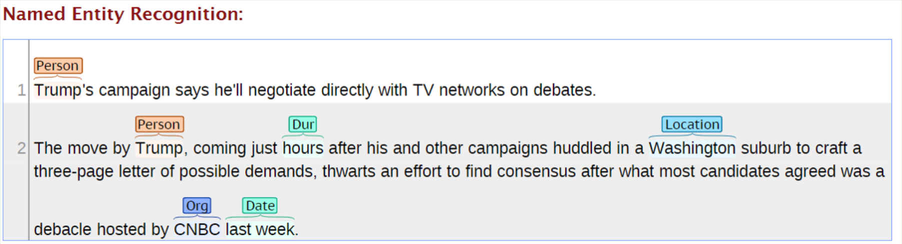
Alice is 170cm high and Bob is 180cm high.

$$\Rightarrow$$

Bob is taller than Alice.

Semantics (2): Named Entities

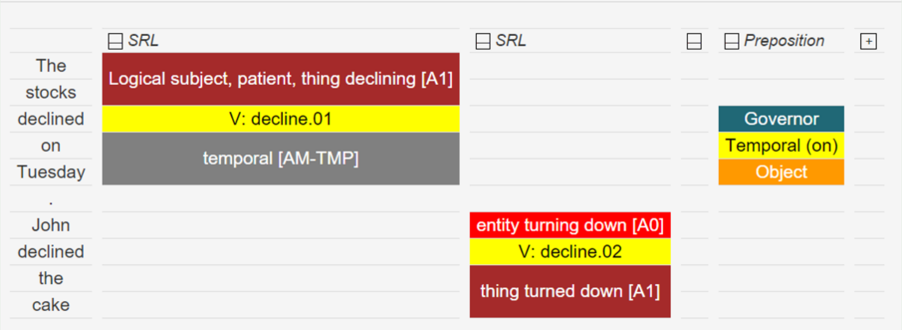
Example (Named Entity Recognition)



- Pers: Person
- Location
- Org: Organization
- Date/time

Semantics (3): Frame based Semantics

Example (Semantic Role Labeling)



- Predicates
- Arguments
- Senses

Semantics (4): Topics

Example (Topics)

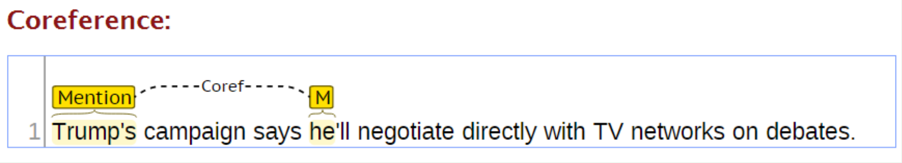
Trump's campaign says he'll negotiate directly with TV networks on debates. The move by Trump, coming just hours after his and other campaigns huddled in a Washington suburb to craft a three-page letter of possible demands, thwarts an effort to find consensus after what most candidates agreed was a debacle hosted by CNBC last week.

Category 1	politics
Category 2	entertainment

- Classification
- Clustering
- Topic modeling

Discourse

Example (General Coreference Problem (Pronoun Resolution))



"The Winograd Schema Challenge" (Levesque, 2011)

- The dog chased the cat, which ran up a tree. It waited at the top.
- The dog chased the cat, which ran up a tree. It waited at the bottom.
- Paul tried to call George on the phone, but he wasn't successful.
- Paul tried to call George on the phone, but he wasn't available.

Easy for humans, can't use surface-level patterns

Discourse

Example (Shallow Discourse Parser for Document-level Analysis)

- S1: Kemper is the first firm to make a major statement with program trading.
- S2: He added that "having just one firm do this isn't going to mean a hill of beans."

We can add a connective "but" between to above two sentences to indicate "Contrast relationship"

- S1: Senator calls this "the first gift of democracy."
- S2: The Poles might do better to view this as a Trojan Horse.

Pragmatics

Conversational implicature: new material suggested (not logically implied) by sentence

Example (Conversational implicature)

- A: What on earth has happened to the roast beef?
- B: The dog is looking very happy.
- Implicature: The dog at the roast beef.

Presupposition: background assumption independent of truth of sentence

Example (Presupposition)

- I have stopped eating meat.
- Presupposition: I once was eating meat.

Pragmatics

Semantics: what does it mean literally?
Pragmatics: what is the speaker really conveying?

- Underlying principle (Grice, 1975): language is cooperative game between speaker and listener
- Implicature and presupposition depend on people and context and involve soft inference (machine learning opportunities here!)

We need a lot of background knowledge and commonsense knowledge!

More about "Commonsense Knowledge"

When we communicate,

- we omit a lot of "common sense" knowledge, which we assume the hearer/reader possesses
- we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve

Knowledge about the everyday world that is possessed by all people

Example (Commonsense Knowledge)

- A lemon is sour.
- To open a door, you must usually first turn the doorknob.
- If you forget someones birthday, they may be unhappy with you.
- A coat is used for keeping warm.
- People want to be respected.
- The last thing you do when you cook dinner is wash your dishes.
- People want good coffee.

Commonsense Knowledge in Sentiment Analysis

Example (Sentiment Analysis)

Figure 2. Empathy Buddy Reacts to an E-mail.

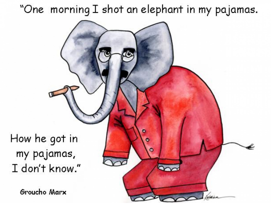
More about Ambiguity

Ambiguity: more than one possible (precise) interpretations

One morning I shot an elephant **in** my pajamas.

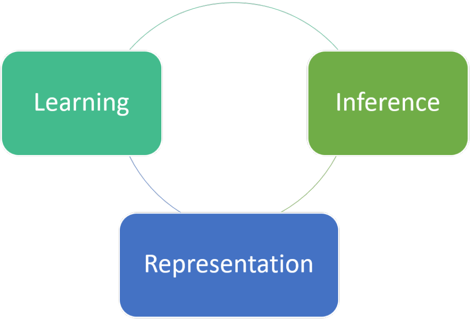
- "One morning I was wearing my pajamas, and I shot an elephant." or
- "One morning, an elephant was wearing my pajamas, and I shot that elephant."

How he got in my pajamas, I don't know. — Groucho Marx



- The joke is based on misdirection, where the listener thinks one thing, and the teller says another

Course Organization



- Representation: language models, word embeddings, topic models
- Learning: supervised learning, semi-supervised learning, sequence models, deep learning, optimization techniques
- Inference: constraint modeling, joint inference, search algorithms

Applications: tasks introduced above

Summary

1. Logistics
2. Introduction to NLP
   - Why is NLP Important?
   - Machine Learning for NLP: Algorithms, Tasks, and Challenges

In this class, we will

- Understand the intuition and motivation of how to model text data
- Know popular and state-of-the-art statistical models for NLP
- Build relationships of different algorithms