Natural Language Processing

Info 159/259
Lecture 1: Intro

Many slides & instruction ideas borrowed from:
David Bamman, Dan Jurafsky & Kemal Oflazar

Introduction

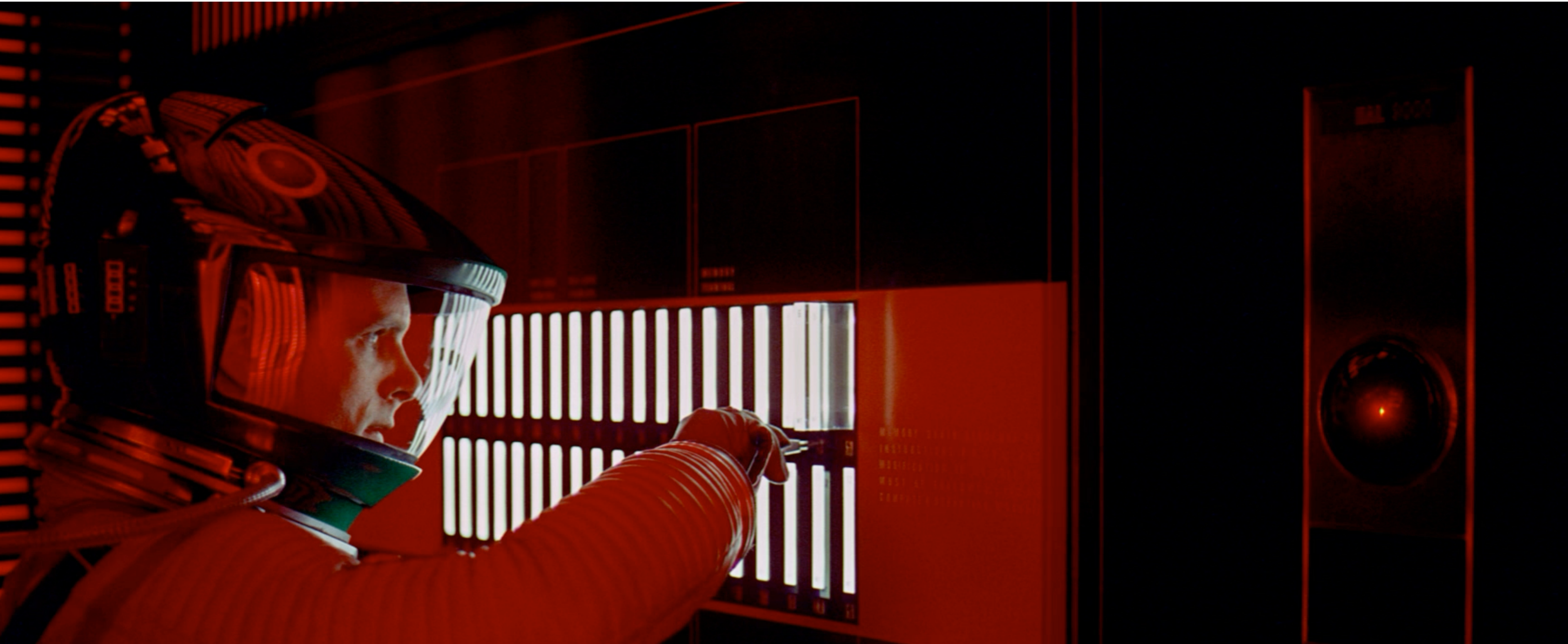
- Instructor: Behrang Mohit
- Teaching Assistants:
    1. Alvin Bao
    2. Cassandra Claire Calciano
    3. Jaewon Lee
    4. Madeleine Tammy Wang
    5. Sanjana Gajendran
    6. Zhihao Du

NLP is interdisciplinary

- Artificial intelligence
- Machine learning (~ 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/ society)

NLP = processing* language with computers

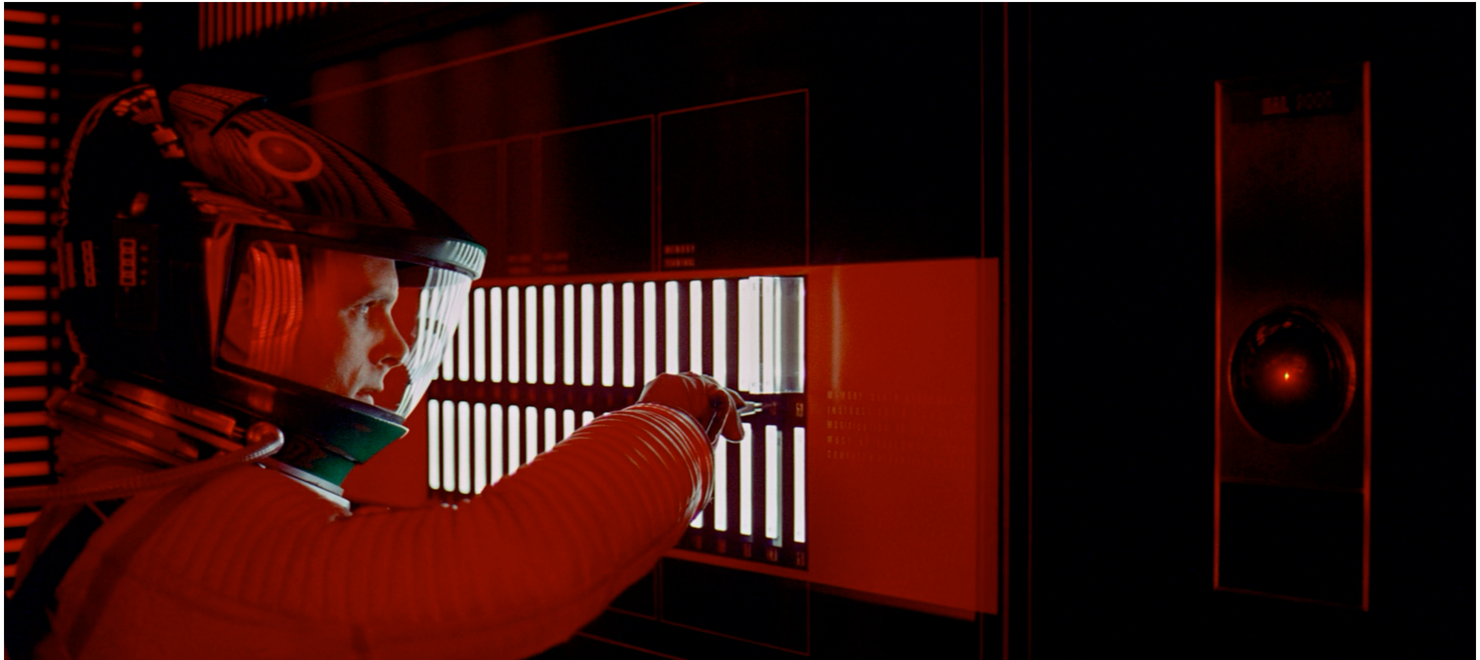
processing as “understanding”



