What is Natural Language Processing (NLP)?

What is Natural Language Processing (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

What is Natural Language Processing (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics

(Christopher Manning)

What is Natural Language Processing (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics

(Christopher Manning)

Make computers to understand natural language to do certain task humans can do such as Machine translation, Summarization, Questions answering

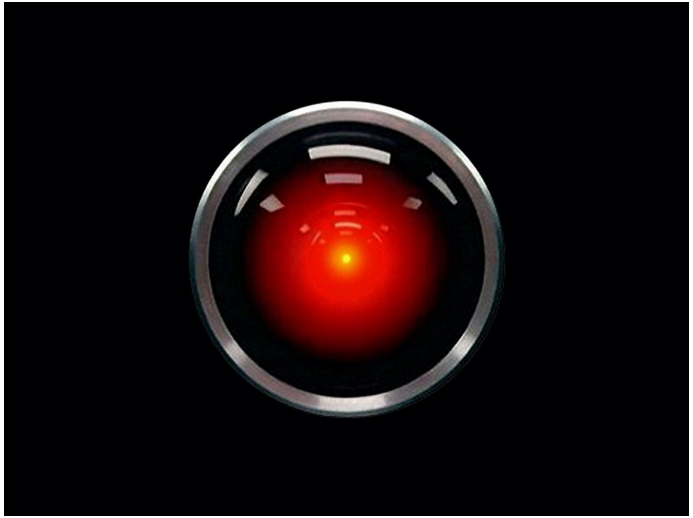
(Behrooz Mansouri)

Example: Conversational Agent

Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
...HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



2001: A Space Odyssey – HAL 9000

HAL is an artificial agent capable of such advanced language-processing behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips

Natural Language Processing: Terms

Natural language refers to the language that humans use to communicate with each other, such as English, Spanish, or Chinese

Processing

As distinguished from data processing

Question: How is data processing and natural language processing different?

Natural Language Processing: Terms

Consider the Unix wc program, which counts the total number of bytes, words, and lines in a text file

- When used to count bytes and lines, wc is an ordinary data processing application
- However, when it is used to count the words in a file, it requires knowledge about what it means to be a word and thus becomes a language processing system

Natural Language Processing vs Computational Linguistics

In linguistics, language is the object of study

- Computational methods may be brought to bear, just as in scientific disciplines like computational biology and computational astronomy, but they play only a supporting role

In contrast, natural language processing is focused on the design and analysis of computational algorithms and representations for processing natural human language

- The goal of natural language processing is to provide new computational capabilities around human language: for example, extracting information from texts, translating between languages, answering questions, holding a conversation, taking instructions

Knowledge Requirement for Machine

Machines require much broader and deeper knowledge of language

What does HAL need?

Knowledge Requirement for Machine

Machines require much broader and deeper knowledge of language

What does HAL need?

- Recognize words from an audio signal and to generate an audio signal from a sequence of words
  - knowledge about phonetics and phonology: how words are pronounced in terms of sequences of sounds
- HAL is capable of producing contractions like *I'm* and *can't*
  - knowledge about morphology, the way words break down into component parts that carry meanings
- HAL must use structural knowledge to properly string together the words that constitute its response
  - knowledge needed to order and group words comes under the heading of syntax
- …

Knowledge Requirement for Machine

- 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

Phonetics and Phonology

- Phonetics and Phonology: knowledge about linguistic sounds
- The study of:

language sounds how they are physically formed;	systems of discrete sounds, e.g. languages' syllable structure

dis-k&-'nekt disconnect

Morphology

- Morphology: knowledge of the meaningful components of words
- The study of the sub-word units of meaning

