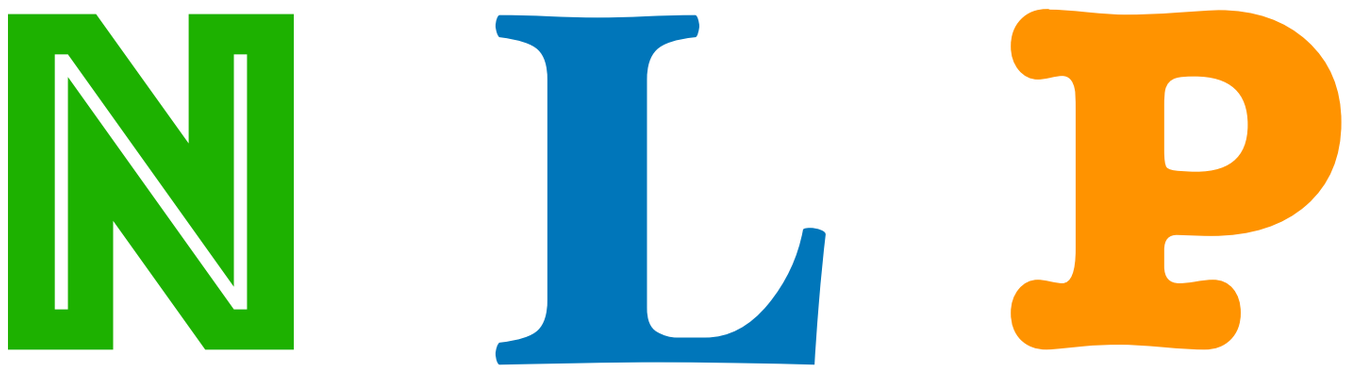
Natural Language Processing

Lecture 1: Course Overview and Introduction.

9/4/2018



COMS W4705
Daniel Bauer

The 4705 Team

- **Instructor:** Daniel Bauer <**bauer@cs.columbia.edu**>
  Office Hours: Mon 2:00pm-3:30pm
  (starting next week)
  704 Shapiro CEPSR

- **Assistants:**
  - Himanshu Aggarwal <ha2467@columbia.edu>
  - Wanlin Xie <wx2161@columbia.edu>
  - Shreya Jain <sj2842@columbia.edu>
  - Robert Netzorg <rgn2112@columbia.edu>

- **IA office hours / recitations start next week.**
  Time/Location TBA by email.

Lectures & Recitation Sessions

- **Lectures:** Tue & Thu 5:40pm-6:55am
  310 Fayerweather Hall

- **Recitation Sessions:**

  - Optional recitation sessions, led by the IAs (schedule TBA)

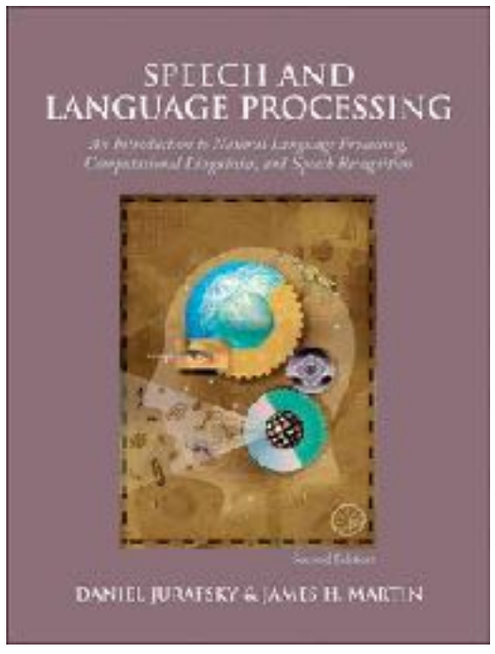
Course Resources

- **Courseworks 2 (a.k.a Canvas):**
  - All course materials: Lecture notes, code, announcements, assignments, reading materials
  - Homework submission, grade book.
- **Piazza** used for Q & A.
  Do not email the instructor or IAs with questions about the course content.

Textbook / Reading

- There is **NO official textbook** for this course.

- Recommended textbook (somewhat outdated, we won't follow too closely):



Dan Jurafsky & James Martin
Speech and Language Processing
2nd Ed. Prentice Hall (2009).

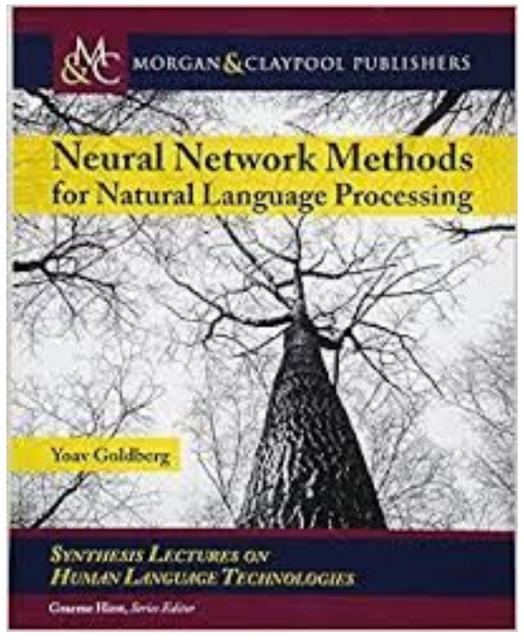
- Draft of most 3rd edition chapters:
https://web.stanford.edu/~jurafsky/slp3/

- We will also read a number of research papers.

Textbook / Reading

- Recommended textbook (mostly relevant later in the course):

Yoav Goldberg
Neural Network Methods for Natural Language Processing
Morgan & Claypool. 2017



- Available as an ebook through the CU library https://clio.columbia.edu/catalog/13420294

Prerequisites

- Data Structures (COMS W3134 or COMS W3137)
- Discrete Math (COMS W3202, recommended)
- Some previous or concurrent exposure to AI and machine learning is beneficial, but not required.
- Some experience with basic probability/statistics.
- Some experience with Python is helpful.