The dream of intelligent machines

- 2001 Space Odyssey (Stanley Kubrik, 1968)
- HAL: Intelligent system capable of:
    - Understanding & Generating human lang.
    - And more

Dave (human): *Open the pod bay doors, HAL*
HAL: *I'm sorry Dave. I'm afraid I can't do that*

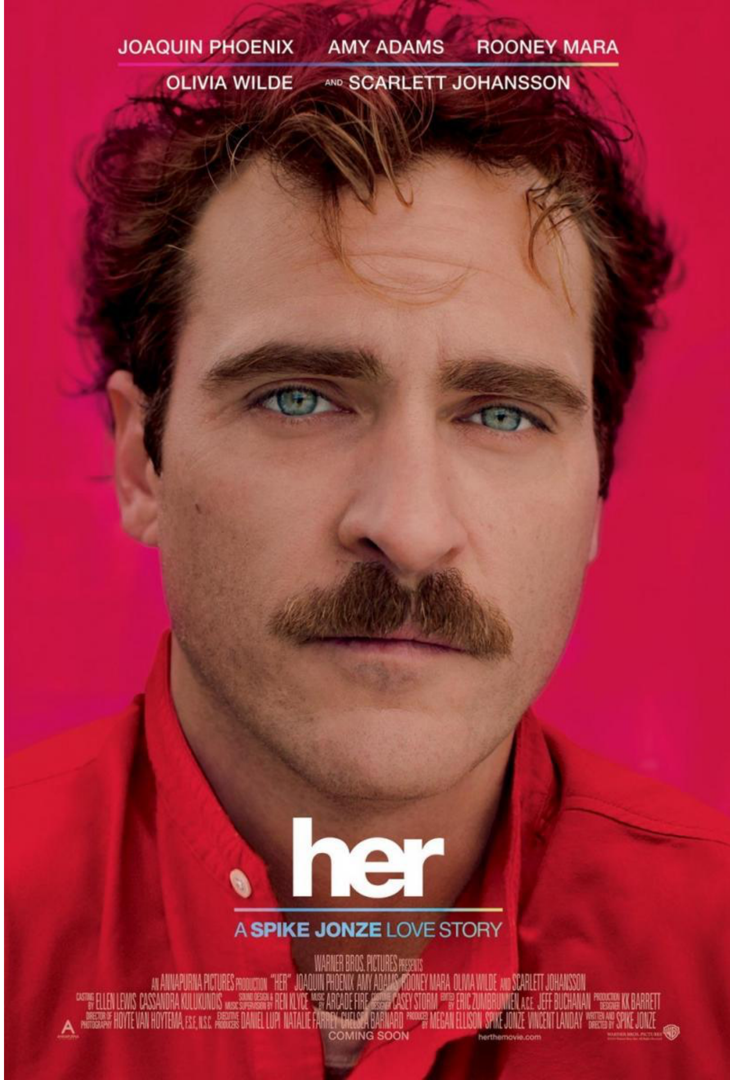


Turing test

Distinguishing human vs. computer only through written language

Turing 1950

- **Samantha:** Is that weird? You think I'm weird?
- **Theodore:** Kind of.
- **Samantha:** Why?
- **Theodore:** Well, you seem like a person but you're just a voice in a computer.
- **Samantha:** I can understand how the limited perspective of an un-artificial mind might perceive it that way. *You'll get used to it.*



What makes language hard?

- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Modeling it is AI-complete (requires first solving general AI)

What makes language hard?

- Speech acts (“can you pass the salt?)
  [Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
  [Grice 1975]

What makes language hard?

You

If I were to say "the opera singer was amazing. she sang all of the notes". did i think the opera singer was good

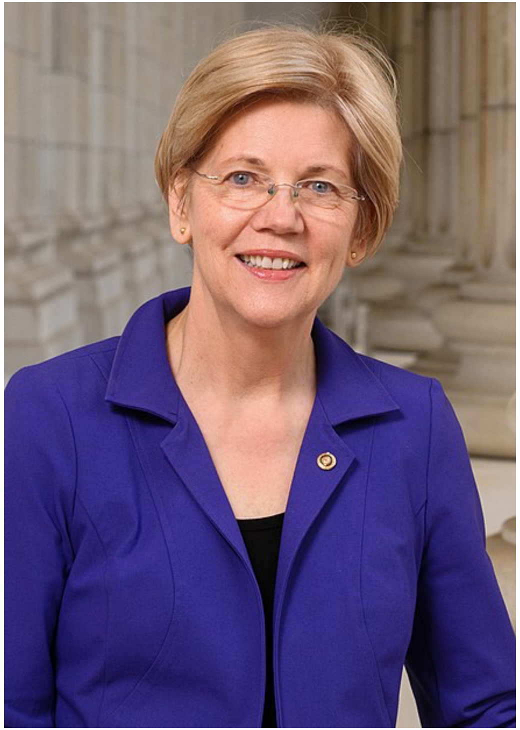
ChatGPT

Based on the statement "the opera singer was amazing. she sang all of the notes," it seems that you have a positive impression of the opera singer. Describing the singer as "amazing" and highlighting that she sang all of the notes suggests that you think highly of her performance.