Even more necessary in some other languages,

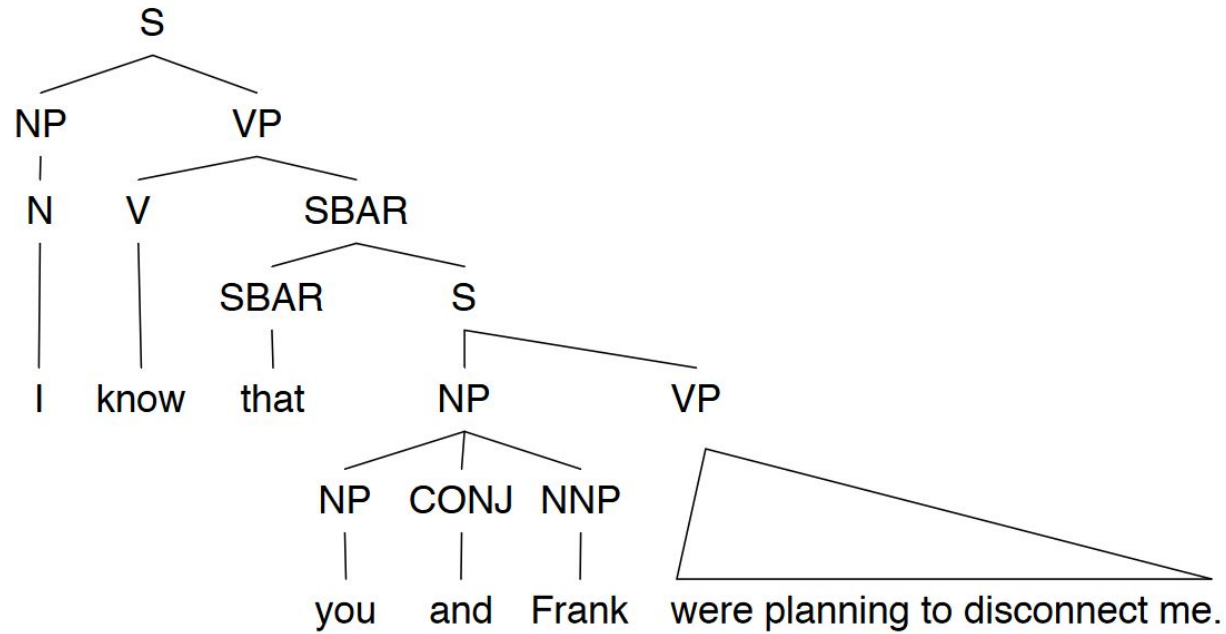
e.g. Turkish:

uygarlastiramadiklarimizdanmissinizcasina

uygar las tir ama dik lar imiz dan mis siniz casina

Syntax

- Syntax: knowledge of the structural relationships between words
- The study of the structural relationships between words
  - I know that you and Frank were planning to disconnect me.



Semantics

- Semantics: knowledge of meaning
- The study of the literal meaning
  - I know that you and Frank were planning to disconnect me.
  - ACTION = disconnect
  - ACTOR = you and Frank
  - OBJECT = me

Pragmatics

- Pragmatics: knowledge of the relationship of meaning to the goals and intentions of the speaker
- The study of how language is used to accomplish goals
  - What should you conclude from the fact I said something?
  - How should you react?
    - I'm sorry Dave, I'm afraid I can't do that.
    - Includes notions of polite and indirect styles