Grading

- Midterm 20%
- Final 30%
- 5 Homework assignments, each contains an analytical and a programming part, 10% each

Homework

- Homework uploaded through Courseworks. Do not email!
- Analytical part: Must be a plain txt or pdf documents (give LaTeX a shot).
- Programming part: We will use Python 3.

Homework Late Policy

- Written homework and programming problems may be submitted up to four days late for a 20 point penalty.

- No homework will be accepted more than four days after the deadline.

- Other extensions will only be granted in exceptional circumstances.

Academic Honesty

- Submit your own answers and code.
- Review academic honesty policy on the syllabus (Courseworks).
- When in doubt, ask.
- When in trouble, ask for help (and early).

NLP in the Movies

Natural Language Processing

- Important and active research area within AI.

- Timely: Most of our activities online are text based (web-pages, email, social media, blogs, news, product descriptions and reviews, medical reports, course content, …)

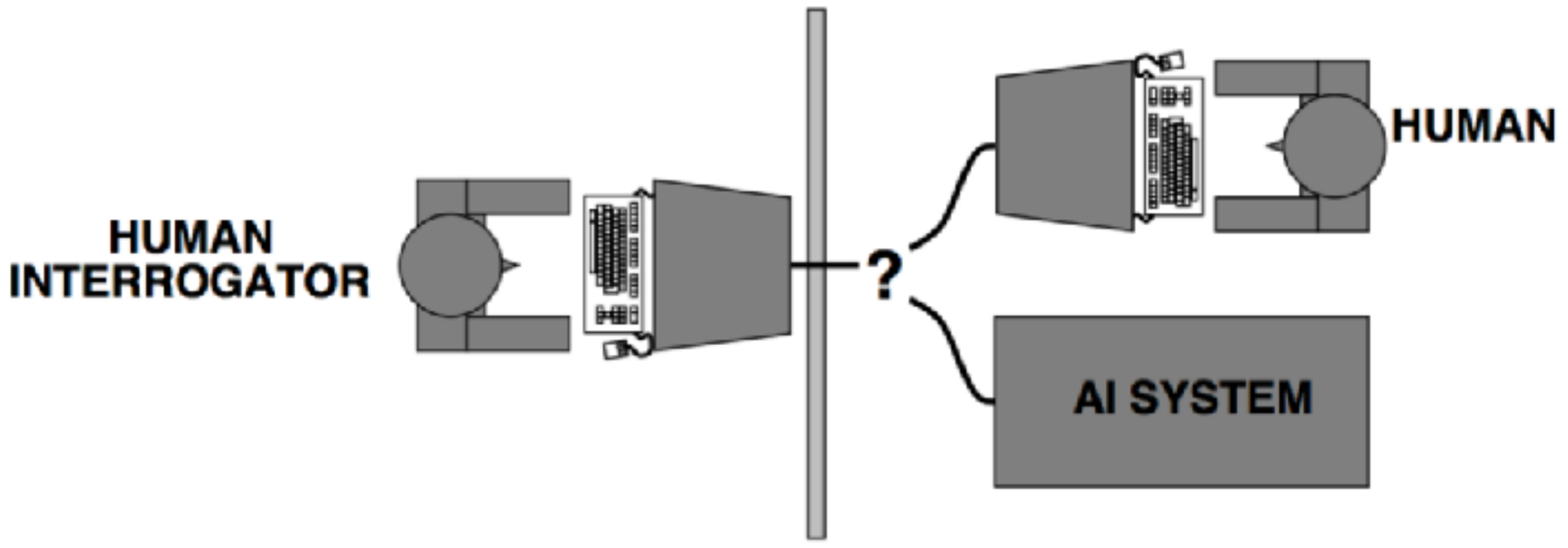
- NLP leverages more and more available training data and modern Machine Learning techniques.

- Communicating with computers is the “holy grail” of AI.

Turing Test

(Alan Turing, 1950)

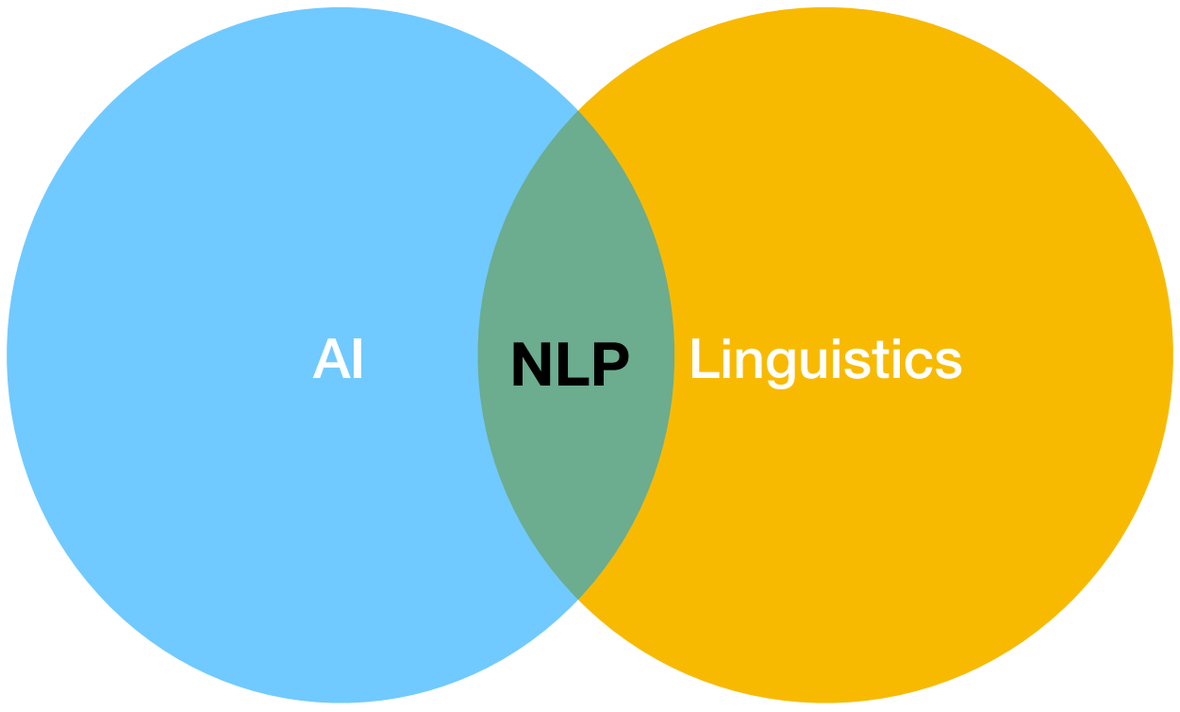
- A computer passes the test of intelligence if it can fool a human interrogator into believing it is human.



