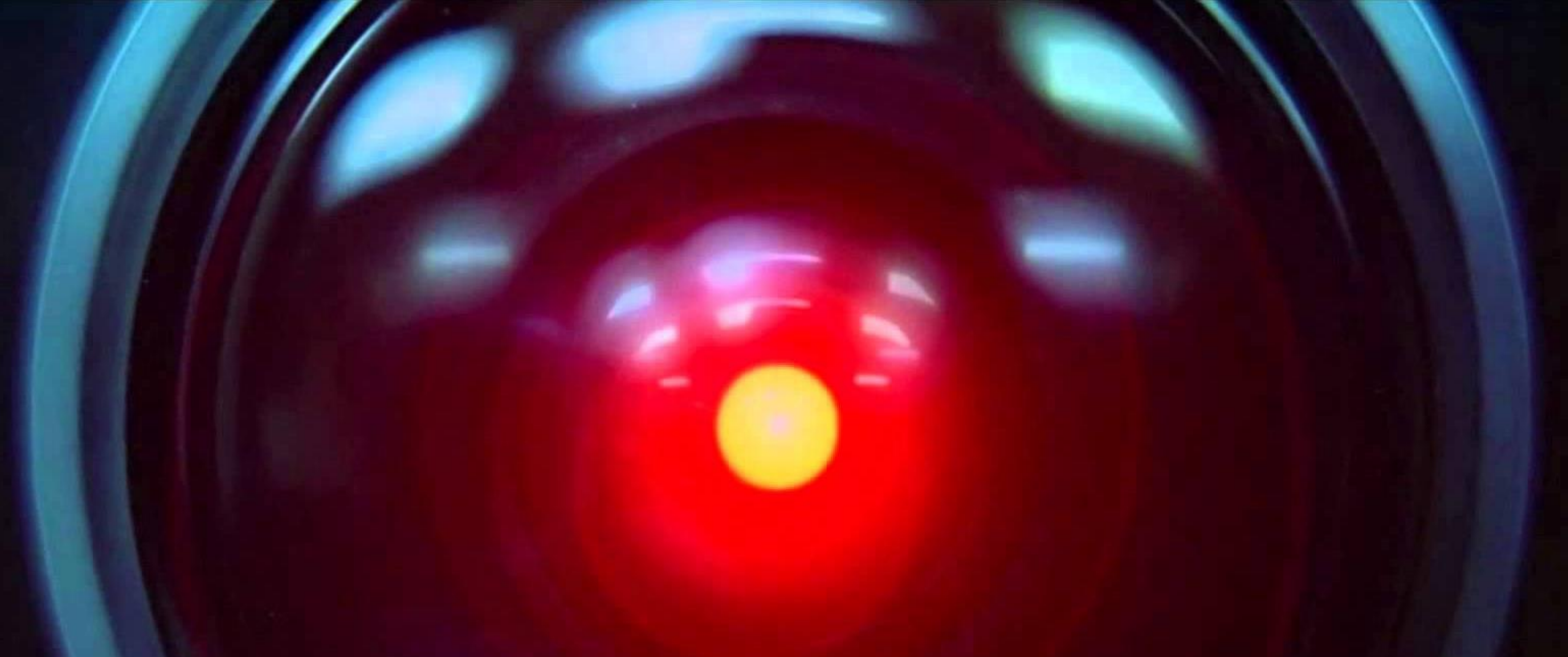


2001: A Space Odyssey (1968), *Stanley Kubrick and Arthur C. Clarke*

- A Conversation with HAL! *https://www.youtube.com/watch?v=r13I-TuDcWI*

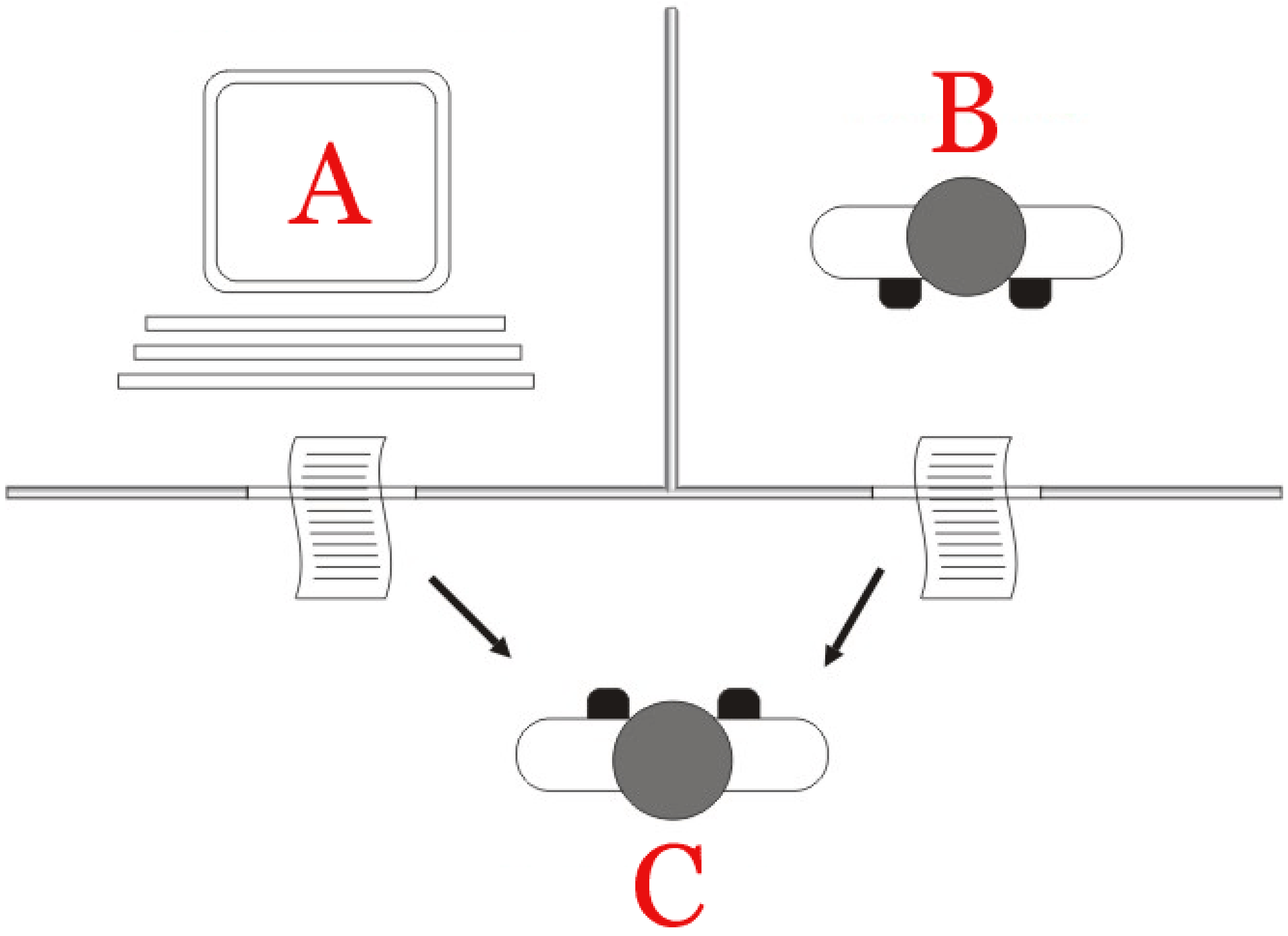
- HAL Reads Lips! *https://www.youtube.com/watch?v=XDO8OYnmkNY*

– HAL: I'm Sorry, Dave! *https://www.youtube.com/watch?v=Wy4EfdnMZ5g*

The Turing Test

by Alan Turing in 1950

Player C, the interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination.



Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments

https://www.youtube.com/watch?v=D5VN56jQMWM

Natural Language Understanding

Speech, *Text*, Emotion, ...

Research Priorities for Artificial Intelligence

The capacity for language is one of the central features of human intelligence and is therefore a prerequisite for artificial intelligence.

Despite its many practical applications, language is perhaps number 300 in the priority list for AI research. It would be a great achievement if AI could attain the capabilities of an orangutan, which do not include language!
- Yann LeCun (computer vision researcher)

Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research priorities for robust and beneficial artificial intelligence." *Ai Magazine* 36.4 (2015): 105-114.

Handle: Boston Dynamics' newest design. Jumps 4 feet in the air and zips around at 9 miles per hour. https://www.youtube.com/watch?v=7h8mX9ZMs7g

NLP is AI-hard (AI-Complete)