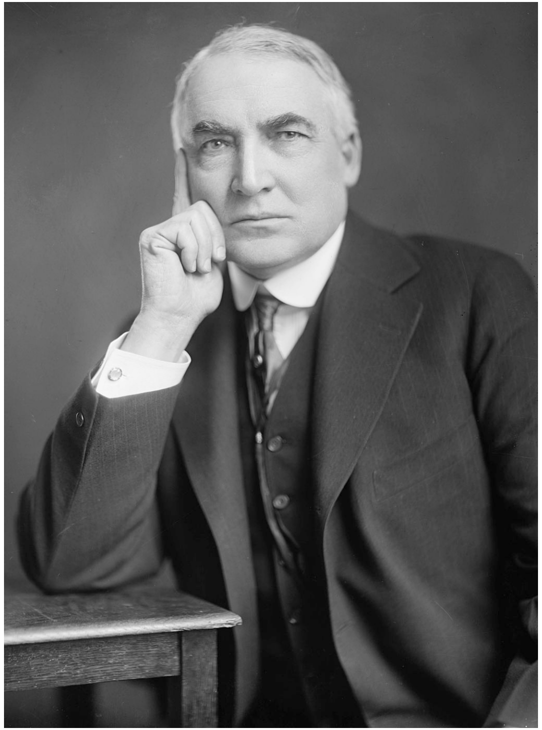
https://chat.openai.com/chat

What makes language hard?

- Speech acts (“can you pass the salt?)
  [Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
  [Grice 1975]
- Shared knowledge (“Warren ran for president”)



Elizabeth Warren
2020



Warren G. Harding
1920

What makes language hard?

- Speech acts (“can you pass the salt?)
  [Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
  [Grice 1975]
- Shared knowledge (“Warren ran for president”)
- Variation (“This homework is wicked hard”)
  [Labov 1966, Eckert 2008]

Ambiguity

"One morning I shot
an elephant in my pajamas"

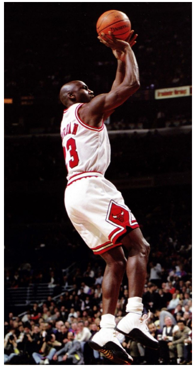
Animal Crackers

Ambiguity

"One morning I shot
an elephant in my pajamas"

Animal Crackers

Ambiguity



"One morning I shot
an elephant in my pajamas"

Ambiguity

verb noun

"One morning I shot
an elephant in my pajamas"

Animal Crackers

Ambiguity

A computer that understands you like your mother.

- *Computer understands you as well as your mother understands you.*

- *Computer understands (that) you like your mother*

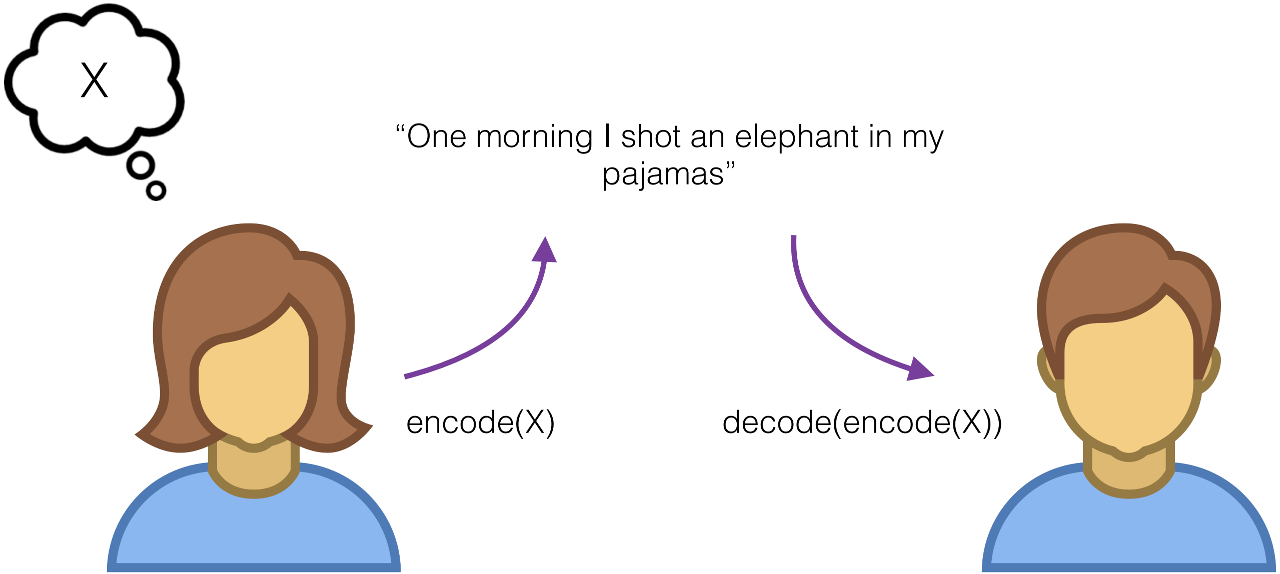
- *Computer understands you (as well as) it understands your mother*