- What skills are needed to build such a system?
    - **Language processing,** knowledge representation, reasoning, learning.

Image source: Russel & Norvig, Artificial Intelligence - A Modern Approach

Natural Language Processing



“Every time I fire a linguist, my performance goes up” (Fred Jelinek)

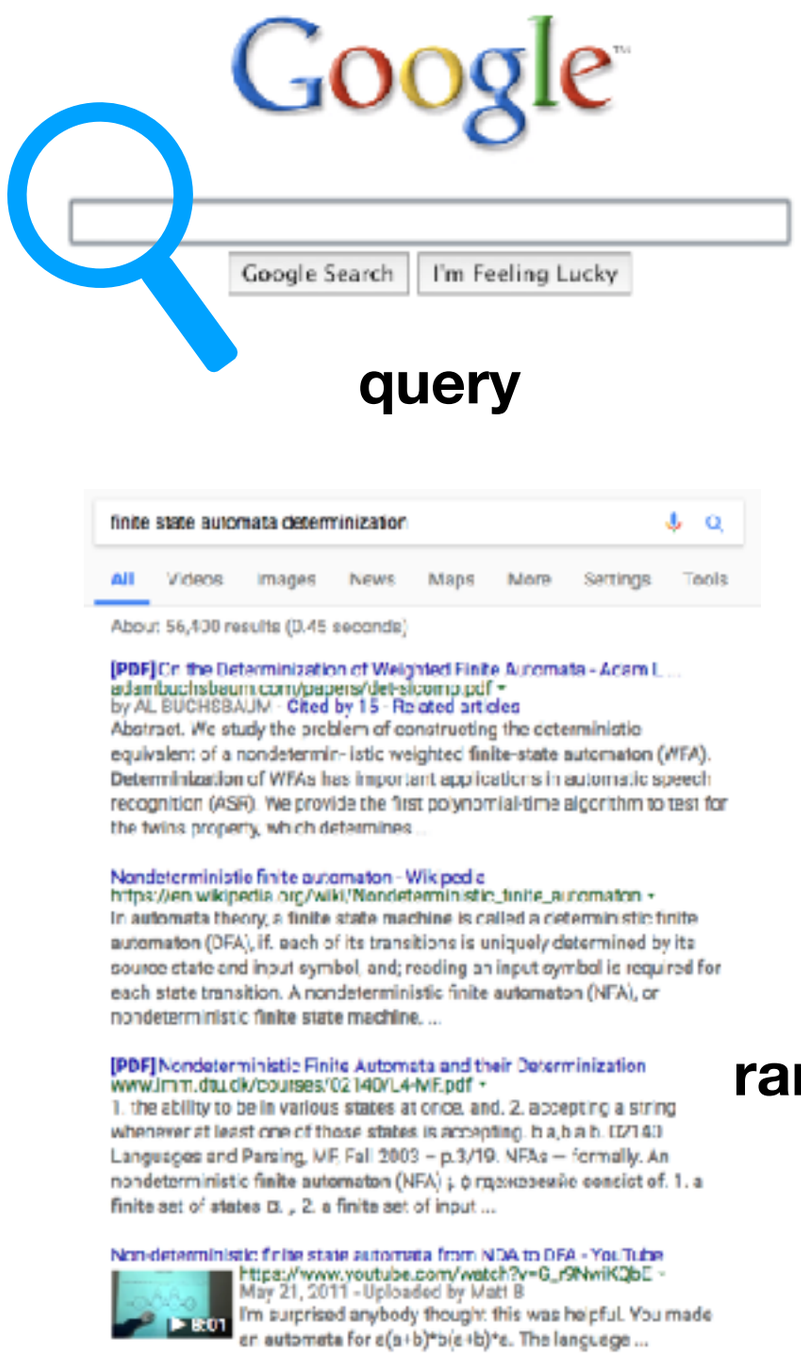
Natural Language Processing vs. Computational Linguistics

- **NLP:** Build systems that can understand and generate natural language. Focus on applications.

- **Computational Linguistics:** Study human language using computational approaches.

- Many overlapping techniques.