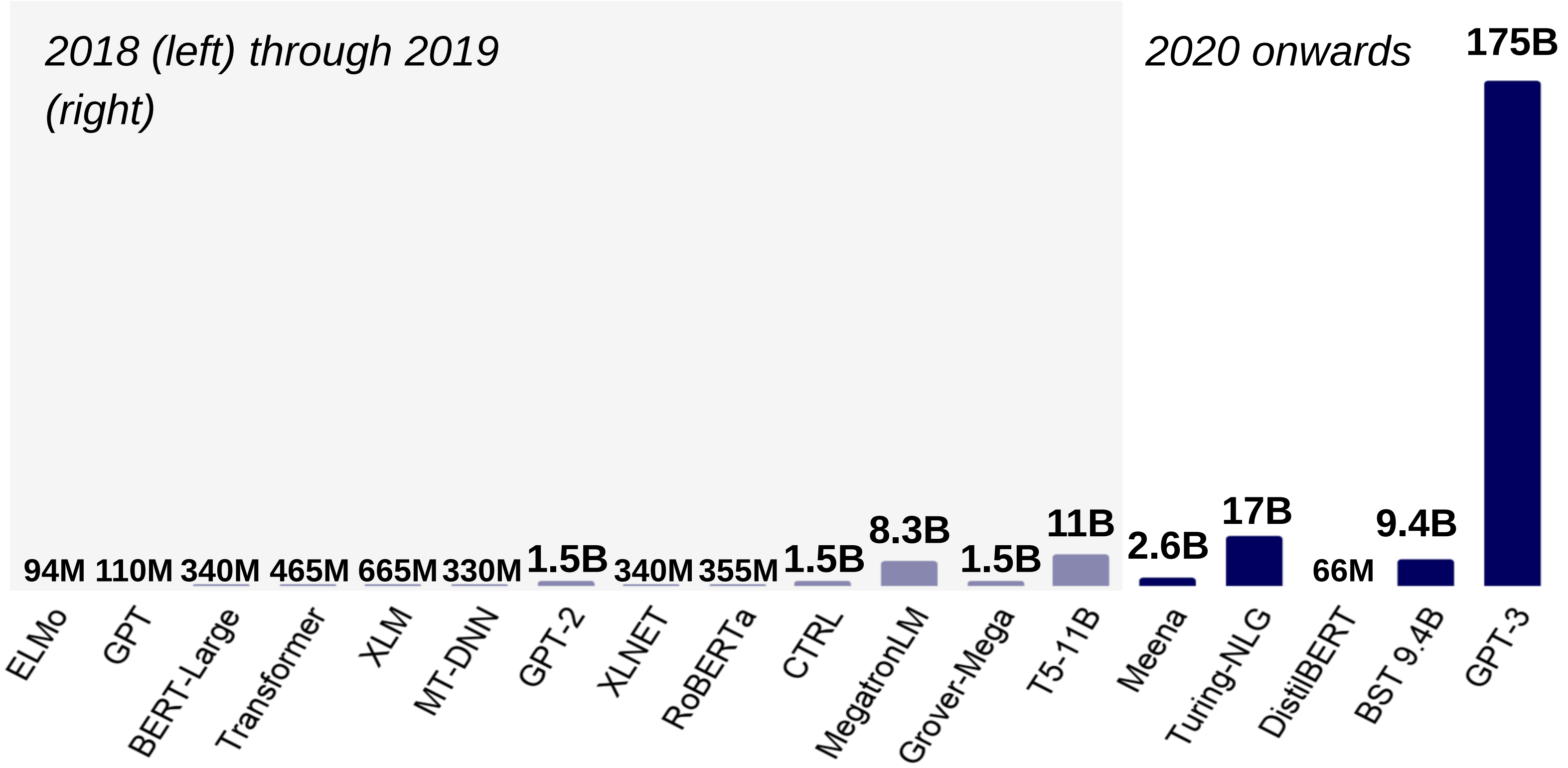
State of AI Report

October 1, 2020

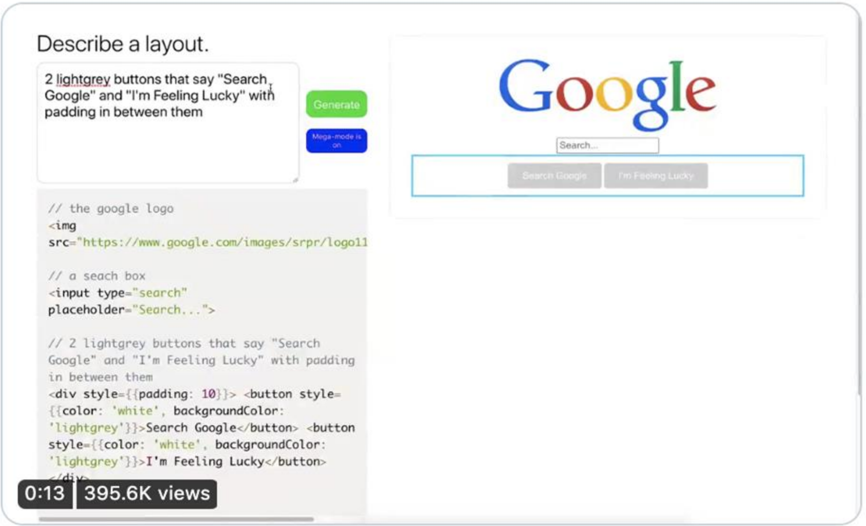
Nathan Benaich Ian Hogarth

Language models: Welcome to the Billion Parameter club

- **Huge models, large companies and massive training costs dominate the hottest area of AI today, NLP.**



Note: The number of parameters indicates how many different coefficients the algorithm optimizes during the training process.



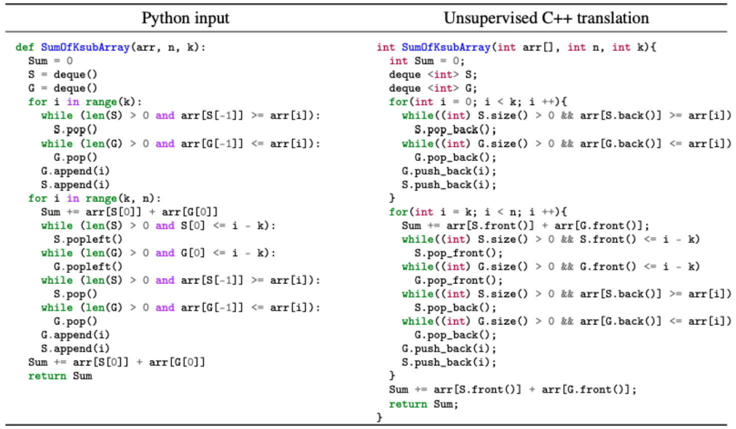
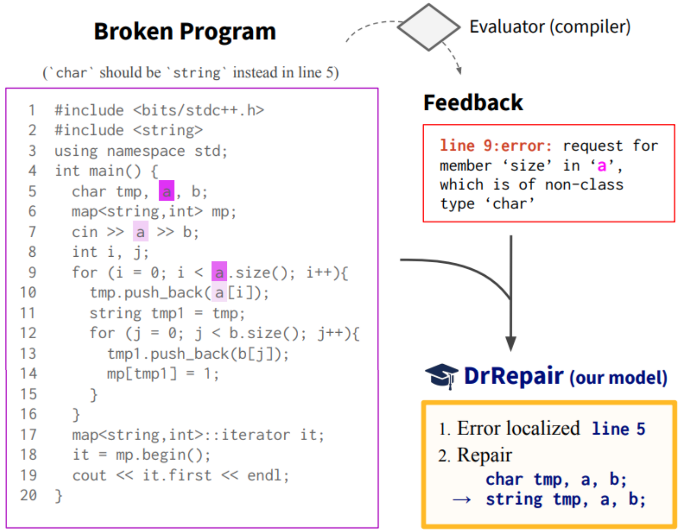


