Natural Language Processing

Lecture 1: Course Overview and Introduction.

9/10/2021

N L P

COMS W4705 (2) – Fall B 2021
Yassine Benajiba

The 4705 Team

- **Instructor:** Yassine Benajiba <**yb2235@columbia.edu**>
  Office Hours: Friday after class
  Virtusl upon request**: time slot to be communicated soon**

- **Assistants:**
  - Let’s look at Courseworks

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

Lectures & Recitation Sessions

- **Lectures:**

Fri 4:10pm-6:40pm, CSB 451

Recitation Sessions:

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

I am requesting to have lectures and recitation sessions to be recorded. Zoom will be used only in case we switch from in-person

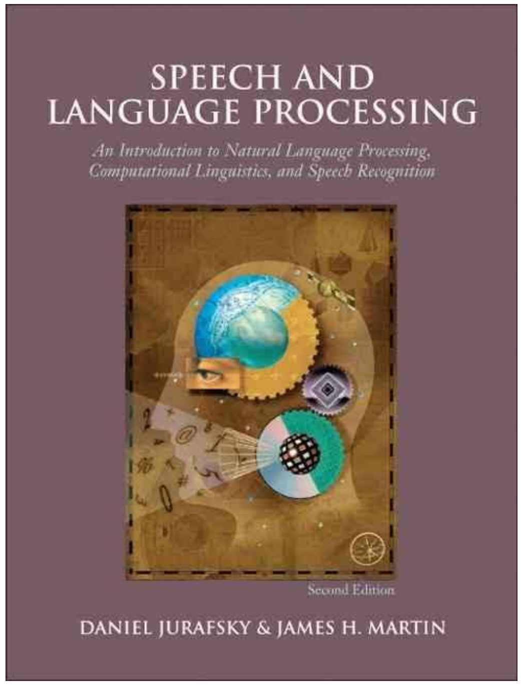
Course Resources

- **Gradescope/Courseworks 2 (a.k.a Canvas):**
  - Courseworks: All course materials: Lecture notes, code, announcements, assignments, reading materials
  - Homework submission, grade book on both
- **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- Recommended textbook (somewhat outdated, we won’t follow too closely):


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

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

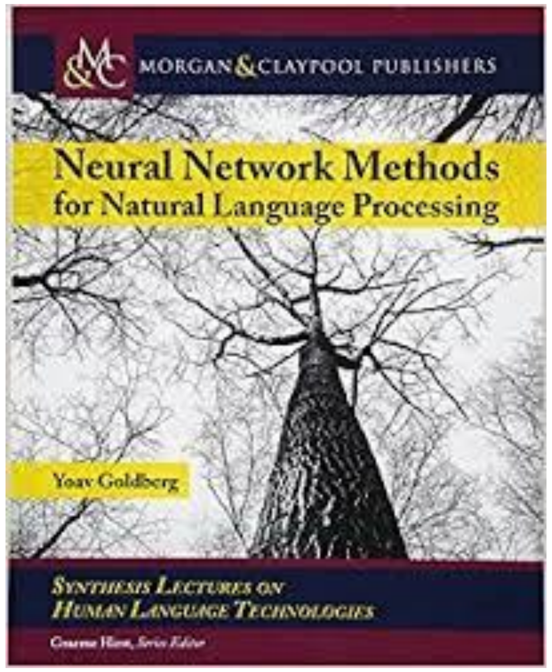
- Draft of most 3rd edition chapters: https://web.stanford.edu/~jurafsky/slp3/

- 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


MORGAN & CLAYPOOL PUBLISHERS
Neural Network Methods
for Natural Language Processing
Yoav Goldberg
Synthesis Lectures on
Human Language Technologies
Graeme Hirst, Series Editor

- 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%
- 4 Homework assignments, each contains an analytical and a programming part, 12.5% each
- Regrade requests should be submitted on Gradescope within 3 days!

Homework

- Homework uploaded through Courseworks AND Gradescope. Do not email!
- PDF: analytical part + copy/paste code programming part (only for comment and discussion on Gradescope)
- Python 3: programming part on Courseworks only

Homework Late Policy

- Written homework and programming problems may be submitted up to 3 days late for a 20 point penalty.
- No homework will be accepted more than 3 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).

Quick discussion of course title

Natural Language Processing

NLP in the Movies

Open the pod bay doors HAL!
I'm sorry Dave, I'm afraid I can't do that!

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, …)

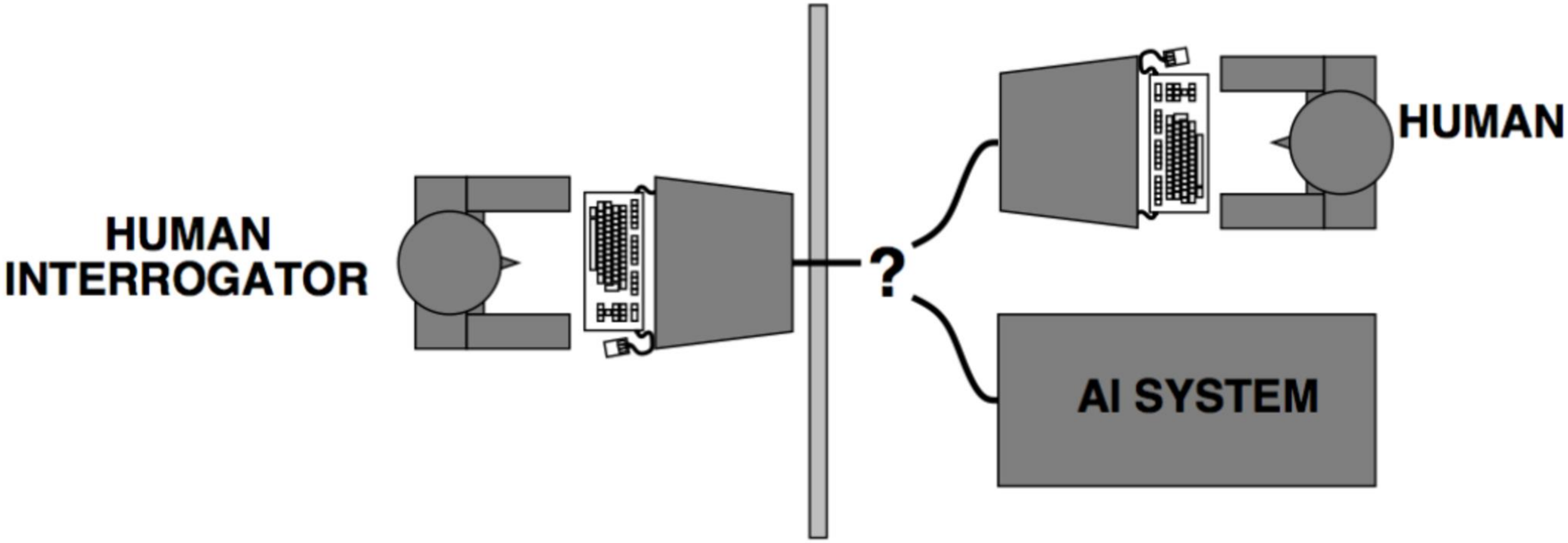
- NLP leverages more and more available training data and modern Machine Learning techniques.