Applications: Information Retrieval



query

indexed document corpus

ranked results

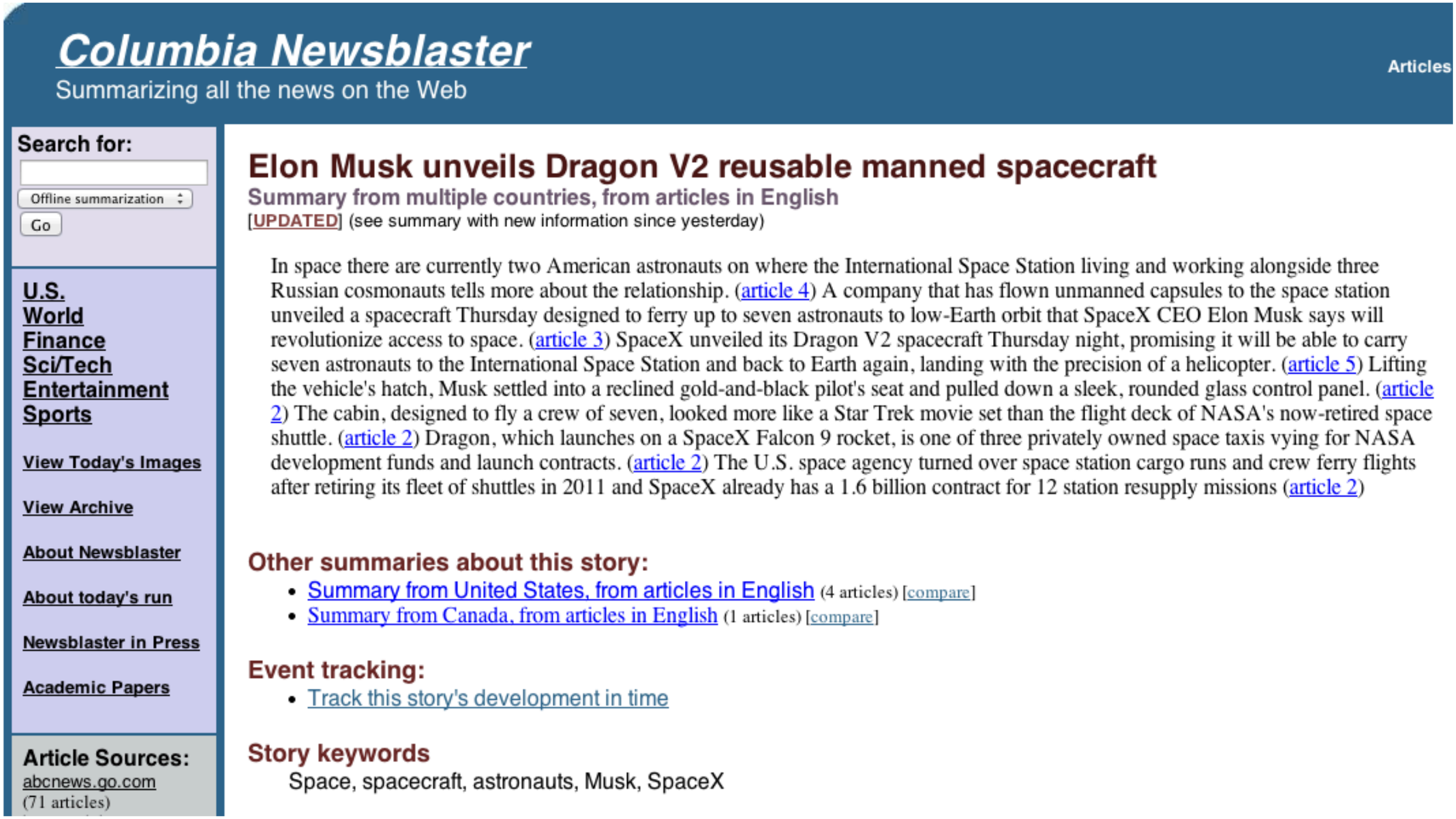
Applications: Text Classification

- Spam filtering.
- Detecting topics / genre.
- Sentiment analysis, author recognition, forensic linguistics, …

Applications: Sentiment Analysis

Fantastic... truly a wonderful family movie	★★★★★
I have a mixed feeling about this movie.	★★★
Well it is fun for sure but definitely not appropriate for kids 10 and below	★★★
My kids loved it!!	★★★★★
The movie is very funny and entertaining. Big A+	★★★★★
I got so booooored...	★
Disappointed. They showed all fun details in the trailer	★★
Cute but not for adults	★★★

Applications: News Summarization



Columbia Newsblaster
Summarizing all the news on the Web

Articles

Search for:
Offline summarization
Go

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images
View Archive
About Newsblaster
About today's run
Newsblaster in Press
Academic Papers

Article Sources:
abcnews.go.com
(71 articles)

Elon Musk unveils Dragon V2 reusable manned spacecraft

Summary from multiple countries, from articles in English
[UPDATED] (see summary with new information since yesterday)

In space there are currently two American astronauts on where the International Space Station living and working alongside three Russian cosmonauts tells more about the relationship. (article 4) A company that has flown unmanned capsules to the space station unveiled a spacecraft Thursday designed to ferry up to seven astronauts to low-Earth orbit that SpaceX CEO Elon Musk says will revolutionize access to space. (article 3) SpaceX unveiled its Dragon V2 spacecraft Thursday night, promising it will be able to carry seven astronauts to the International Space Station and back to Earth again, landing with the precision of a helicopter. (article 5) Lifting the vehicle's hatch, Musk settled into a reclined gold-and-black pilot's seat and pulled down a sleek, rounded glass control panel. (article 2) The cabin, designed to fly a crew of seven, looked more like a Star Trek movie set than the flight deck of NASA's now-retired space shuttle. (article 2) Dragon, which launches on a SpaceX Falcon 9 rocket, is one of three privately owned space taxis vying for NASA development funds and launch contracts. (article 2) The U.S. space agency turned over space station cargo runs and crew ferry flights after retiring its fleet of shuttles in 2011 and SpaceX already has a 1.6 billion contract for 12 station resupply missions (article 2)

Other summaries about this story:

- Summary from United States, from articles in English (4 articles) [compare]
- Summary from Canada, from articles in English (1 articles) [compare]

Event tracking:

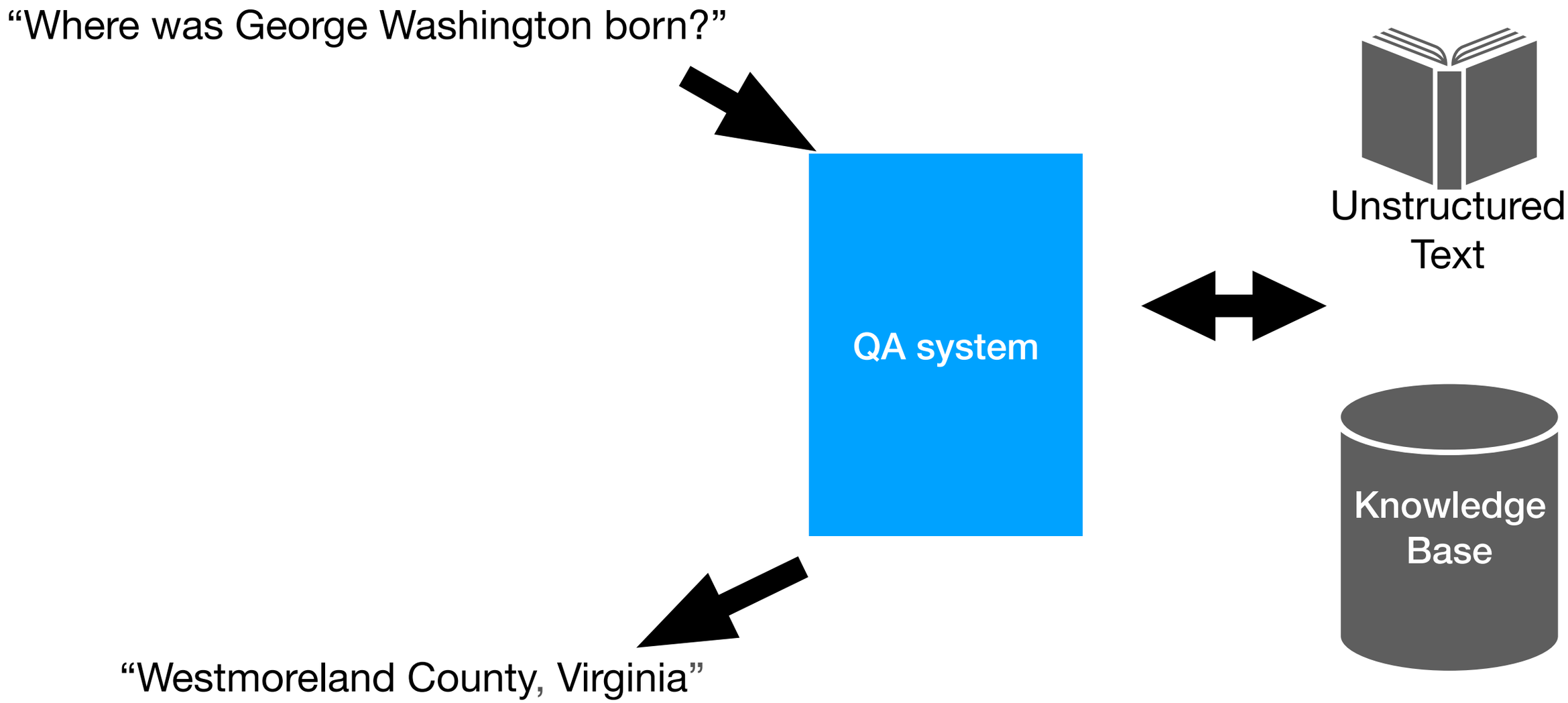
- Track this story's development in time