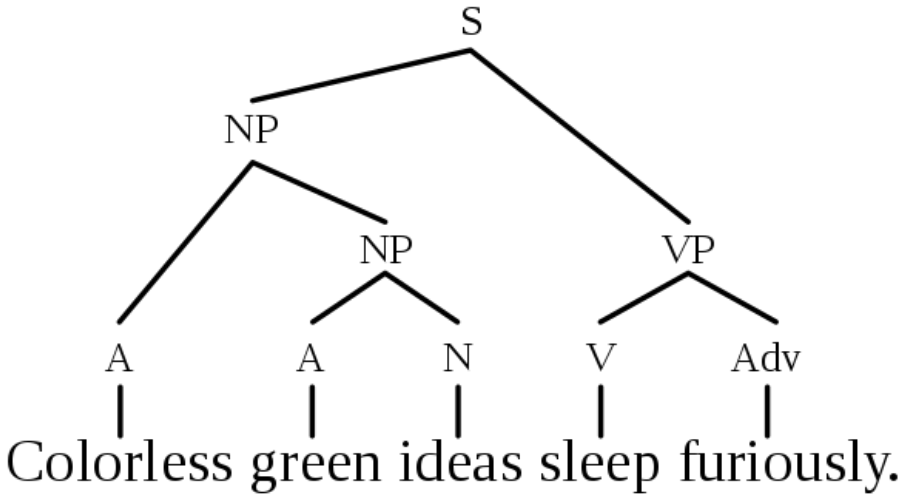
Discourse

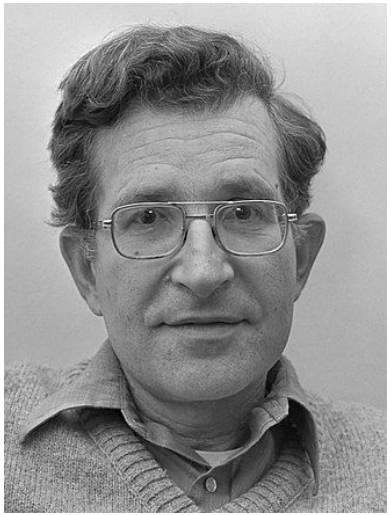
- Discourse: knowledge about linguistic units larger than a single utterance
- The study of linguistic units larger than a single utterance
- The structure of conversations:
  - turn taking, thread of meaning

Syntax vs. Semantics

Colorless green ideas sleep furiously.

(example by Noam Chomsky 1957)





Noam Chomsky
The most cited person alive

Semantics vs. Pragmatics

What does "You have a green light" mean?

- You are holding a green light bulb?
- You have a green light to cross the street?
- You can go ahead with your plan?

Is NLP hard?

What does this sentence mean? “*I made her duck*”

“duck”: noun or verb?
“make”: “cook X” or “cause X to do Y” ?
“her”: “for her” or “belonging to her” ?

Is NLP hard?

What does this sentence mean? “*I made her duck*”

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

These different meanings are caused by a number of ambiguities

Is NLP hard?

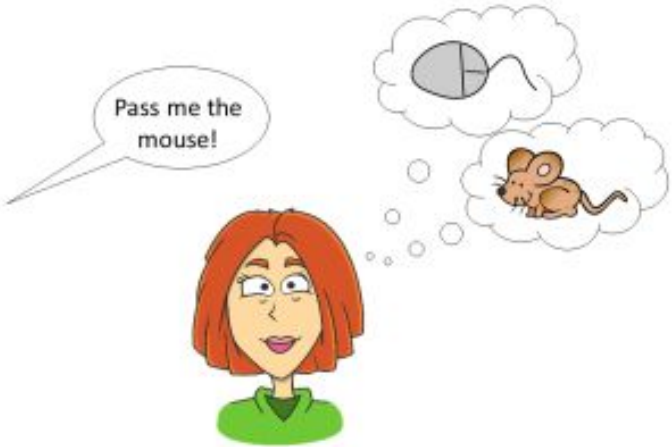
What does this sentence mean? “*I made her duck*”

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

These different meanings are caused by a number of ambiguities

- First, the words duck and her are morphologically or syntactically ambiguous in their part-of-speech
  - Duck can be a verb or a noun, while her can be a dative pronoun or a possessive pronoun
- Second, the word make is semantically ambiguous; it can mean create or cook
- Finally, the verb make is syntactically ambiguous in a different way

We Need to Disambiguate



Disambiguation

Models and algorithms in this course are ways to resolve or disambiguate these ambiguities

- Deciding whether duck is a verb or a noun can be solved by part-of-speech tagging
- Deciding whether make means “create” or “cook” can be solved by word sense disambiguation