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

Turing Test

(Alan Turing, 1950)

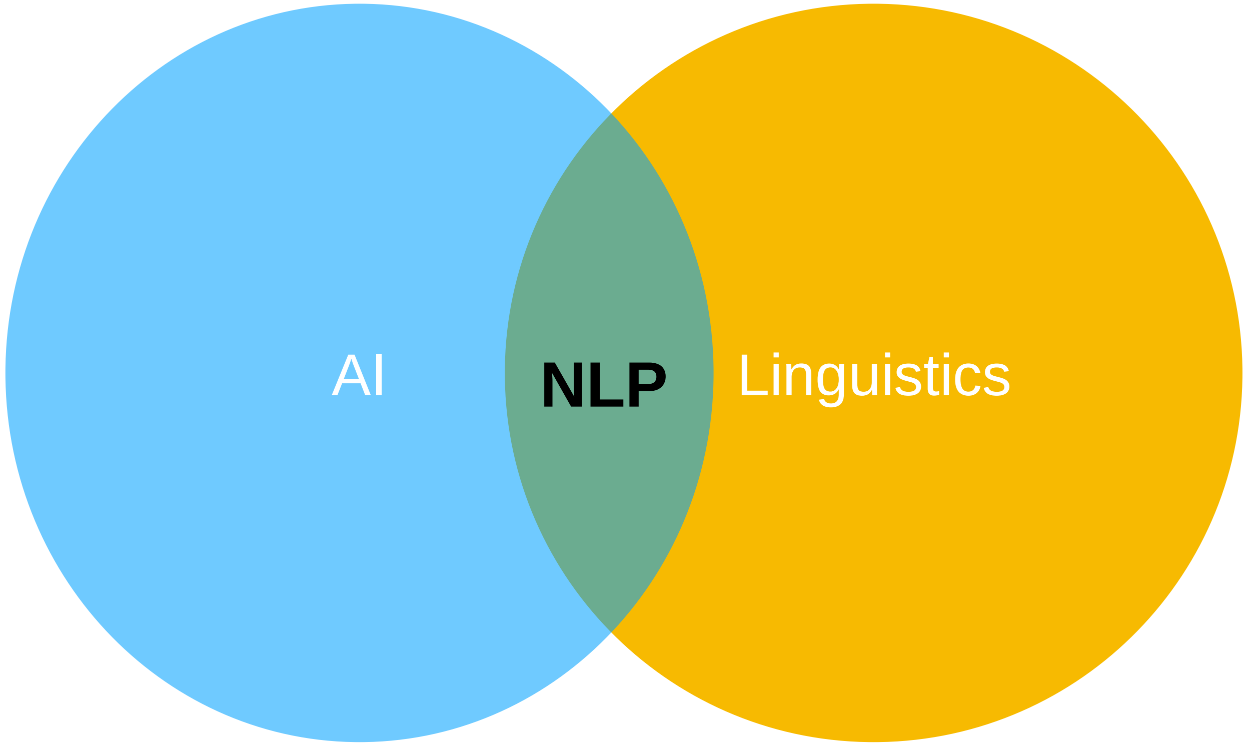
- A computer passes the test of intelligence if it can fool a human interrogator into believing it is human.



- 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

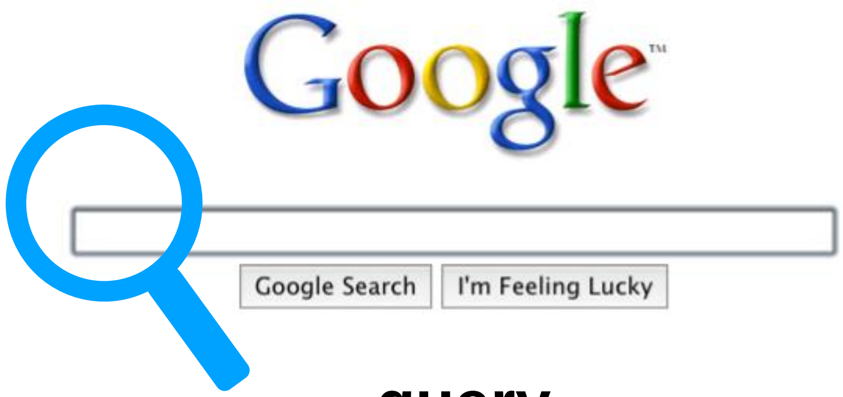


“Every time I fire a linguist, my performance goes up” (Fred Jelinek)