Story keywords

Space, spacecraft, astronauts, Musk, SpaceX

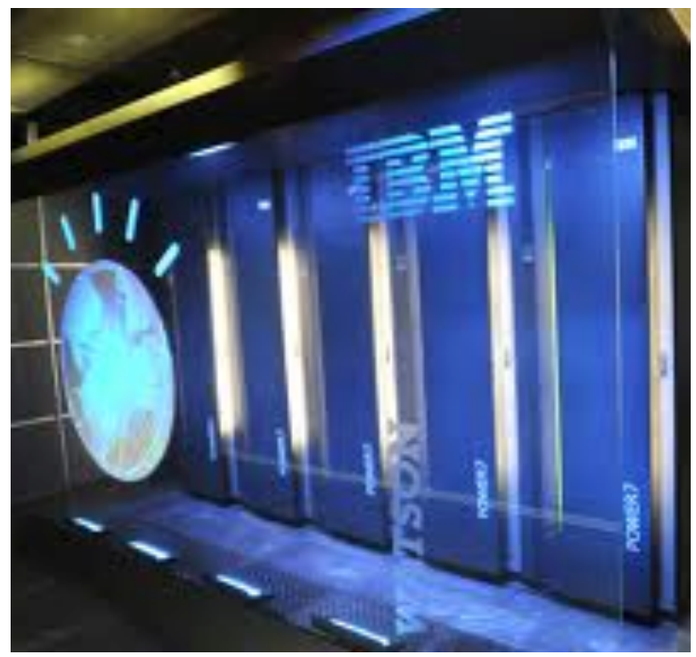
Credit: Prof. Kathleen Mc Keown

Application: Question Answering

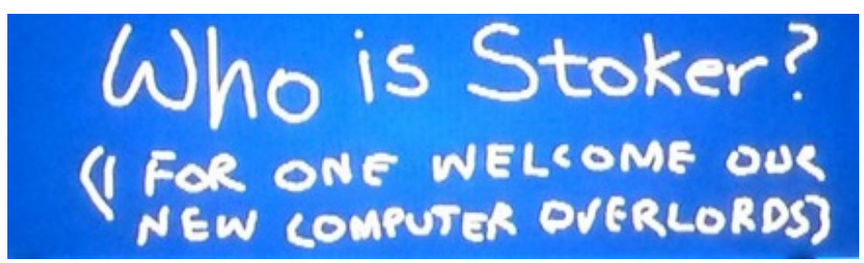


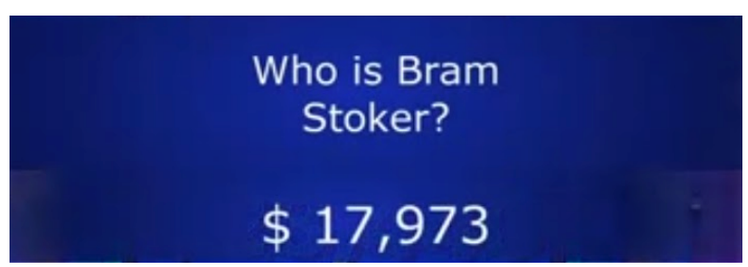
Applications: Playing Jeopardy!

IBM Watson [2011]



*William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired **this author's** most famous novel.*





Combines information extraction & natural language understanding.

Applications: Machine Translation

Machine Translation

- One of the main research areas in NLP, and one of the oldest. Historical motivation: Translate Russian to English.

- MT is really difficult:
  - “Out of sight, out of mind” → “Invisible, imbecile”
  - *“The spirit is willing, but the flesh is weak”*
    English → Russian → English
    “The vodka is good, but the meat is rotten”

- Challenges: Word order, multiple translations for a word (need context), want to preserve meaning.

Machine Translation

- Until recently phrase-based translation was the predominant framework.

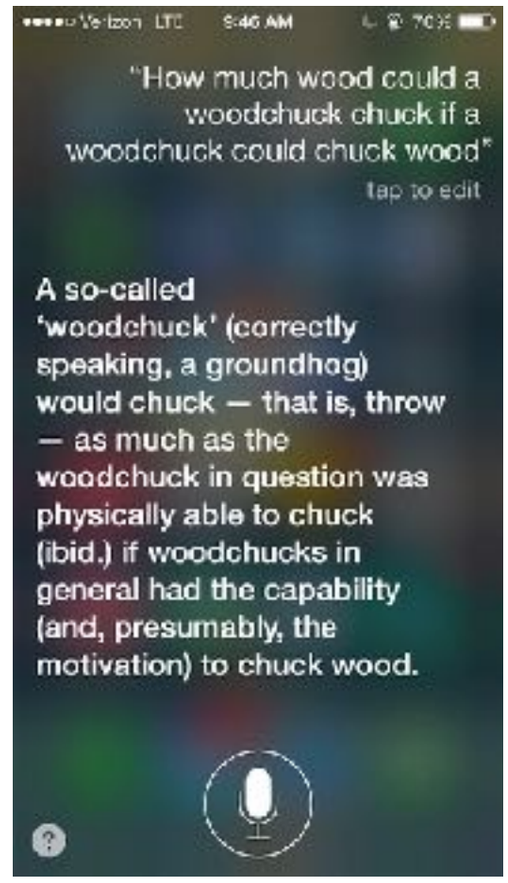
- Today neural network sequence-to-sequence models are used.

- Google Translate supports > 100 languages.

Applications: Virtual Assistants

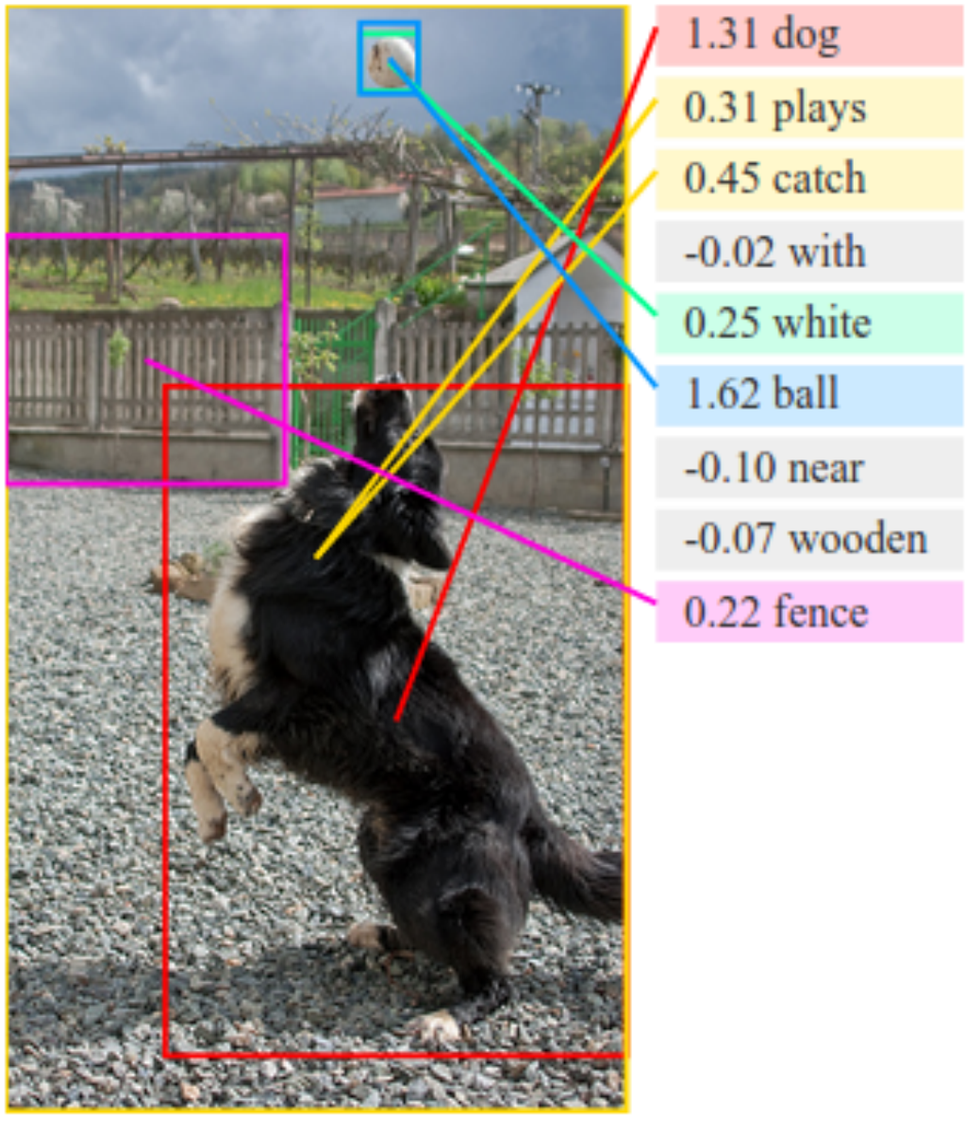
- Siri (Apple), Google Now, Cortana (Microsoft), Alexa (Amazon).