Example by Stuart Shiebert

Processing as Representation

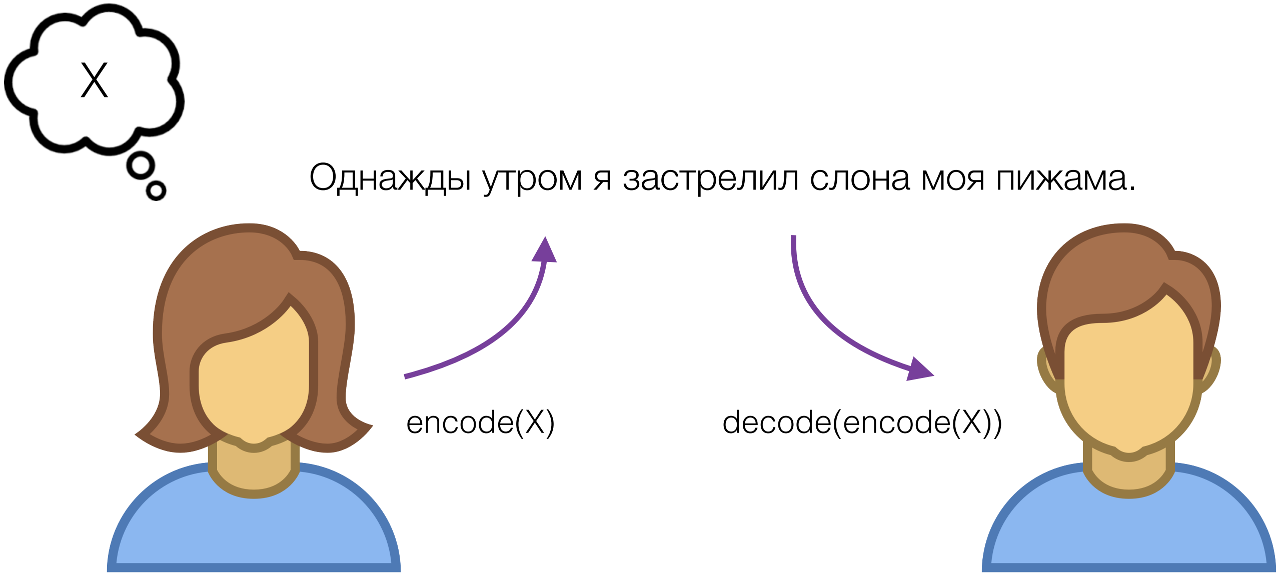
- NLP generally involves representing language for some end, e.g.:
    - Translation
    - Summarization
    - Analysis
    - Dialogue

Information theoretic view



Shannon 1948

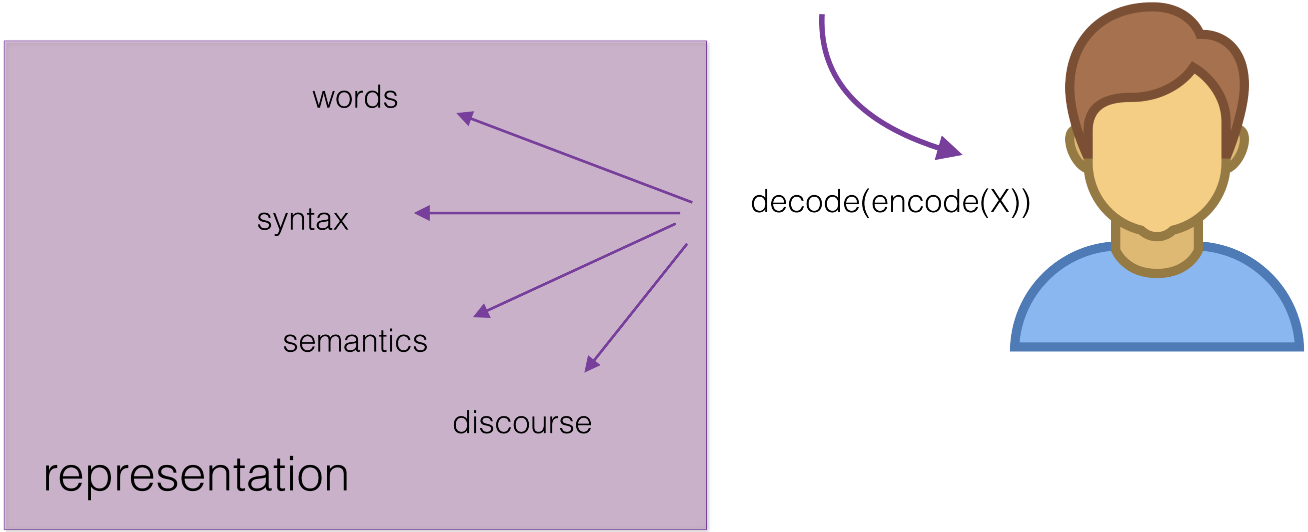
Information theoretic view

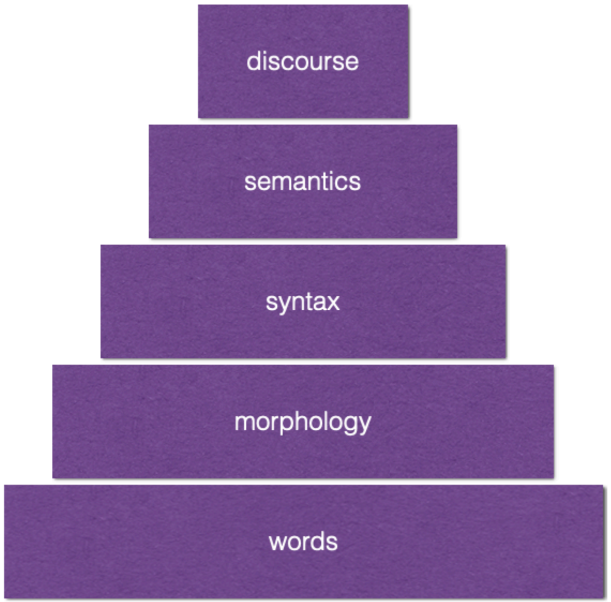


When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'
Weaver 1955

Decoding

"One morning I shot an elephant in my pajamas"




discourse
semantics
syntax
morphology
words

Words

- One morning I shot an elephant in my pajamas
- I didn’t shoot an elephant
- Imma let you finish but Beyonce had one of the best videos of all time
- 一天早上我穿着睡衣射了一只大象