Figure 2: **Example of unsupervised Python to C++ translation.** TransCoder successfully translates the Python input function SumOfKsubArray into C++. TransCoder infers the types of the arguments, of the variables, and the return type of the function. The model maps the Python deque() container, to the C++ implementation deque<>, and uses the associated front, back, pop_back and push_back methods to retrieve and insert elements into the deque, instead of the Python square brackets [], pop and append methods. Moreover, it converts the Python for loop and range function properly.

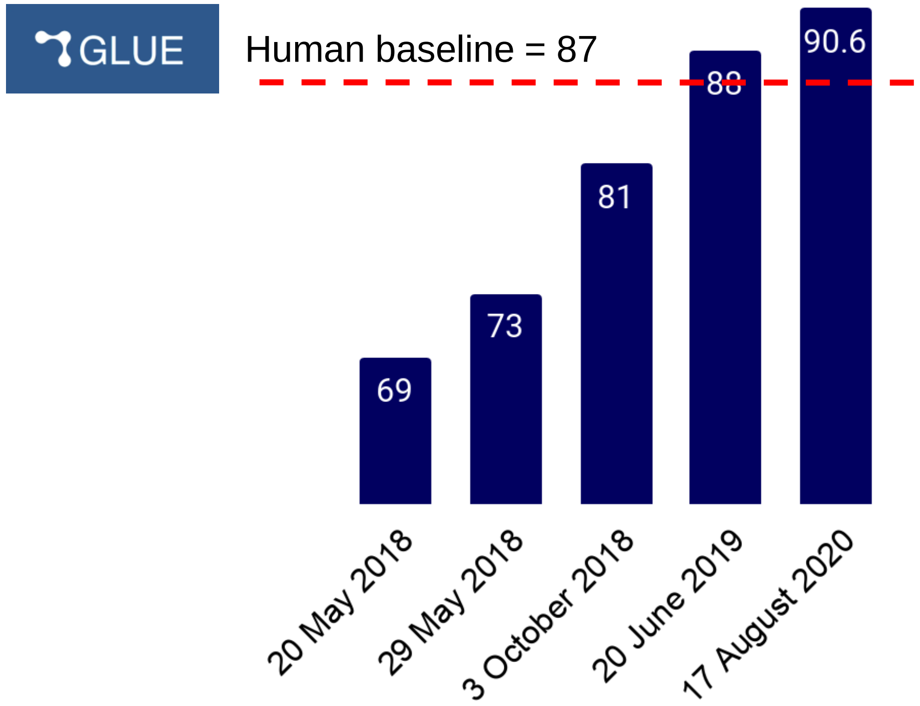


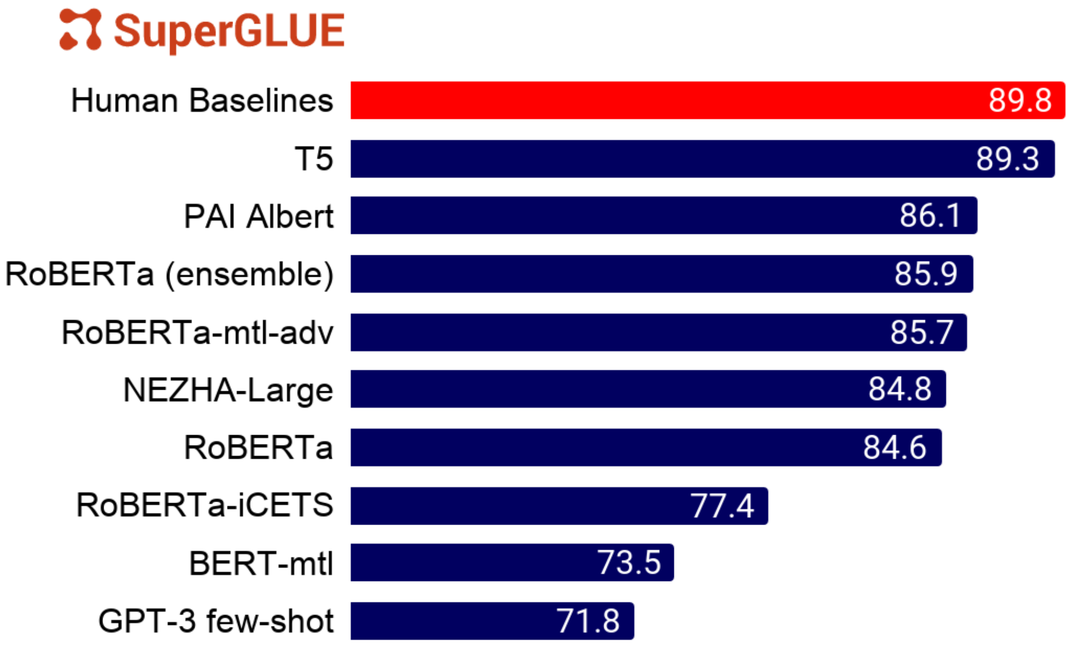
Example taken from SPoC dataset (909A-45398788.cpp)

NLP benchmarks take a beating: Over a dozen teams outrank the human GLUE baseline

It was only 12 months ago that the human GLUE benchmark was beat by 1 point. Now SuperGLUE is in sight.

