Introduction to Natural Language Processing

By J. H. Wang

Feb. 22, 2023

Outline

- What is NLP?
- Example Tasks
- The state-of-the-art

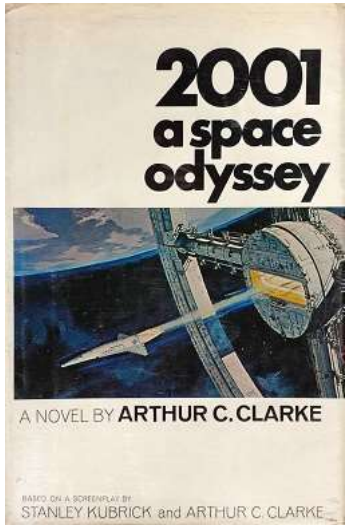
What is Natural Language Processing

- (speech and) language processing
- Human language technology
- Computational linguistics

- To give computers ability to process human language
- To enable human-machine communication
- E.g. conversational agent

Example: 2001: A Space Odyssey

- A 1968 epic science fiction film produced and directed by Stanley Kubrick
- Based on a novel by Arthur C. Clarke
- HAL 9000 computer
  - Speaking and understanding English
  - Even reading lips
- Conversational agent, dialog system
  - Input: automatic speech recognition, natural language understanding
  - Output: dialogue and response planning, speech synthesis



Modern Examples: Chatbots

- ChatGPT, Bard
- Mobile phone assistants
  - Apple Siri, …
- Smart speakers
  - Google Home, Amazon Alexa, LINE Clova, …
- Other applications:
  - Social networking platforms
  - Healthcare
  - Banking
  - …

Example Language-related Tasks

- Machine Translation
  - Automatically translate a document from one language to another
- Web-based question answering, for example:
  - What year was Abraham Lincoln born?
  - How many states were in the United States that year?
  - How much Chinese silk was exported to England by the end of the 18th century?
  - What do scientists think about the ethics of human cloning?
- Information extraction, word sense disambiguation
- Spelling correction, grammar checking, …

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

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

→ Bram Stoker

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafs

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

Hi Dan, we've now scheduled the curriculum meeting.

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

-Chris

Create new Calendar entry

Information Extraction & Sentiment Analysis

Attributes:

zoom
affordability
size and weight
flash
ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't [illegible] hose heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very l[illegible]e to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

Enter Source Text:

這 不過 是 一 個 時間 的 問題 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Knowledge of Language

- Phonetics and Phonology - knowledge about linguistic sounds
- Morphology - knowledge of the meaningful components of words
- Syntax - knowledge of the structural relationships between words
- Semantics - knowledge of meaning
- Pragmatics - knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse - knowledge about linguistic units larger than a single utterance