Parts of speech

noun verb noun noun

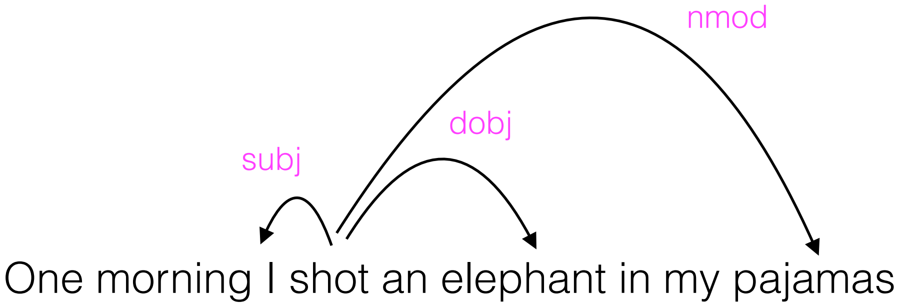
One morning I shot an elephant in my pajamas

Named entities

person Location

Dudley North travels to North London.

Syntax



Industrial NLP

- Machine translation
- Question answering
- Information extraction
- Conversational agents
- Summarization

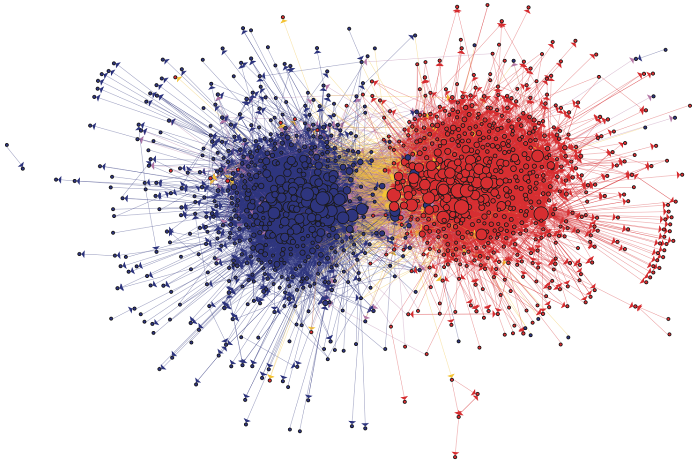
NLP Research Community

- Association for Computational Linguistics (ACL)
- Conferences: *ACL, NAACL, Empirical Methods for NLP (EMNLP), …*
- Publications: *Transaction for ACL (TACL), Computational Linguistics*
- Neurips, AAAI, ICML, ….

NLP + X

Computational Social Science

- Inferring ideal points of politicians based on voting behavior, speeches
- Detecting fake news
- Inferring power differentials in language use



Link structure in political blogs
Adamic and Glance 2005

Computational Journalism

What do Journalists do with Documents?
Field Notes for Natural Language Processing Researchers

Jonathan Stray
Columbia Journalism School
jms2361@columbia.edu

- Robust import
- Robust analysis
- Search, not exploration
- Quantitative summaries
- Interactive methods
- Clarity and Accuracy

NLP Methods

- Finite state automata/transducers (tokenization, morphological analysis)
- Rule-based systems

NLP Methods

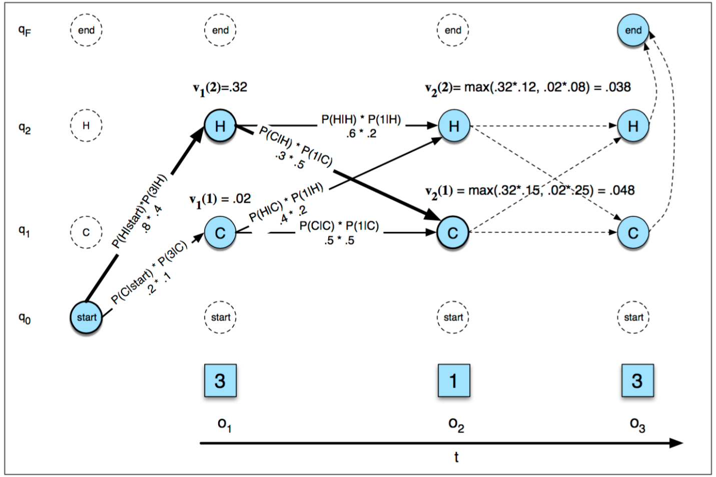
- Probabilistic models
- Naive Bayes, Logistic regression, HMM, MEMM, CRF, language models

$$P(Y=y|X=x) = \frac{P(Y=y)P(X=x|Y=y)}{\sum_y P(Y=y)P(X=x|Y=y)}$$

NLP Methods

- Dynamic programming (combining solutions to subproblems)

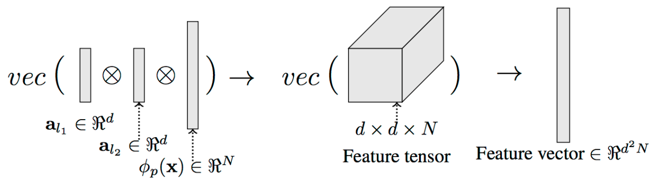
Viterbi algorithm,
CKY



Viterbi lattice, SLP3 ch. 9

NLP Methods

- Dense representations for features/labels (generally: inputs and outputs)