- GLUE and it's more challenging sibling SuperGLUE are benchmarks that evaluate NLP systems at a range of tasks spanning logic, common sense understanding, and lexical semantics. The human benchmark on GLUE is reliably beat today (right) and the SuperGLUE human benchmark is almost surpassed too!





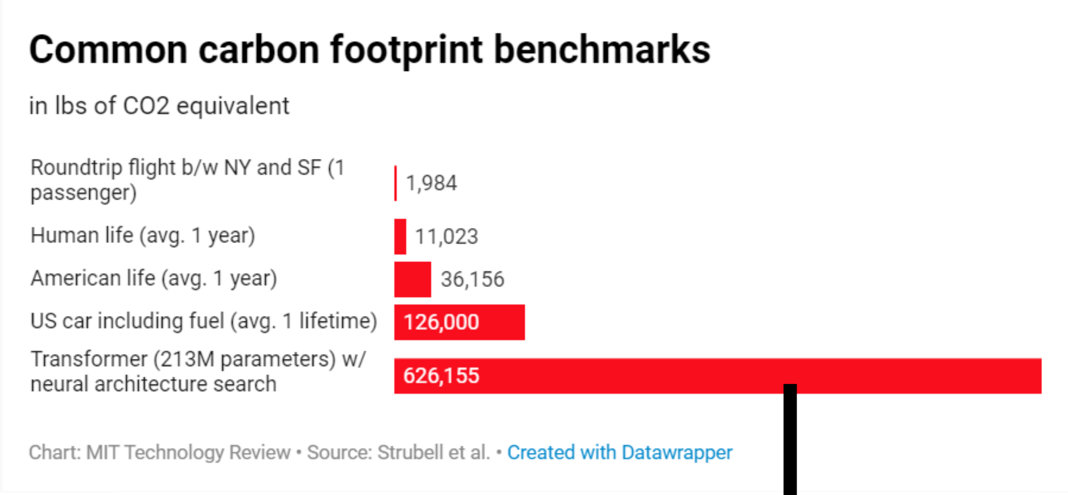
MIT Technology Review Topics

Artificial intelligence / Machine learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by **Karen Hao** June 6, 2019



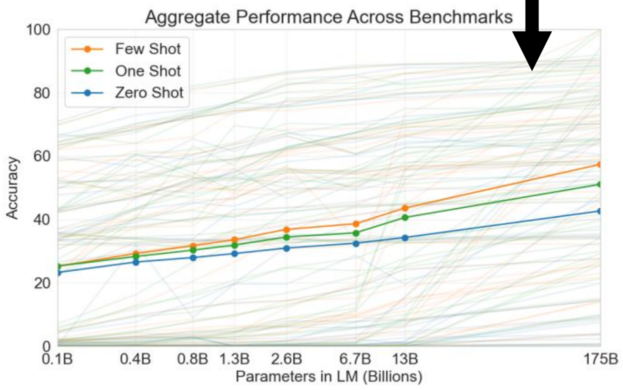


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

Natural
Language
Processing & Understanding

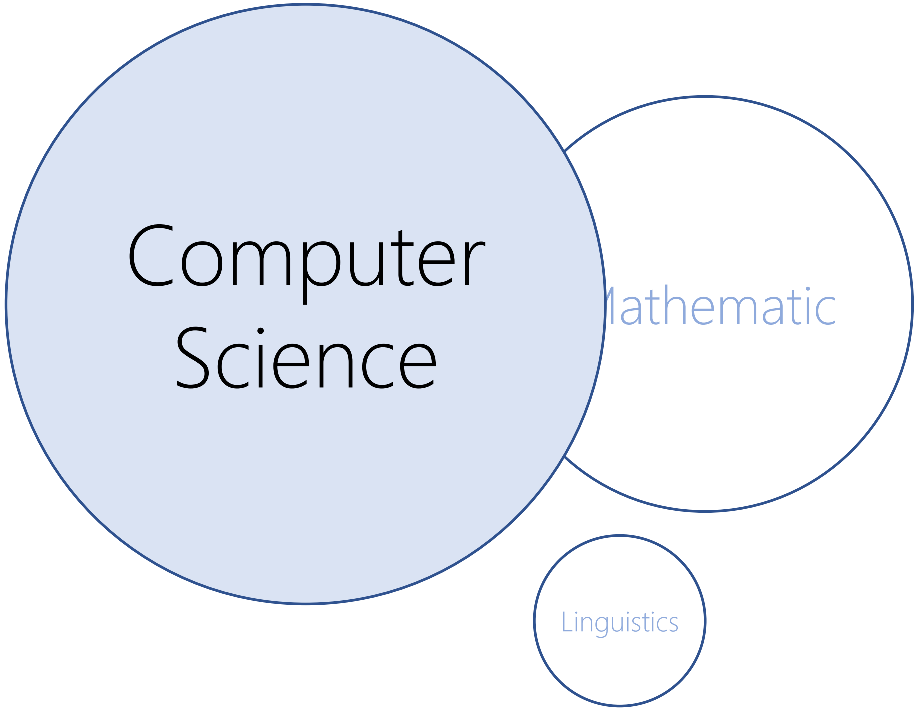
COMP8730 Winter 2021

Image Source: https://unanimous.ai/chat-with-a-different-kind-of-artificial-intelligence/