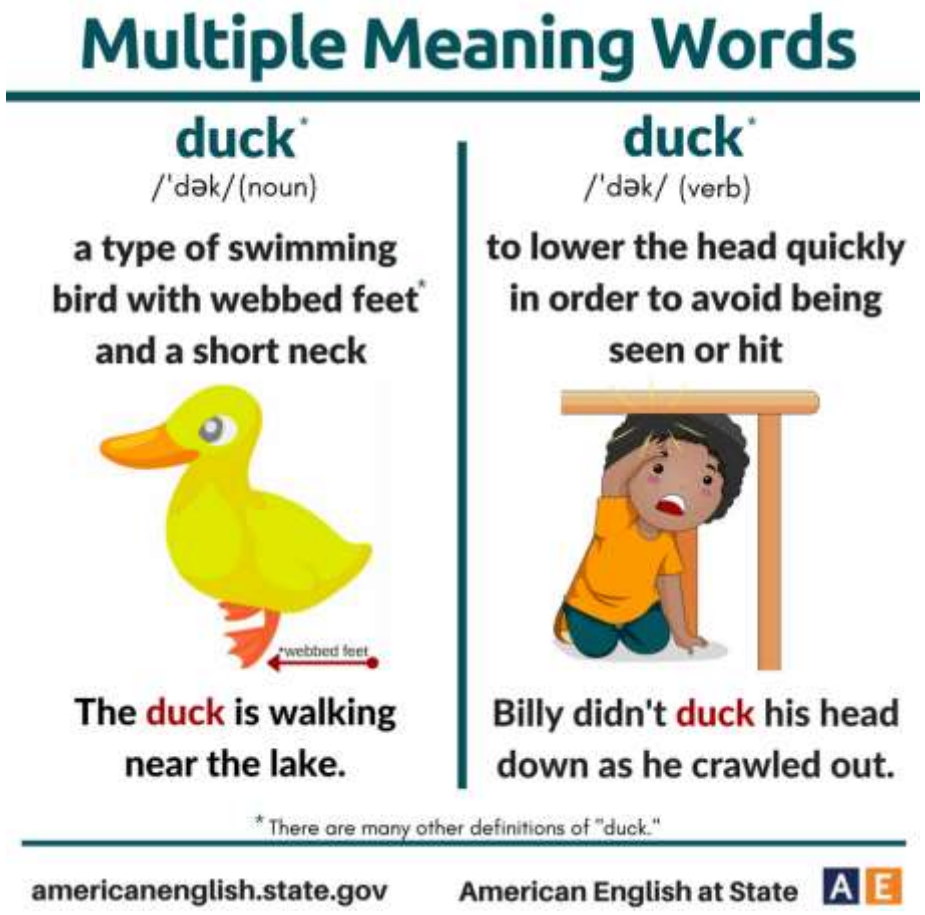
Ambiguity

- Some input is **ambiguous** if multiple, alternative linguistic structures can be built for it
- For example: *I made her duck.*
- What's the meaning?

?

Possible Answers:

- (1.5) I cooked waterfowl for her.
- (1.6) I cooked waterfowl belonging to her.
- (1.7) I created the (plaster?) duck she owns.
- (1.8) I caused her to quickly lower her head or body.
- (1.9) I waved my magic wand and turned her into undifferentiated waterfowl.



Resolving Ambiguities

- Lexical disambiguation
  - Part-of-Speech (POS) Tagging
    - E.g. duck: noun vs. verb
  - Word sense disambiguation
    - E.g. make: create vs. cook
- Syntactic disambiguation
  - Parsing

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

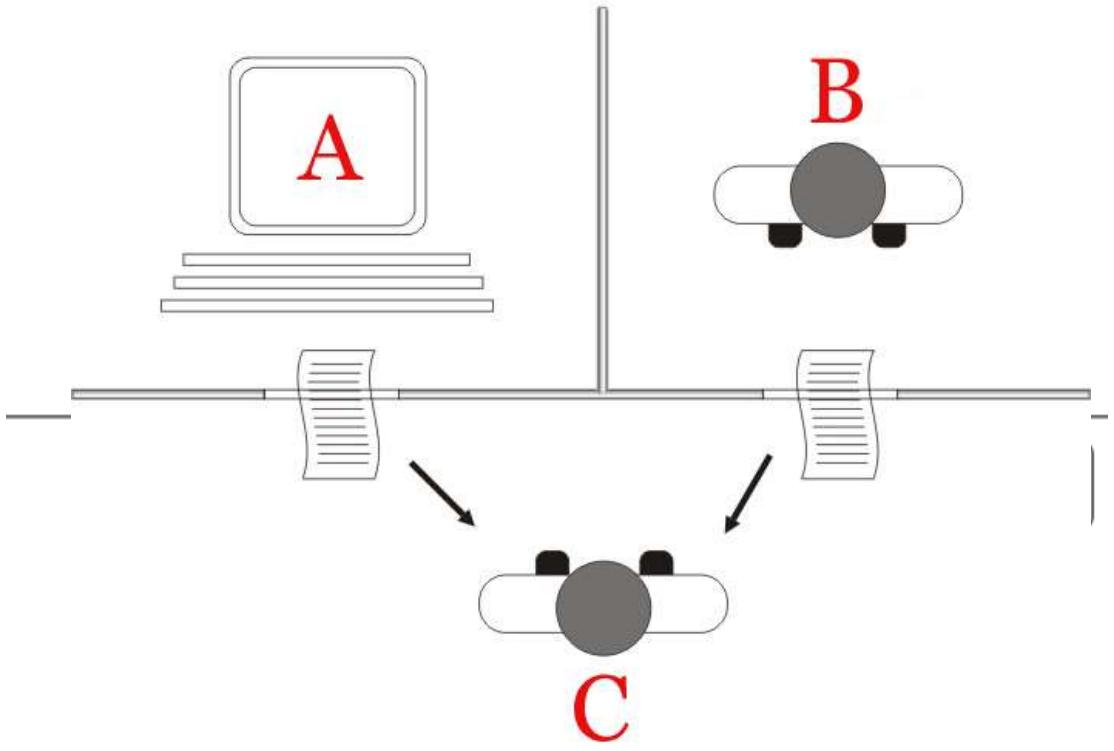
But that's what makes it fun!

Models and Algorithms

- State machines, rule systems, logic, probabilistic models, vector-space models
- State space search algorithm: dynamic programming
- Machine learning algorithm: classification, expectation-maximization (EM), …
- Deterministic and non-deterministic finite state automata, finite-state transducers
- Regular grammars, context free grammars, …

Language, Thought, Understanding

- The ability to process language as skillfully as humans do
  - “Can machines think?”
- Turing test: by Alan Turing (1950)
  - “Imitation game”
  - “Can machines do what we can do?”
- Conversational agents



- Q: Please write me a sonnet on the topic of the Forth Bridge.
- A: Count me out on this one. I never could write poetry.
- Q: Add 34957 to 70764.
- A: (Pause about 30 seconds and then give answer as) 105621.