Natural Language Processing vs. Computational Linguistics

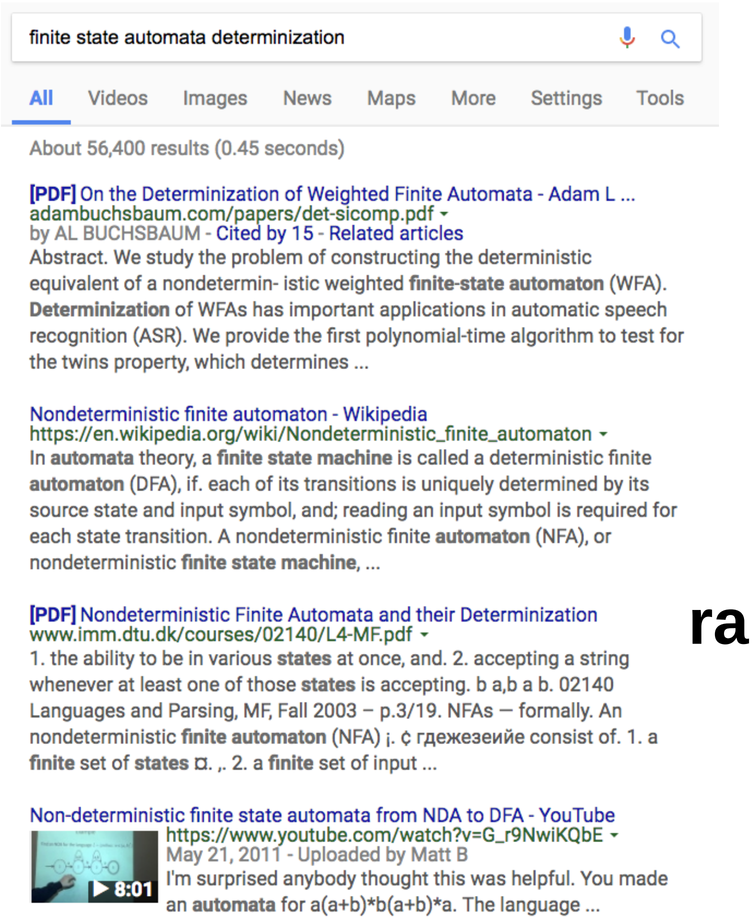
- **NLP:** Build systems that can understand and generate natural language. Focus on applications.
- **Computational Linguistics:** Study human language using computational approaches.
- Many overlapping techniques.

Applications: Information Retrieval



query

indexed document corpus



finite state automata determinization

About 56,400 results (0.45 seconds)

[PDF] On the Determinization of Weighted Finite Automata - Adam L ...
adambuchsbaum.com/papers/det-sicomp.pdf
by AL BUCHSBAUM - Cited by 15 - Related articles
Abstract. We study the problem of constructing the deterministic equivalent of a nondetermin- istic weighted **finite-state automaton** (WFA). **Determinization** of WFAs has important applications in automatic speech recognition (ASR). We provide the first polynomial-time algorithm to test for the twins property, which determines ...

Nondeterministic finite automaton - Wikipedia
https://en.wikipedia.org/wiki/Nondeterministic_finite_automaton
In **automata** theory, a **finite state machine** is called a deterministic finite **automaton** (DFA), if. each of its transitions is uniquely determined by its source state and input symbol, and; reading an input symbol is required for each state transition. A nondeterministic finite **automaton** (NFA), or nondeterministic **finite state machine**, ...

[PDF] Nondeterministic Finite Automata and their Determinization
www.imm.dtu.dk/courses/02140/L4-MF.pdf
1. the ability to be in various **states** at once, and. 2. accepting a string whenever at least one of those **states** is accepting. b a,b a b. 02140 Languages and Parsing, MF, Fall 2003 – p.3/19. NFAs — formally. An nondeterministic **finite automaton** (NFA) ¡. ¢ гдежезеийе consist of. 1. a **finite** set of **states** ¤. ,. 2. a **finite** set of input ...

Non-deterministic finite state automata from NDA to DFA - YouTube
https://www.youtube.com/watch?v=G_r9NwiKQbE
May 21, 2011 - Uploaded by Matt B
I'm surprised anybody thought this was helpful. You made an **automata** for a(a+b)*b(a+b)*a. The language ...