Resolution of part-of-speech and word sense ambiguities are two important kinds of lexical disambiguation

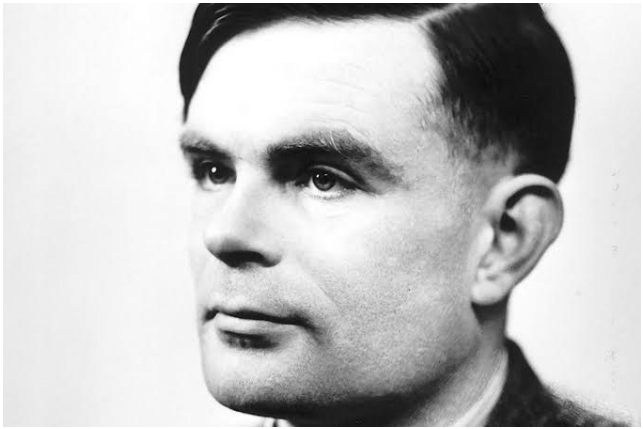
A wide variety of tasks can be framed as lexical disambiguation problems

- A text-to-speech synthesis system reading the word lead needs to decide whether it should be pronounced as in lead pipe or as in lead me on
- Deciding whether her and duck are part of the same entity or are different entities is an example of syntactic disambiguation and can be addressed by probabilistic parsing