Background: *the course is targeted at computer scientists!*

Assumed to know

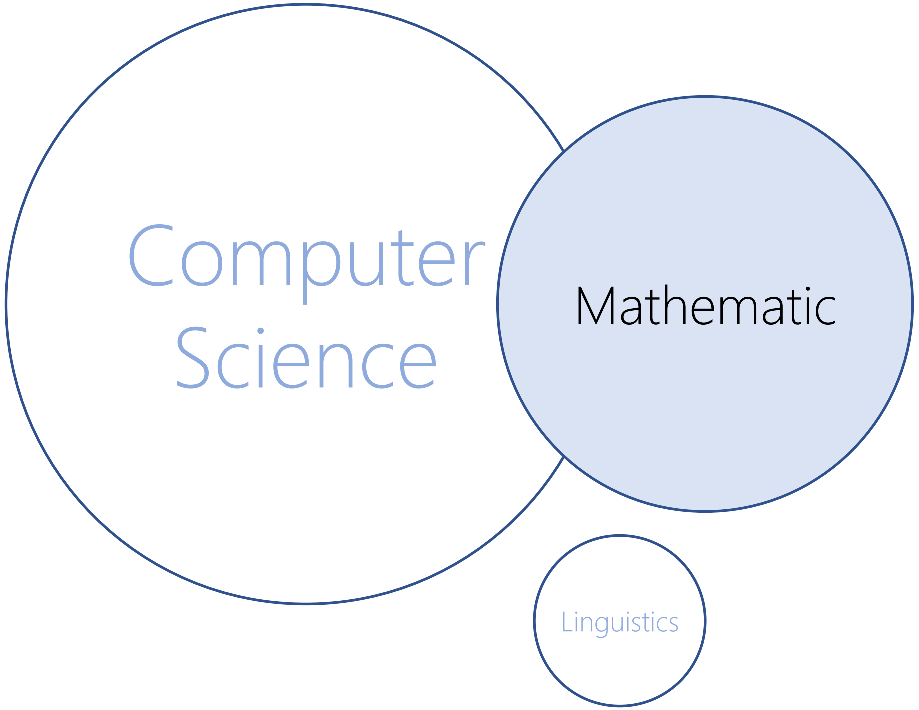
- Design of Algorithms
  - Greedy
  - Dynamic Programming
  - Divide-Conquer
  - Recursion
  - Backtracking
- Analysis of Algorithms
  - Time & Space (memory)
  - Big O
  - Complexity Theory
- Data Modeling
  - Data Structure



Background: *the course is targeted at computer scientists!*

Assumed to know

- Multivariate Calculus
    - Derivatives
    - Partial Derivatives
- Linear Algebra
    - Vectors
    - Matrices
- Probability & Statistics



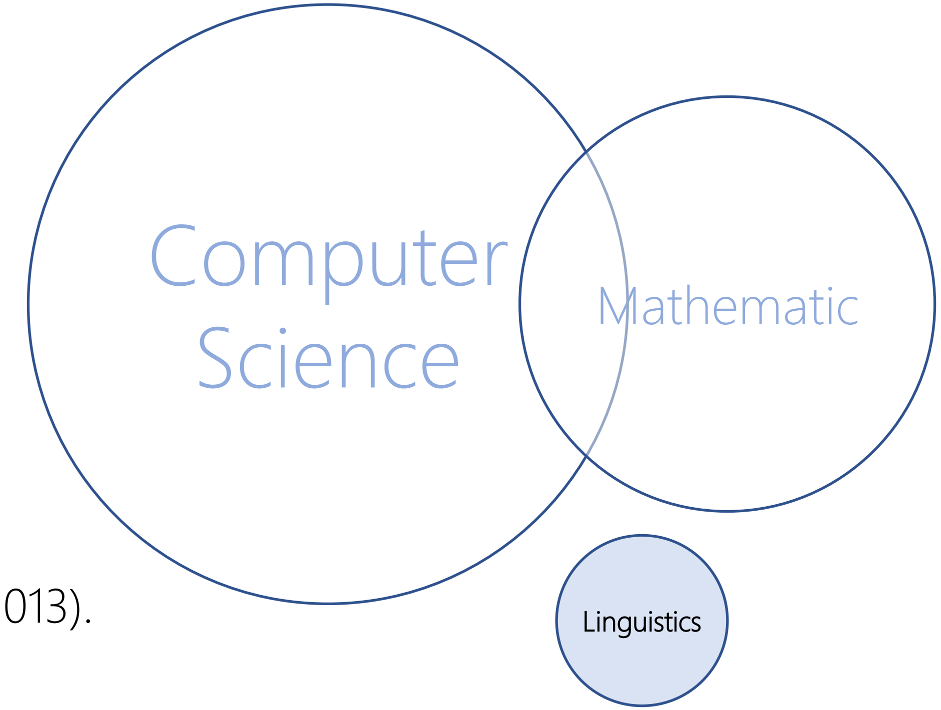
Background: *the course is targeted at computer scientists!*

Assumed to know

- Elementary concepts
    - Part-of-Speech (Nouns, Verbs, …)
    - English Grammar

References

- Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax, Bender, E. M. (2013).