ranked results

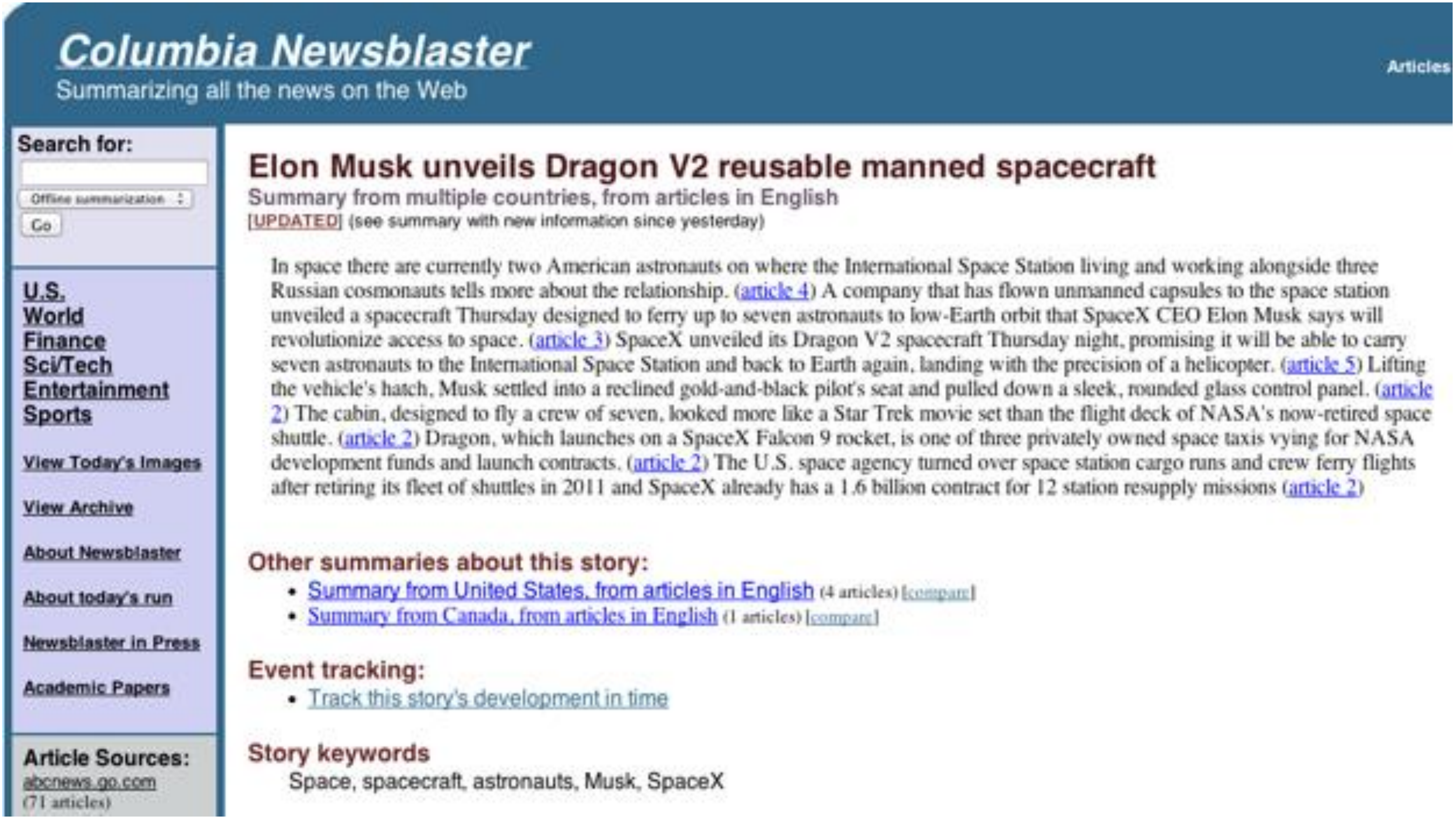
Applications: Text Classification

- 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 boooored...	★
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

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

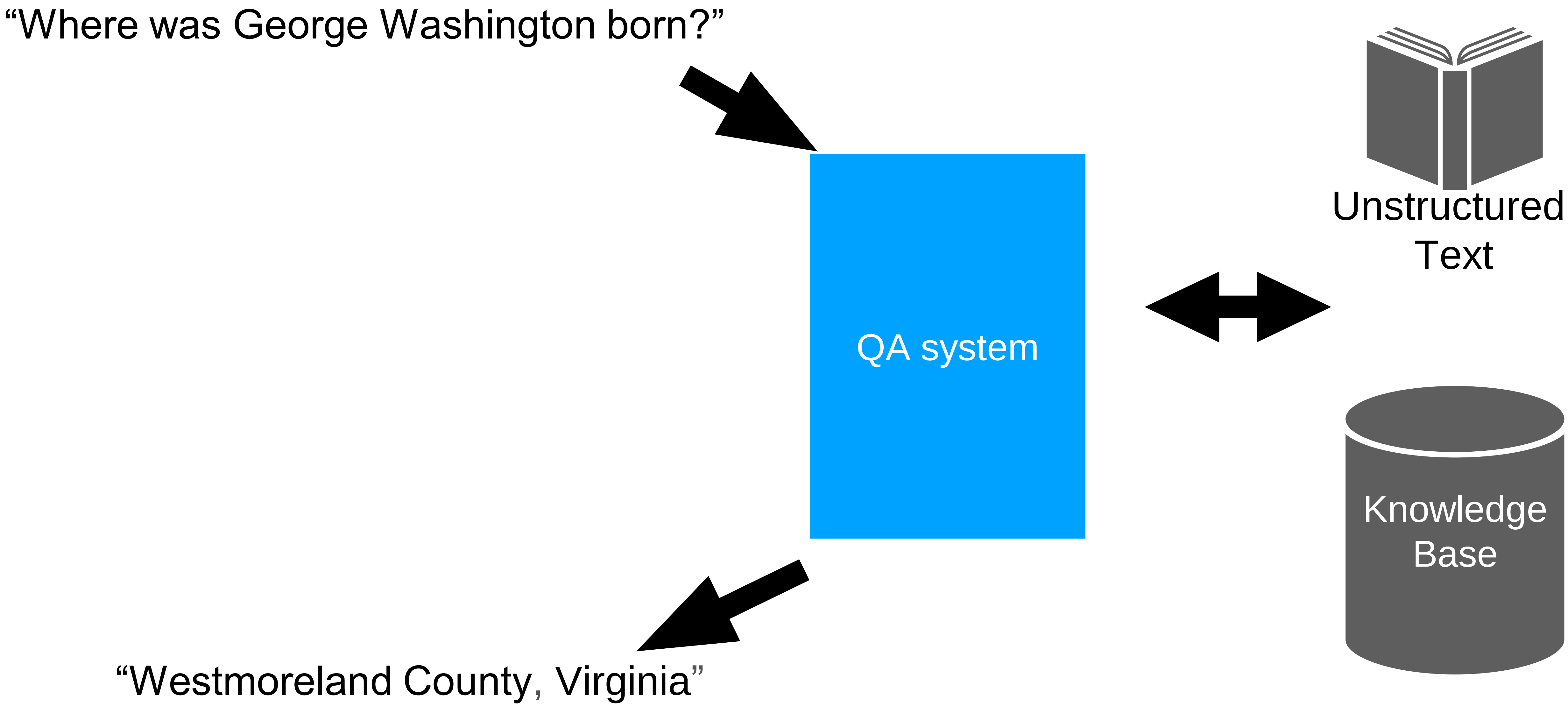
Article Sources:

abcnews.go.com

(71 articles)

Credit: Prof. Kathleen Mc Keown

Application: Question Answering



“Where was George Washington born?”