Srikumar and Manning (2014), "Learning Distributed Representations for Structured Output Prediction" (NIPS)

- Neural networks: multiple, highly parameterized layers of (usually non-linear) interactions mediating the input/output

Vaswani et al. (2017), "Attention is All You Need" (NeurIPS)

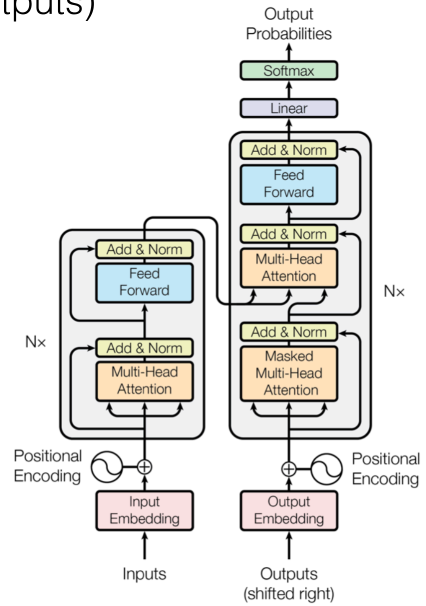
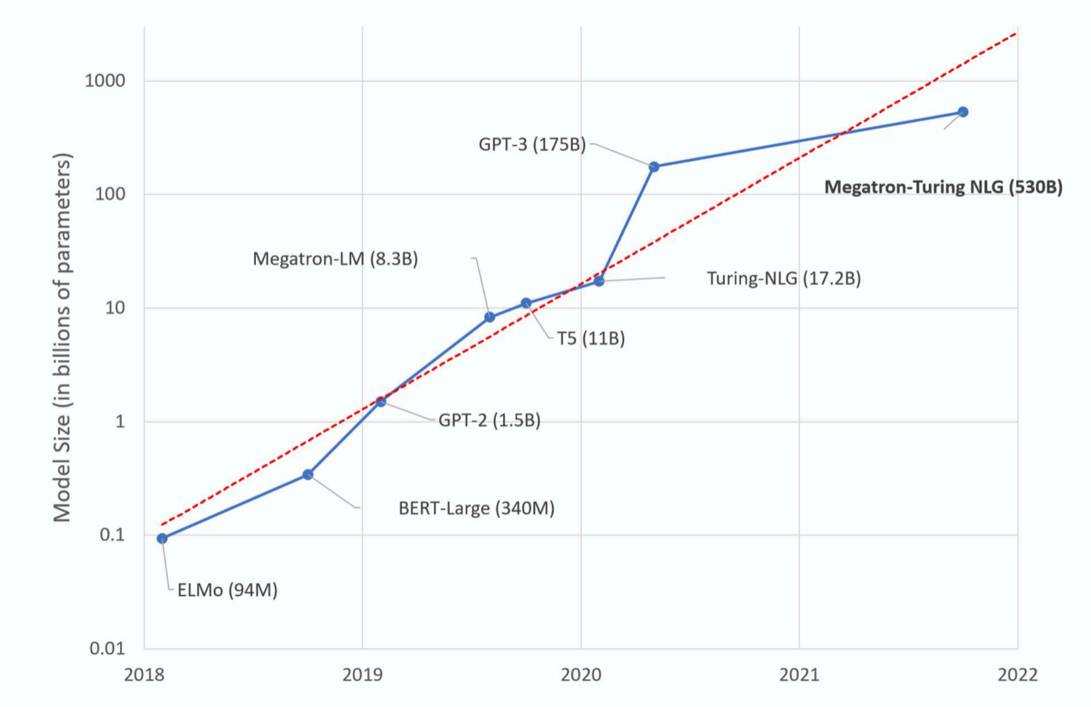


Figure 1: The Transformer - model architecture.

NLP Methods

- Pretraining highly parametrized large language models.



Info 159/259

- This is a class about NLP models and methods.
    - You'll learn and implement algorithms to solve NLP tasks efficiently and evaluate.
- This is a class about the linguistic representation of text.
    - You'll annotate texts so you'll understand the phenomena you'll be modeling

Three Themes

- Intro to statistical and neural methods:
    - Text Classification and Language modeling
- Modeling different linguistic layers
    - Syntax, Semantics, Discourse
- Applications
    - Information Extraction, Machine Translation, Question Answering, …

Why Classic NLP?

- With neural models we have moved towards abstraction of linguistic representation.
- Why do we study classic NLP methods?
    - We are far away from solving the language problem.
        - Learn about the SOA and their shortcomings
    - Not every NLP scenario affords using heavy data, infrastructure, etc.
    - Ethical and legal concerns

Prerequisites

- Strong programming skills
    - Translate pseudocode into code (Python)
    - Analysis of algorithms (big-O notation)
- Basic probability/statistics
- Calculus

function VITERBI(*observations* of len T, *state-graph* of len N) **returns** *best-path*

create a path probability matrix *viterbi[N+2,T]*
for each state s **from** 1 **to** N **do** ; initialization step
$viterbi[s,1] \leftarrow a_{0,s} * b_s(o_1)$
$backpointer[s,1] \leftarrow 0$
for each time step t **from** 2 **to** T **do** ; recursion step
for each state s **from** 1 **to** N **do**

$$viterbi[s,t] \leftarrow \max_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$$

$$backpointer[s,t] \leftarrow \operatorname*{argmax}_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s}$$

$$viterbi[q_F,T] \leftarrow \max_{s=1}^{N} \; viterbi[s,T] * a_{s,q_F}$$; termination step

$$backpointer[q_F,T] \leftarrow \operatorname*{argmax}_{s=1}^{N} \; viterbi[s,T] * a_{s,q_F}$$; termination step

return the backtrace path by following backpointers to states back in time from $backpointer[q_F,T]$