History of NLP

Turing Test

"Computing Machinery and Intelligence"
Mind, Vol. 59, No. 236, pp. 433-460, 1950

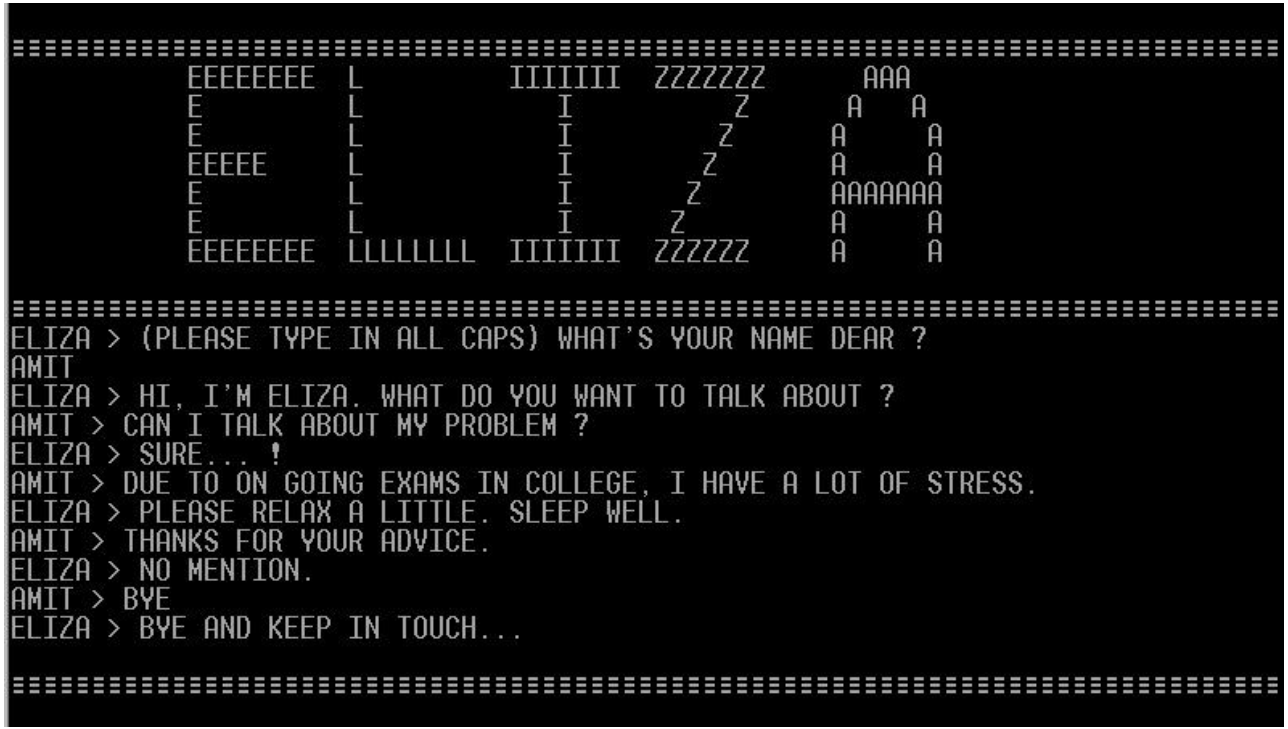


I propose to consider the question
"Can machines think?"...
We can only see a short distance ahead, but
we can see plenty there that needs to be done

In Turing's game, there are three participants: two people and a computer.
One of the people is a contestant who plays the role of an interrogator. To win, the interrogator must determine which of the other two participants is the machine by asking a series of questions via a teletype. The task of the machine is to fool the interrogator into believing it is a person by responding as a person would to the interrogator's questions. The task of the second human participant is to convince the interrogator that the other participant is the machine and that she is human.

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.

ELIZA



ELIZA was an early natural language processing system capable of carrying on a limited form of conversation with a user

1950 – 1970

Mid 1950's – Mid 1960's: Birth of NLP and Linguistics

- At first, people thought NLP is easy! Researchers predicted that "machine translation" can be solved in 3 years or so
- Mostly hand-coded rules / linguistic-oriented approaches
- The 3-year project continued for 10 years, but still no good result, despite the significant amount of expenditure

Mid 1960's – Mid 1970's: A Dark Era

- After the initial hype, a dark era follows
- People started believing that machine translation is impossible, and most abandoned research for NLP

1970 – 2000

1970's and early 1980's – Slow Revival of NLP

- Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation

Late 1980's and 1990's – Statistical Revolution!

- By this time, the computing power increased substantially
- Data--driven, statistical approaches with simple representation win over complex hand-coded linguistic rules
- "Whenever I fire a linguist, our machine translation performance improves." (Jelinek, 1988)

2000's – Statistics Powered by Linguistic Insights

- With more sophistication with the statistical models, richer linguistic representation starts finding a new value

Recent Years

2010's – Emergence of embedding model and deep neural networks

- Several embedding models for text using neural networks and deep neural networks were proposed including Word2Vec, Glove, fastText, Elmo, BERT, COLBERT, GTP[1-3.5]
- New techniques brought attention to more complex tasks

Tasks/Applications in NLP

A few of the NLP Tasks

- Spell Checking, Keyword Search, Finding Synonyms
- Part of Speech Tagging
- Extracting information from a website
  - Location, people, temporal expressions
- Classifying text
  - Sentiment analysis
- Machine translation
- Complex question answering
- Spoken dialog systems