QA system

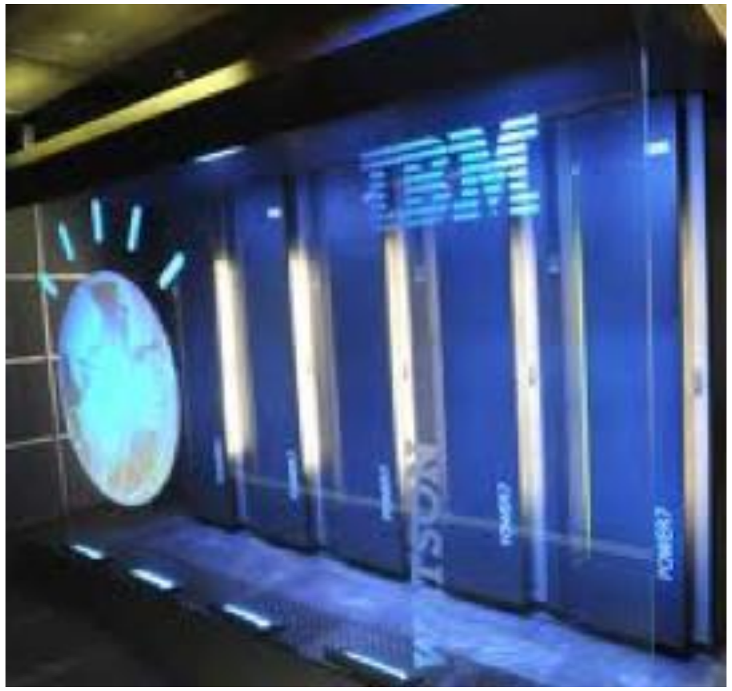
Unstructured Text

Knowledge Base

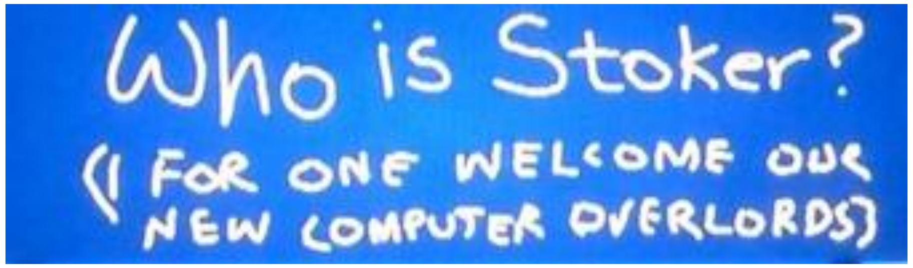
“Westmoreland County, Virginia”

Applications: Playing Jeopardy!

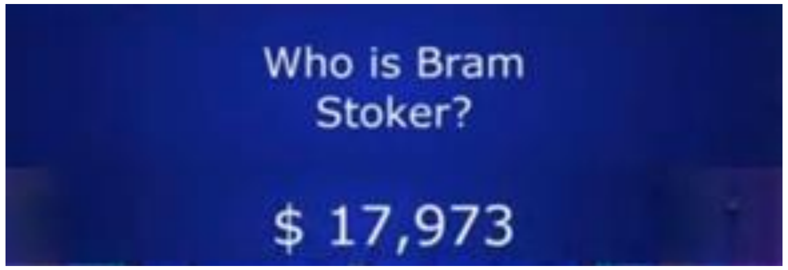
IBM Watson [2011]



William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" i



Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)

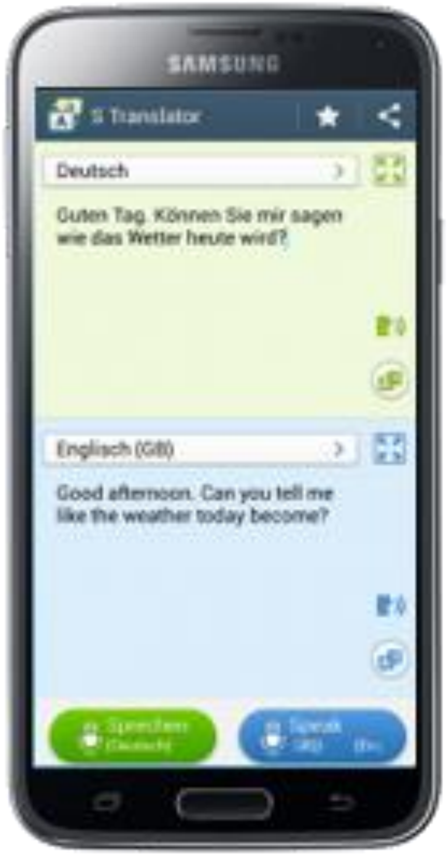


Who is Bram Stoker?

$ 17,973

Combines information extraction & natural language understanding.