Natural Language Processing & Text Mining

The goal is to provide new computational capabilities for applications

E.g., *predict next form of a word for branding!*

- extracting information from texts,
- translating between languages,
- answering questions,
- holding a conversation,
- taking instructions,
- ....

Computational Linguistics

Here, language is the object of study.
Computational methods are to support.
Just as in computational biology

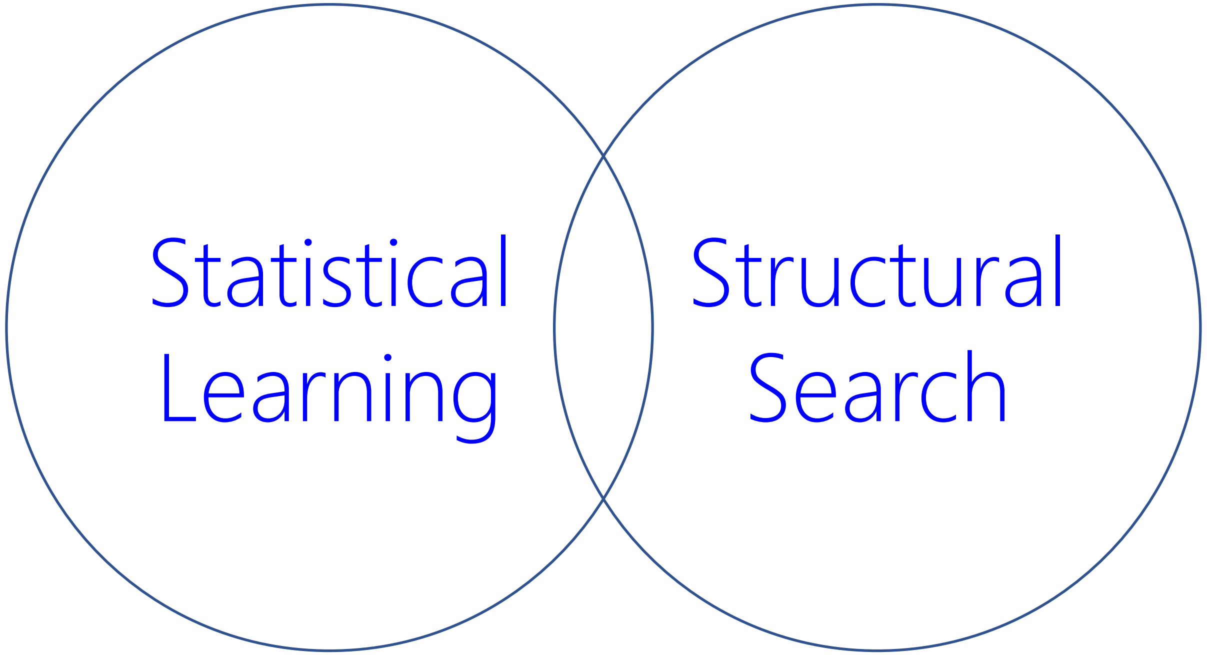
E.g, *how a word is evolving in time?*

- Discourse analysis
- Parsing

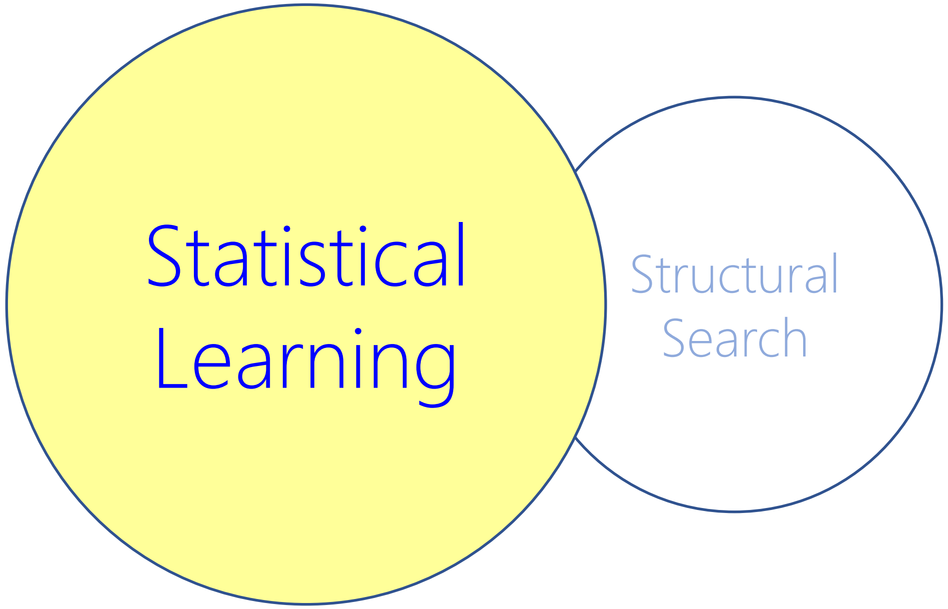
Machine Learning & Data Mining

Methods of learning from data.
Text data needs special care! Why?