- Subtasks: Speech recognition, language understanding (in context?), speech generation, …



Applications:Image Captioning

“Man in black t-shirt is playing guitar.”



- Neural Networks for Object Detection and Language Generation.
- “Multi-modal” embeddings.
- Microsoft COCO data set.

A. Karpathy, L. Fei-Fei. D*eep Visual-Semantic Alignments for Generating Image Descriptions.* CVPR 2015.

What You Will Learn In This Course

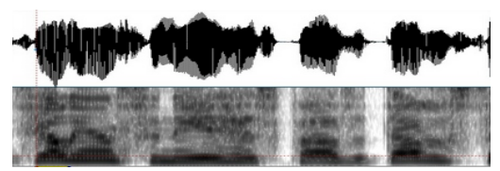
- How can machines **understand** and **generate** natural language?
- Theories about language (linguistics).
- Algorithms.
- Statistical / Machine Learning Methods.
- Applications.

Course Overview

- Part I: Core NLP techniques.
    - Language modeling, part-of-speech tagging, syntactic parsing, word-sense disambiguation, semantic parsing, text similarity.
- Part II: Applications.
    - text classification, information retrieval, question answering, text generation, summarization, machine translation, image captioning, dialog systems.
- Machine Learning Techniques:
  Supervised machine learning, bayesian models, sequence models (n-gram models, HMMs), neural networks, recurrent neural networks,...

Levels of Linguistic Representation

phonetics phonology	sounds and sound patterns of language



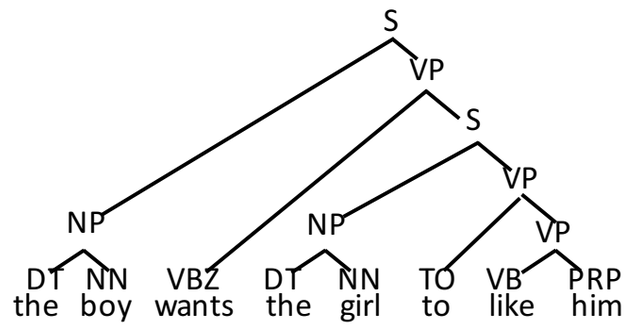
/bɔɪ/

morphology	formation of words

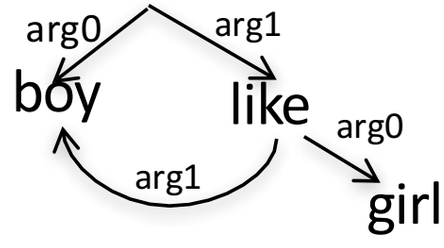
in- + validate + -ed

syntax	word order

```
DT  | NN  |   VBZ     |  DT  | NN   | TO | VB   | PRP |.
the | boy | want+s    | the  | girl | to | like | him |.
```



semantics	word and sentence meaning



pragmatics	influence of context and situation

Natural Language Processing as Translation

- Most NLP techniques can be understood as translation tasks from one structure into another.

- For each translation step:
    - Construct search space of possible translations.
    - Find best paths through this space (decoding) according to some performance measure.
- Modern NLP relies on Machine Learning to figure out these translation steps.

NLP is hard: Ambiguity

- Unlike artificial languages, natural language is full of ambiguity.
- This can happen **on all levels of representation.**
    - *“Wreck a Nice Beach” , “Recognize Speech”*
    - *“inflammable” = **in** + **-flammable***
    - *“Enraged Cow Injures Farmer with Axe”*
    - “*Stolen Painting Found by Tree”*
    - *“Red Tape Holds Up New Bridges”*
    - *“Mouse”*

More Real Headlines

- *Ban on nude dancing on Governor's desk*
- *Eye drops off shelf*
- *Kids Make Nutritious Snacks*
- *Drunk gets nine months in violin case*
- *Government head seeks arms*
- *Patient at death's door – doctors pull him through*
- *In America a woman has a baby every 15 minutes*

Syntactic Structure

- What is the **part-of-speech** of each word? (noun, verb, adjective, adverb, determiner, …)

- What are the **constituents**:
    - Noun phrase: *“Enraged cow”*, *“The cat with the hat”*, *“Columbia University”*

- What are the **subjects and objects:**
    - “*Dog bites man*” vs. “*Man bites dog*”

- **Modification:**
    - *“John saw the man in the park with a telescope”*

Structural Ambiguity

- Interplay between constituent structure and modification.
- Prepositional Phrase (PP) attachment:

Enraged cow injures farmer with axe.