Applications: Machine Translation

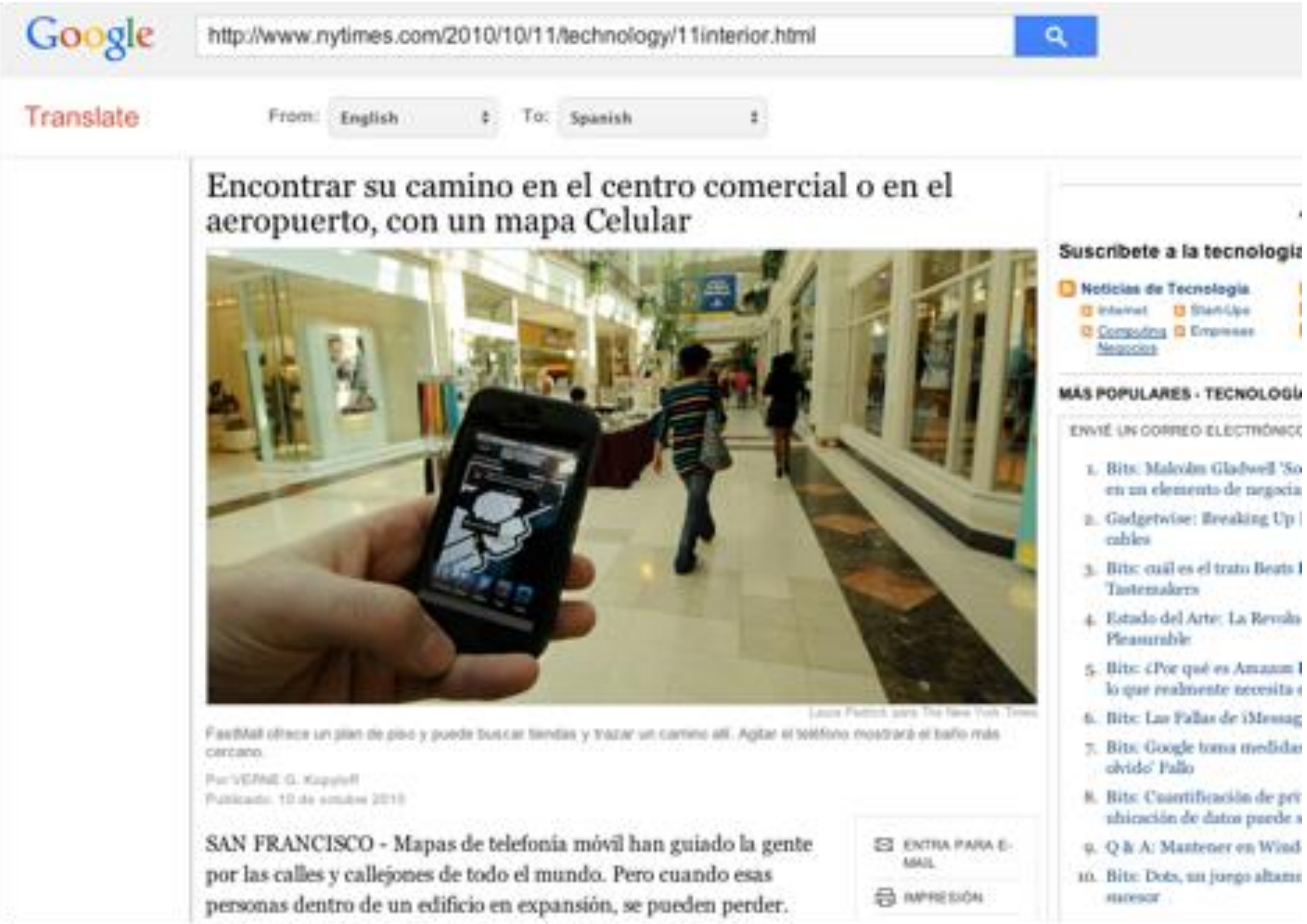


Deutsch

Guten Tag. Können Sie mir sagen wie das Wetter heute wird?

Englisch (GB)

Good afternoon. Can you tell me like the weather today become?



http://www.nytimes.com/2010/10/11/technology/11interior.html

From: English To: Spanish

Encontrar su camino en el centro comercial o en el aeropuerto, con un mapa Celular

Laura Pedrick para The New York Times

FastMall ofrece un plan de piso y puede buscar tiendas y trazar un camino allí. Agitar el teléfono mostrará el baño más cercano.

Por VERNE G. Kopytoff
Publicado: 10 de octubre 2010

SAN FRANCISCO - Mapas de telefonía móvil han guiado la gente por las calles y callejones de todo el mundo. Pero cuando esas personas dentro de un edificio en expansión, se pueden perder.

Suscribete a la tecnologia

Noticias de Tecnología
Internet Start-Ups
Computing Empresas
Negocios

MÁS POPULARES - TECNOLOGÍA

ENVIÉ UN CORREO ELECTRÓNICO

1. Bits: Malcolm Gladwell 'So en un elemento de negocia
2. Gadgetwise: Breaking Up cables
3. Bits: cuál es el trato Beats Tastemakers
4. Estado del Arte: La Revolu Pleasurable
5. Bits: ¿Por qué es Amazon lo que realmente necesita
6. Bits: Las Fallas de iMessag
7. Bits: Google toma medidas olvido' Fallo
8. Bits: Cuantificación de pri ubicación de datos puede
9. Q & A: Mantener en Wind
10. Bits: Dots, un juego altam menesor

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”

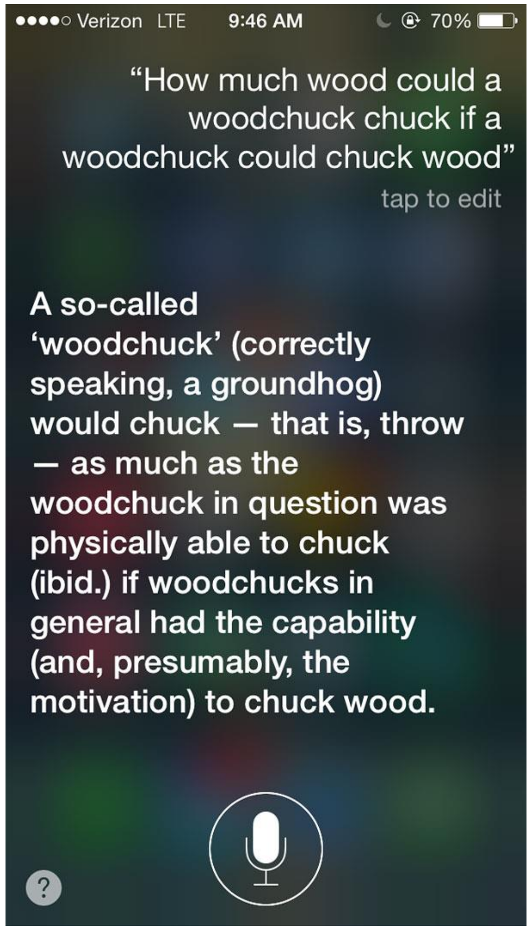
- Challenges: Word order, multiple translations for a word (need context), want to preserve meaning.