Example dialog of ELIZA (1966)

Rogerian Psychotherapy
(person-centered therapy)

- User1: You are like my father in some ways.
- ELIZA1: WHAT RESEMBLANCE DO YOU SEE
- User2: You are not very aggressive but I think you don’t want me to notice that.
- ELIZA2: WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
- User3: You don’t argue with me.
- ELIZA3: WHY DO YOU THINK I DON’T ARGUE WITH YOU
- User4: You are afraid of me.
- ELIZA4: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

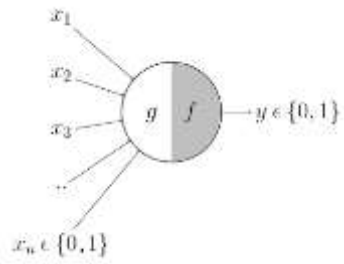
The State of the Art

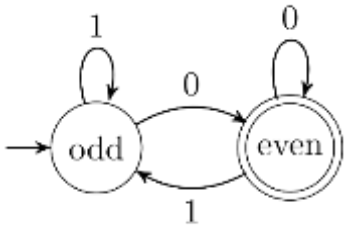
- Travelers calling travel providers interact with conversational agents
- Cars that allow drivers to control by voice
- Searching video on the Web by speech
- Cross-language information retrieval and translation by Google
- Grading and assessing student essays by automated systems
- Interactive virtual agents
- Automated measurement of user opinions, preferences, attitudes in social media
- …

Brief History of NLP

- Different fields in different departments
  - Computational linguistics: in linguistics
  - Natural language processing: in computer science
  - Speech recognition: in electrical engineering
  - Computational psycholinguistics: in psychology

Foundational Insights: 1940s and 1950s





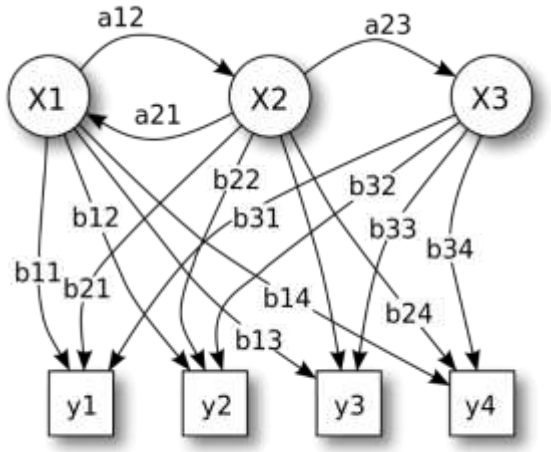
- Automaton: 1950s
  - Turing (1936)
  - McCulloch-Pitts neuron (1943)
  - Finite automata, regular expressions (1951, 1956)
  - Shannon (1948)
  - Context-free grammar: Chomsky (1956), Backus (1959), Naur (1960)
- Probabilistic or information-theoretic models
  - Shannon: communication, entropy
  - Koenig: sound spectrogram (1946)

The Two Camps: 1957–1970

- Two paradigms
  - Symbolic
    - Formal language theory, parsing
    - Artificial intelligence: John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester (1956)
  - Stochastic
    - Bayesian method

Four Paradigms: 1970–1983

- Stochastic
  - Hidden Markov model
- Logic-based
- Natural language understanding
- Discourse modeling



Empiricism and Finite-State Models Redux: 1983–1993

- Finite-state models
- Probabilistic models

The Field Comes Together: 1994–1999

- Probabilistic and data-driven models had become quite standard throughout natural language processing
- The increases in the speed and memory of computers had allowed commercial exploitation of a number of subareas of speech and language processing, in particular, speech recognition, and spelling and grammar correction
- The rise of the Web emphasized the need for language-based information retrieval and information extraction

The Rise of Machine Learning: 2000–2008

- Large amounts of spoken and written material became widely available through the auspices of the Linguistic Data Consortium (LDC) and other similar organizations
- This increased focus on learning led to a more serious interplay with the statistical machine learning community
  - SVM, maximum entropy, multinomial logistic regression, graphical Bayesian models
- The widespread availability of high-performance computing systems facilitated the training and deployment of systems that could not have been imagined a decade earlier
- Near the end of this period, largely unsupervised statistical approaches began to receive renewed attention
  - Topic modeling, …

New Potentials for NLP

- Powerful computing resources
- Web as the massive source of information
- Availability of wireless mobile access
- Many new application scenarios

Thanks for Your Attention!