Viterbi algorithm, SLP3 ch. 8

Grading

- Info 159:
    - Homeworks (25%)
    - Annotation project (25%)
    - Quizzes (10%)
    - Exams (20%)
    - NLP subfield survey (20%)

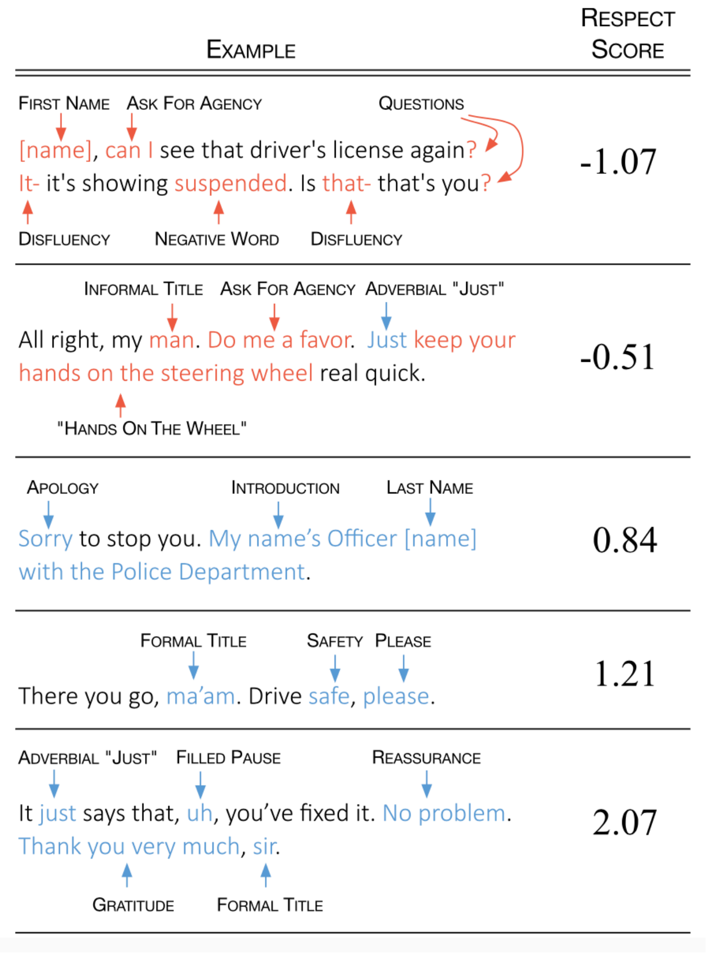
Annotation project

- This course covers many of the methods and existing tasks in NLP →
- But the most exciting applications of NLP have yet to be invented.
- Design a new NLP task and annotate data to support it, working in groups of exactly 3 students.

Existing tasks
Question answering
Named entity recognition
Sentiment analysis
Machine translation
Syntactic parsing
Coreference resolution
Text generation
Word sense disambiguation
…

Respect

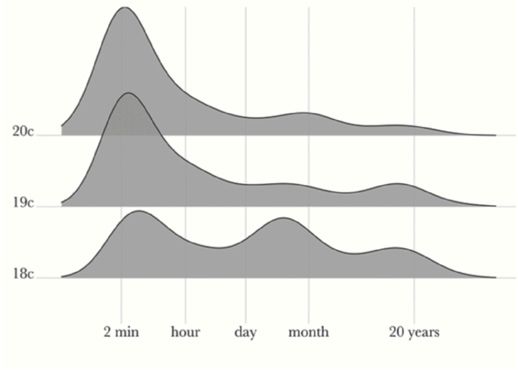
- Present one dialogue turn (police/driver) to be rated by people for respect (4-point Likert scale). High IAA.
- Build a predictive model mapping text to respect.



Voigt et al. 2017, "Language from police body camera footage shows racial disparities in officer respect"

Time

Underwood (2018), “Why Literary Time is Measured in Minutes”. Measuring how much time has passed in 250-word chunks of text.



Underwood (2018), “Why Literary Time Is Measured in Minutes”

Passage	Mins
I fear then, Emma, Sewell is a knave, and joined in mean collusion with his brother, to distress your husband, who looks upon him as his friend. You are deceived, Charles, I am sure he is Sir James's friend, and mine, by his perpetually dissuading him from play. It may be so; but tell me, Emma, all you know, and all you think of Lady Juliana's sudden departure, what can it mean? …	5.0
At length we reached the gates of this noble edifice, and had the pleasure to find the family not retired to rest, by perceiving lights in the hall. … In a few minutes all was hushed, and a man, whom I believed to be an upper servant, was sent to reconnoitre my person, and enquire my name and business. I told him I should not reveal either, but to his master. …	15.0

AP deliverables

- AP1. Design a new document classification task and gather data to support it (must be shareable with the public — nothing private or in copyright).
- AP2. Create robust set of annotation guidelines, annotate the data, creating at least 1000 labeled examples + reporting inter-annotator agreement rates.
- AP3. In a separate assignment, a different group will annotate your data only using your annotation guidelines (are your guidelines comprehensive enough that an independent third party could reproduce your judgments?).
- AP4. Build a classifier to automatically predict the labels using the data you've annotated.

NLP subfield survey

- 4-page survey for a specific NLP subfield of your choice (e.g., coreference resolution, question answering, interpretability, narrative generation, etc.), synthesizing at least 25 papers published at ACL, EMNLP, NAACL, EACL, AACL, *Transactions of the ACL* or *Computational Linguistics*.

- This survey should be able to provide a newcomer (such as yourself at the start of the semester) a sense of the current state of the art in that subfield in 2023, the major historical papers that have defined that area, and the different schools of thought within it.