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, …

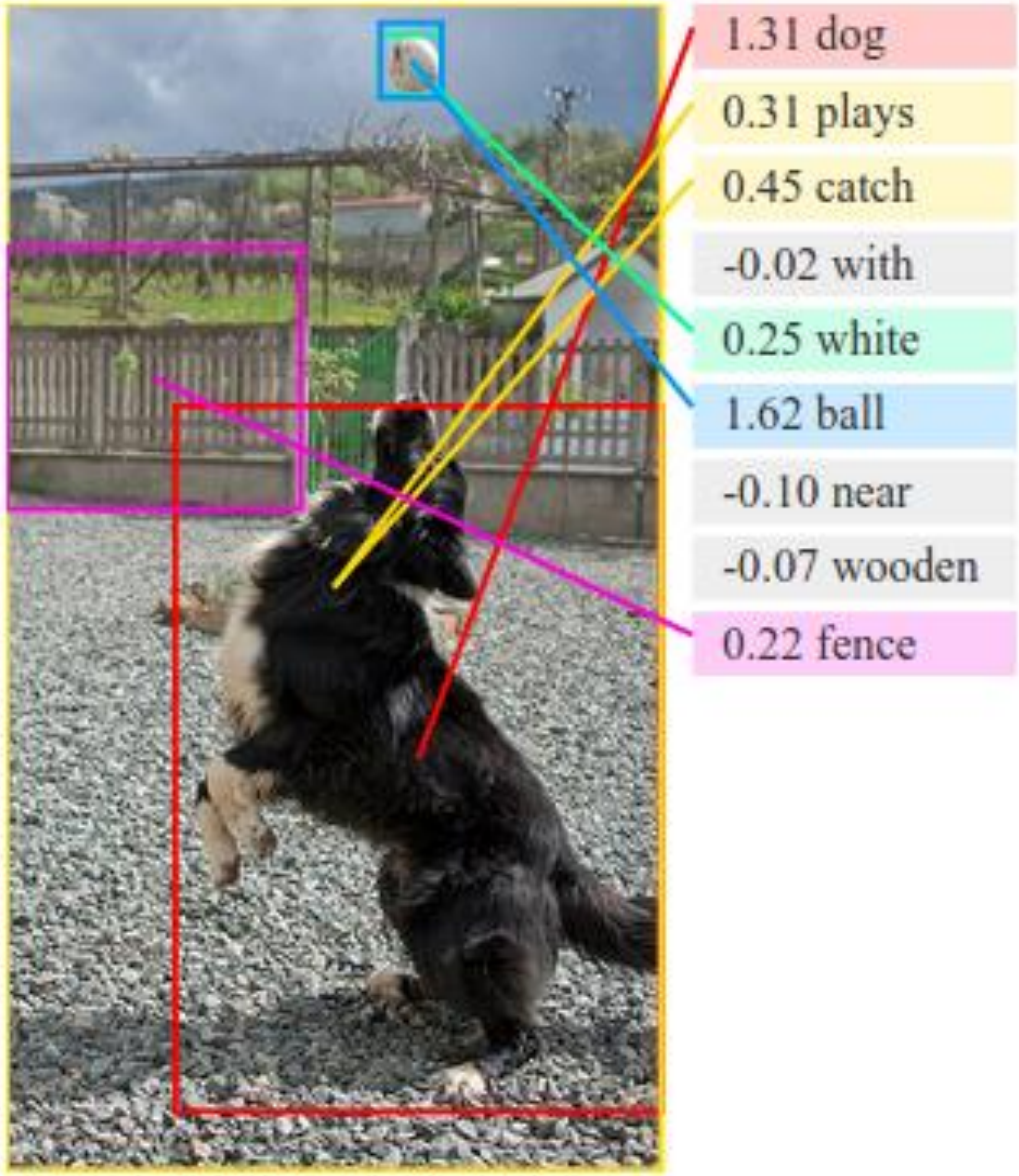


"How much wood could a woodchuck chuck if a woodchuck could chuck wood"

A so-called 'woodchuck' (correctly speaking, a groundhog) would chuck — that is, throw — as much as the woodchuck in question was physically able to chuck (ibid.) if woodchucks in general had the capability (and, presumably, the motivation) to chuck wood.

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 20

What You Will Learn In This Course

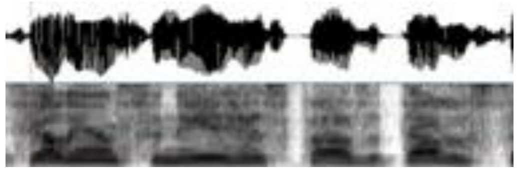
- How can machines **understand** and **generate** natural language?
  - Theories about language (linguistics).
  - Algorithms.
  - Statistical / Machine Learning Methods.
  - Applications.

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, 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

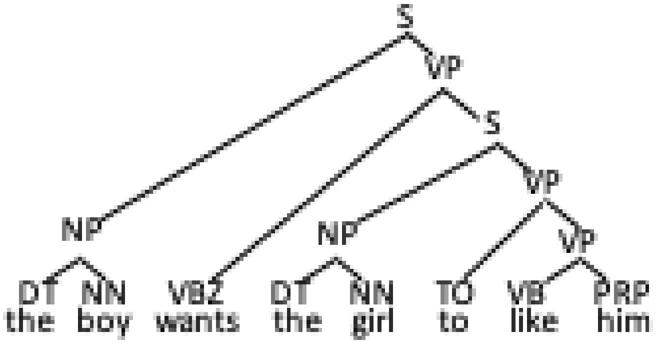
phonetics phonology	sounds and sound patterns of language
morphology	formation of words
syntax	word order
semantics	word and sentence meaning
pragmatics	influence of context and situation

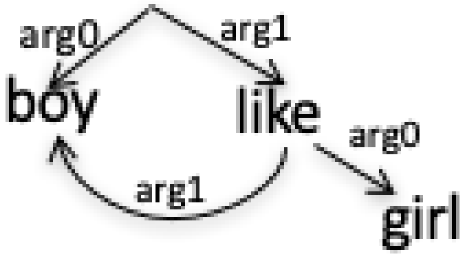


/bɔɪ/

in- + validate + -ed

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





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”*

Real Headlines

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

Syntactic Structure

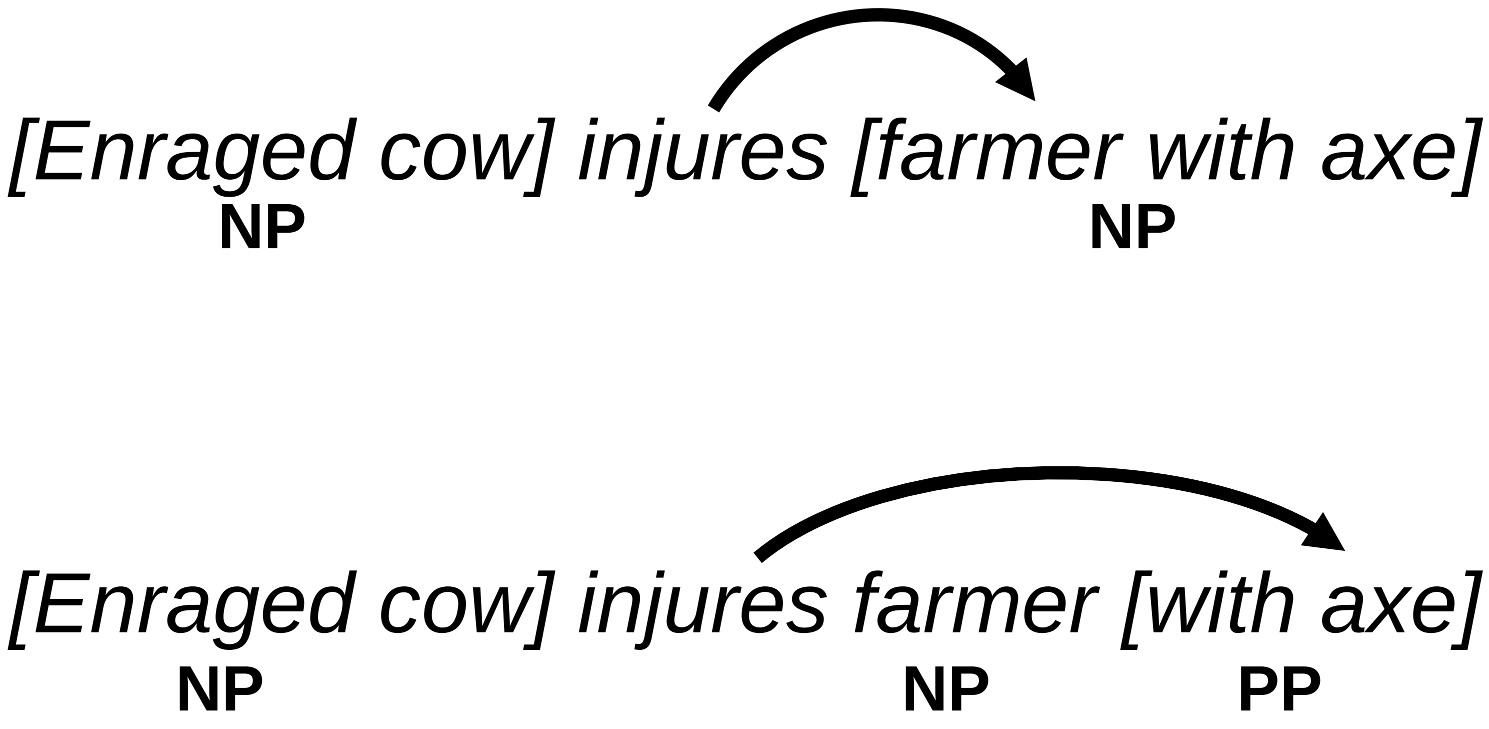
- What is the **part-of-speech** of each word? (noun, verb, adjective, adverb, determiner, …)
- What are the **constituents**:
  - Noun phrase: *“Enraged cow”, “The cat with the hat”, “Columbia University”*
- What are the **subjects and objects:**
  - “*Dog bites man*” vs. “*Man bites dog*”
- **Modification:**
  - *“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.