Representing Modification with Brackets

[Enraged cow] [injures [farmer [with axe]]]
NP **NP** **PP**

[Enraged cow] injures [farmer] [with axe]]
NP **NP** **PP**

More PP attachment

[Ban] on [nude dancing] [on governor’s desk]

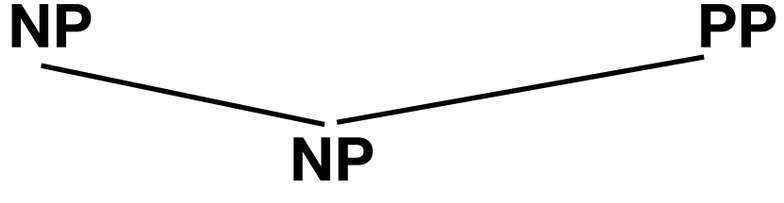
NP **NP** **NP**

- What are the possible modifications? Which one is correct?

[[Ban] on [nude dancing]] [on governor’s desk]

NP **PP**

[Ban] on [[nude dancing] [on governor’s desk]]



Noun-Noun Modification

- Compound nouns also have internal structure:

country song platinum album

1. [country [song [platinum album]]]

2. [country [[song platinum] album]]

3. [[country song] [platinum album]]

4. [[country [song platinum]] album]

5. [[[country song] platinum] album]

Noun-Noun Modification

- What is the *semantic* relationship between nouns in a noun compound?

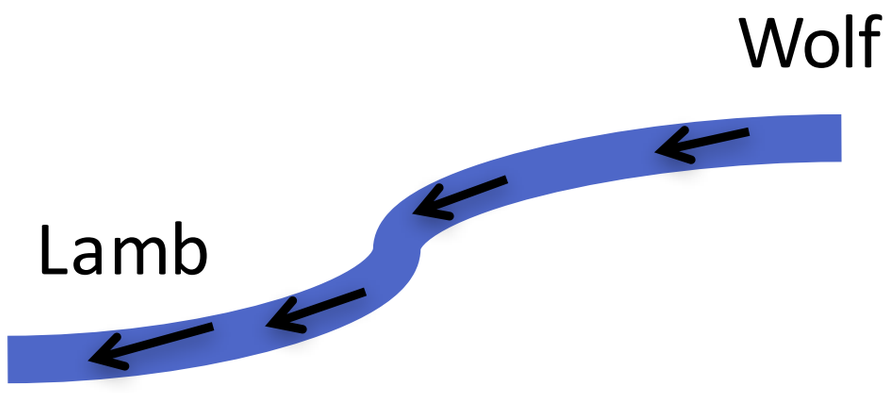
  - *Water fountain:* A fountain that **supplies** water.
  - *Water ballet:* A ballet that **takes place** in water.
  - *Water meter:* A device that **measures** water.
  - *Water barometer:* A barometer that **uses** water (instead of mercury) to measure air pressure.
  - *Water glass:* A glass that is meant to **hold** water.

Other tricky phenomena

- Need for semantic representation.

There was once a Wolf who saw a Lamb drinking at a river and wanted an excuse to eat it.

For that purpose, ***even though*** *he himself was* ***upstream****, he accused the Lamb of stirring up the water and keeping him from drinking. . .*



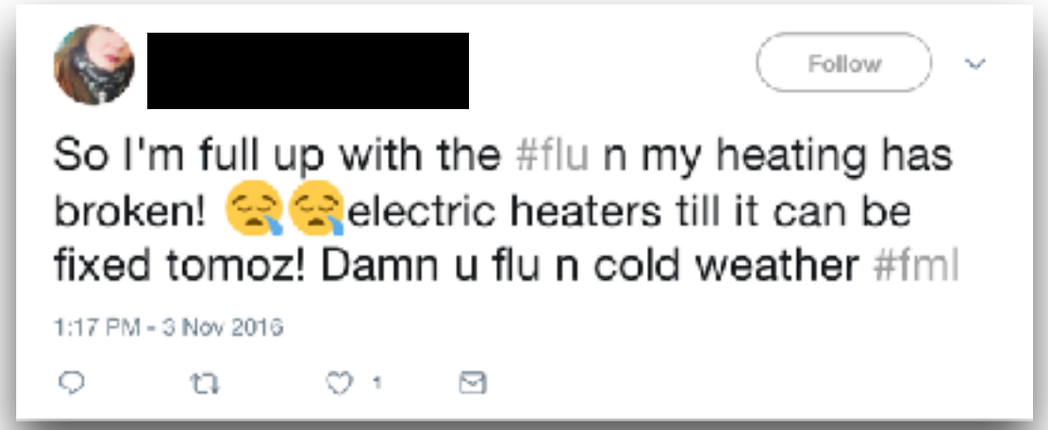
Minsky 1975

Other tricky issues: Language Variety

- Problem: Most NLP techniques were developed on English (specifically financial news written in American English in the 1980s), or other languages with many resources.

- Languages use different mechanisms to express meaning (morphology vs. word-order).

Other tricky issues: Domains and Language Change



- Non-standard English
- Idioms: *throw in the towel, get cold feet, kick the bucket*
- Neologisms (fixed lexicon doesn't work)
  - *noob, crowdsource, unfriend, retweet, bromance, …*

Morphology

- Structure and formation of words.

- **Derivational** morphology: Create new words from old words (can also change the part-of-speech).

  anti- + dis- + **establish** + -ment + -arian + -ism

- **Inflectional** morphology:

  - Convey information about number, person, tense, aspect, mood, voice, and the role a word plays in the sentence (case).

  - English has few morphological categories, but many languages are morphologically rich.

Morphology

- Morphological categories in English
  - Number (“*dog*”, “*dog +s*”)
  - Person (“*I run*”, “*She runs*”)
  - Tense (*“He waited”*)
  - Voice (“*The issue was decided”)*
- Other examples from other languages?

Acknowledgments

- Some slides and examples from Kathy McKeown, Dan Jurafsky, Dragomir Radev.