Grading

- Info 259:
    - Homeworks (20%)
    - Annotation project (20%)
    - Quizzes (10%)
    - Exams (20%)
    - NLP Project (30%)

259 NLP Project

- Semester-long project (involving 1-3 students) involving natural language processing -- either focusing on core NLP methods or using NLP in support of an empirical research question
    - Project proposal/literature review
    - Midterm report
    - 6-page final report, workshop quality
    - Poster presentation

ACL 2024 workshops

- Teaching Natural Language Processing
- Workshop on Computational Approaches to Historical Language Change
- Workshop on Privacy in NLP
- Natural Language Reasoning and Structured Explanations
- The Machine Learning for Ancient Languages Workshop
- Language + Molecules
- Social Media Mining for Health Research and Applications Workshop
- Workshop on Data Contamination
- Cross-Cultural Considerations in NLP
- Workshop on Scholarly Document Processing
- Wordplay: When Language Meets Games Workshop
- Natural Language Processing meets Climate Change
- Workshop on Spatial Language Understanding and Grounded Communication for Robotics

Exams

- We’ll have two exams:
  - Exam 1 (2/21, 6:30-8, remote).
  - Exam 2 (3/20, 6:30-8, remote).
- We will not be offering alternative exam dates, so if you anticipate a conflict, don’t take this class!
- Your exam grade will be the max(exam1, exam 2) — you will drop your lowest-scoring exam grade.

Late submissions

- All homeworks and quizzes are due on the date/time specified.
- You have 3 late days total over the semester to use when turning in homeworks/quizzes (not group annotation project deliverables or 259 project deliverables); each day extends the deadline by 24 hours. If all late days have been used up, homeworks/quizzes can be turned in up to 48 hours late for 50% credit; anything submitted after 48 hours late = 0 credit.
- Late days are assessed immediately once homeworks or quizzes are submitted late and can't be retroactively changed (if you submit 2 homeworks and 2 quizzes late, for example, you can't decide after the fact which ones to apply your 3 slip days to -- they apply to whichever homeworks or quizzes use them up first).

Academic integrity

- We’ll follow the UC Berkeley code of conduct http://sa.berkeley.edu/code-of-conduct
- You may discuss homeworks at a high level with your classmates (if you do, include their names on the submission), but each homework deliverable must be completed independently -- all writing and code must be your own; and all quizzes and exams must be completed independently.

Academic integrity

- If you mention the work of others, you must be clear in citing the appropriate source: http://gsi.berkeley.edu/gsi-guide-contents/academic-misconduct-intro/plagiarism/

- This holds for source code as well: if you use others' code (e.g., from StackOverflow), you must cite its source.

- We have zero tolerance policy for cheating and plagiarism; violations will be referred to the Center for Student Conduct and will likely result in failing the class.

AI Assistants

- If you use the output of automatic writing assistants (e.g. ChatGPT, Bard) or code suggestions (e.g. Copilot) you must cite the source and be clear what code/text came form it — just as you would with anything else you did not create yourself.
- You retain responsibility for anything you submit and should be prepared to demonstrate your understanding of it
- Be honest about your use of these tools; if multiple students submit the same text/code, easy to find plagiarism.
- The work you submit should largely be your own, representing your own ideas, code and words; submissions that rely too heavily on AI tools will not be graded favorably.

Curve

Grades in this class will not be curved.

Lectures

- Recordings of lectures will be available on bCourses through Course Capture.
- Course capture can fail due to technical/human errors. In that case, recordings won’t be repeated.
- Active attendance is highly recommended.

Ed Discussion

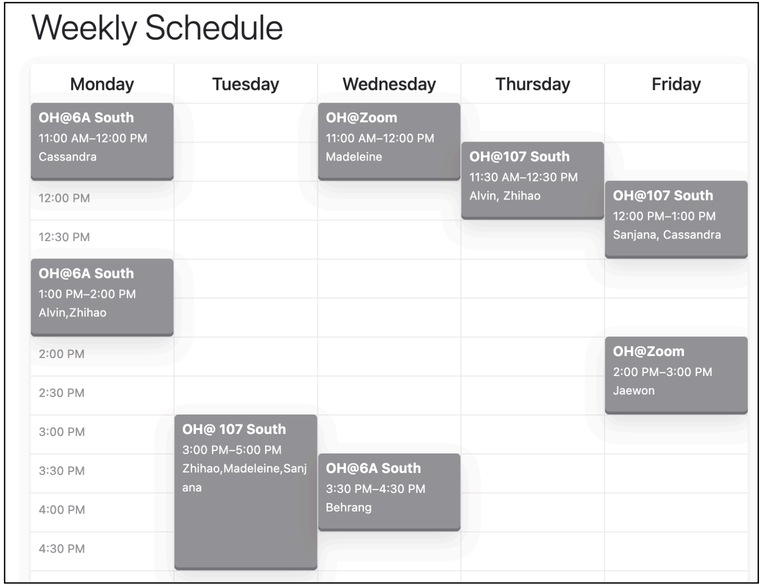
- We'll use Ed Discussion as a platform for asking and answering questions about the course material, including homeworks.

- Students are encouraged to actively participate on this forum and help others by answering questions that arise (helpful students can see a grade bump across a threshold (e.g., B+ to A-) for this participation.

- When helping with homework questions, keep the discussion to the high-level concepts; don't post answers to homeworks or quiz/exam questions.

TAs’ Office Hours

- Visit TA office hours for help with homeworks/quizzes/exams/projects or just to chat about NLP.
- Office hours start next Tuesday (Jan 23)



TAs’ Office Hours

- Keep academic integrity in mind during TA office hours:
  - Discuss hw questions at a high level with others present
  - don't discuss specific answers or share screens with code solutions. Neither the TA office hours nor Ed Discussion should be used for pre-grading

Behrang's office hours

- In Person: Wednesdays 3:30-4:30, South Hall 6A

Next time:

Words/Lexical semantics/Word embeddings