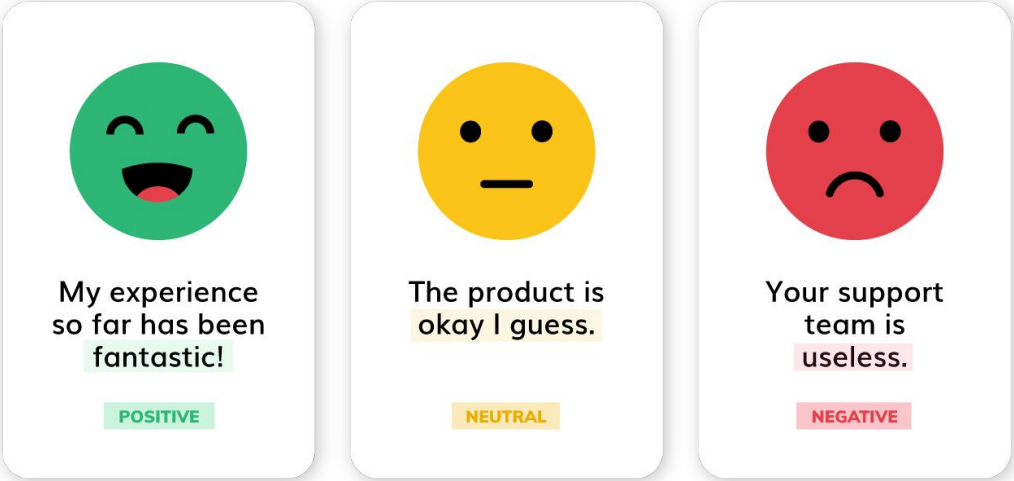
Knowledge & Information Extraction

Knowledge graphs (KGs) organize data from multiple sources, capture information about entities of interest in a given domain or task (like people, places or events), and forge connections between them

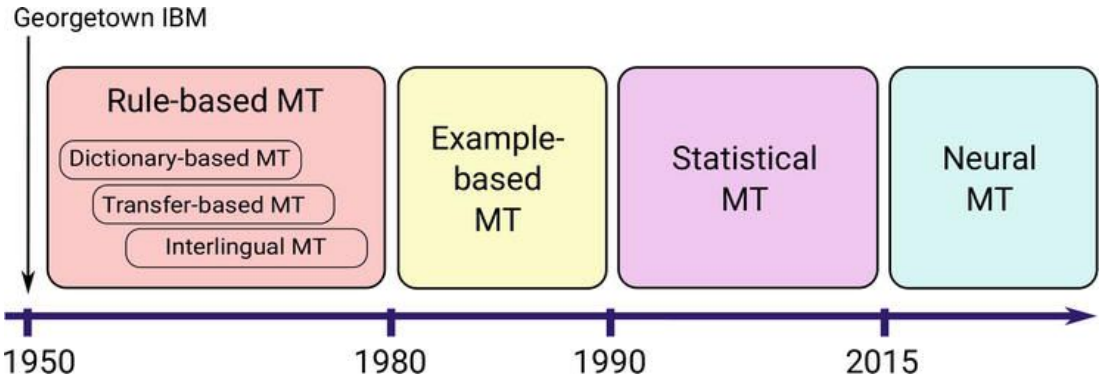
The Google Knowledge Graph is an enormous database of information that enables Google to provide immediate, factual answers to your questions

Sentiment Analysis

Determine whether the meaning behind data is positive, negative, or neutral



Machine Translation



Low resource languages can be challenging?

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Question Answering

IBM-Watson Defeats Humans in "Jeopardy!"

Spoken Dialog Systems

Where to find Tasks and Test Collections?

EMNLP: Conference on Empirical Methods in Natural Language Processing https://2022.emnlp.org/

ACL: Association for Computational Linguistics https://2023.aclweb.org/

NAACL: Annual Conference of the North American Chapter of the Association for Computational Linguistics https://2022.naacl.org/

CoNLL: Conference on Computational Natural Language Learning https://conll.org/2022

COLING: International Conference on Computational Linguistics https://coling2022.org/

CLEF: Conference and Labs of the Evaluation Forum https://clef2022.clef-initiative.eu/index.php

SemEval: Workshop on Semantic Evaluation https://semeval.github.io/SemEval2023/tasks.html


WHAT HAVE
YOU LEARNED?

Summary

In previous session we learned about:

- ✓ What is Natural Language Processing
- ✓ What makes Natural Language Processing hard
- ✓ Natural Language Processing Tasks

Next Session

Python Refresher

You will be reminded of python programming

We will review:

- Basic programming concepts in Python
  - https://docs.python.org/3/tutorial/
- Using libraries, including those needed for this course

To do:

- You should also review the Getting Started page of Google Colab Notebooks
- Bring laptop for testing
- **Reading**: Chapter 1 of Jurafsky Book (here)
- **Question**: How can we crawl data from the internet? Your first assignment is related to this!

Note: You may use other IDEs and editors