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

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

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]

Ban on nude dancing on governor’s desk

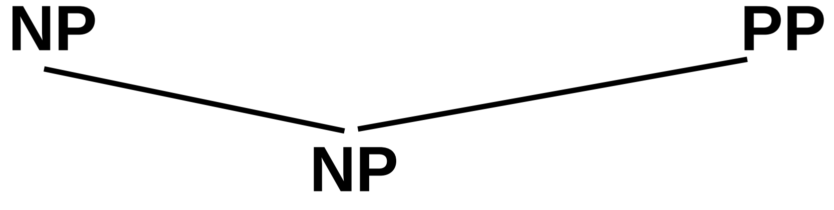
NP **NP** **NP**

- 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]]



NP **PP**

NP

Noun-Noun Modification

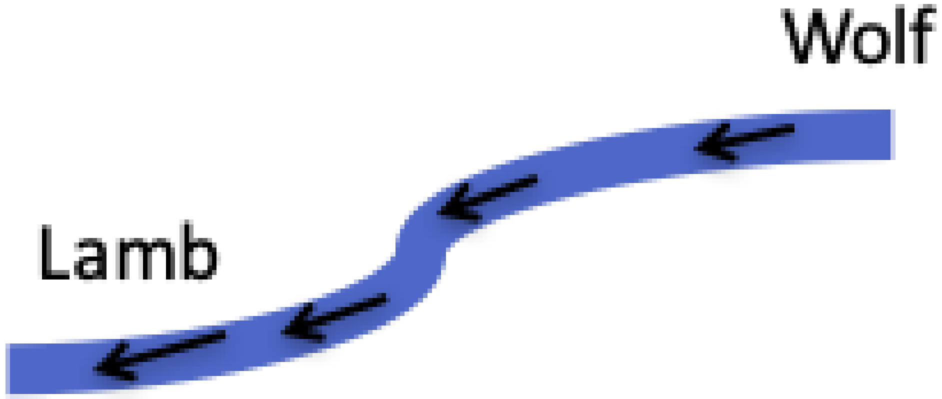
- What is the *semantic* relationship between nouns in a noun compound?
  - *Water fountain:* A fountain that **supplies** water.
  - *Water ballet:* A ballet that **takes place** in water.
  - *Water meter:* A device that **measures** water.
  - *Water barometer:* A barometer that **uses** water (instead 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

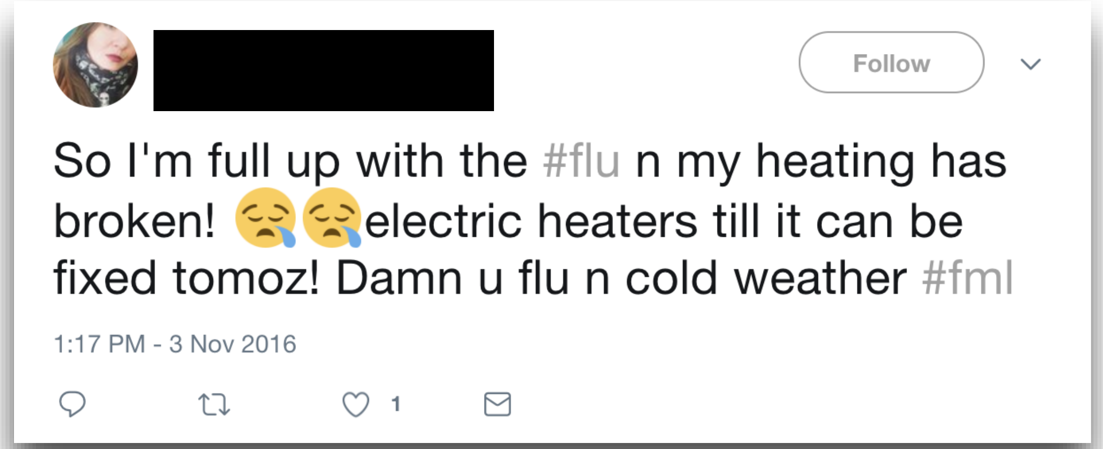
Other tricky issues: Language Variety

- 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



[redacted]

So I'm full up with the #flu n my heating has broken! 😪😪electric heaters till it can be fixed tomoz! Damn u flu n cold weather #fml

1:17 PM - 3 Nov 2016

- 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.