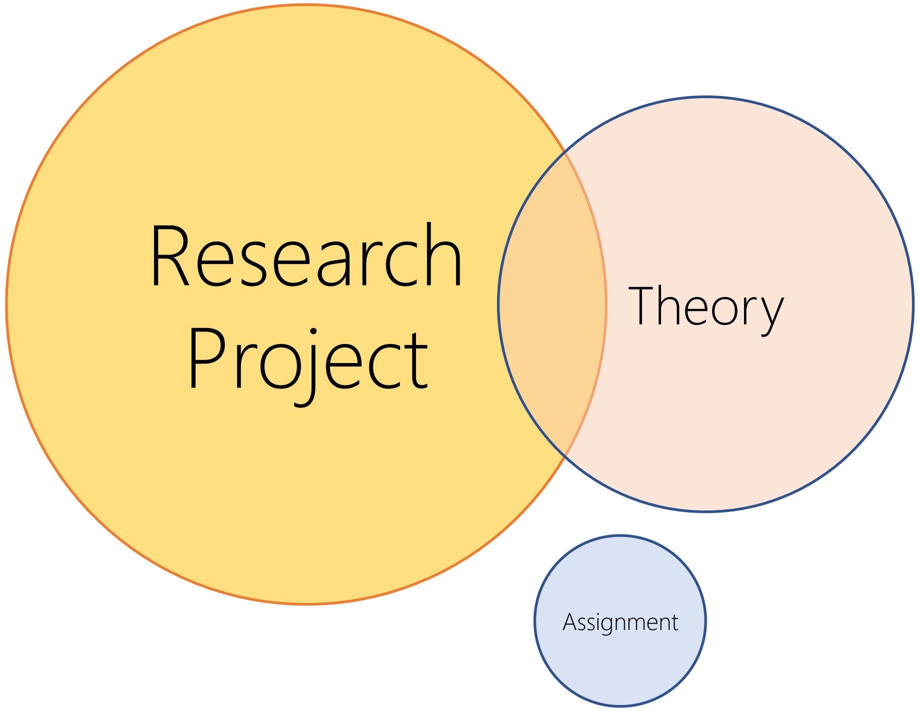
Approach: *unstructured vs. structured*



Approach: *unstructured vs. structured*



Course: Research-oriented, Project-driven

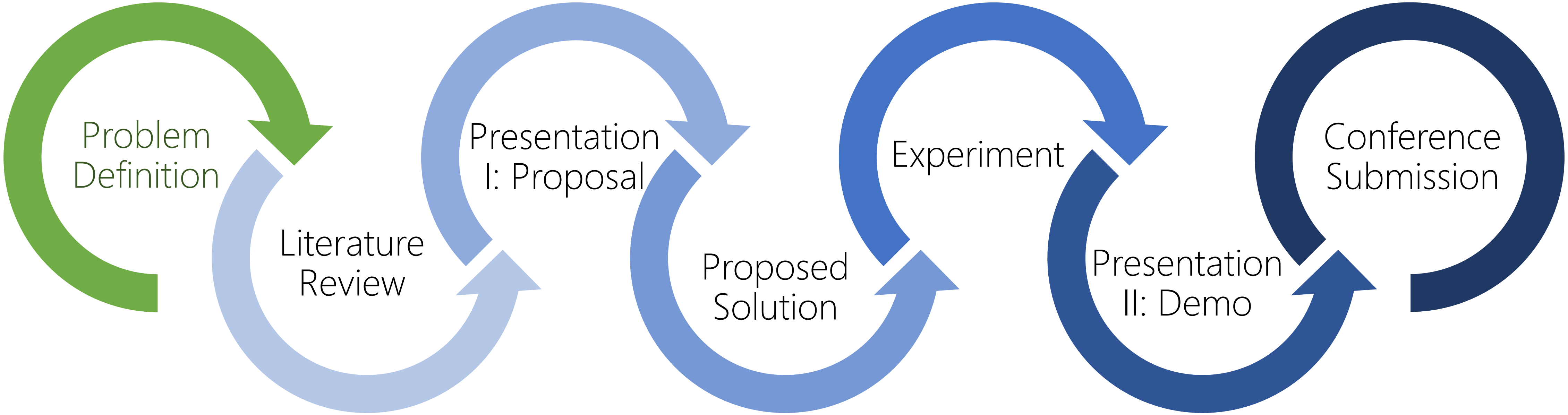


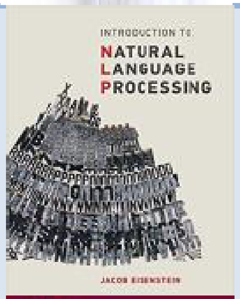
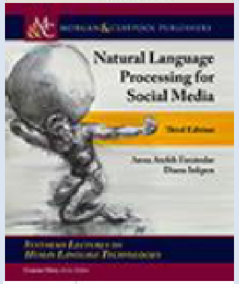
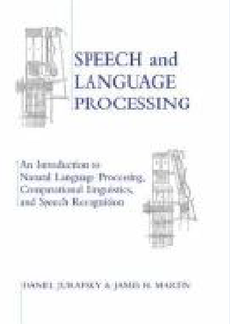
RESEARCH PROJECT

vs.

SOFTWARE PROJECT

Research Project



Books	Introduction to Natural Language Processing Jacob Eisenstein ISBN: 9780262042840 536 pages October 2019 http://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf	
	Natural Language Processing for Social Media, Third Edition *Synthesis Lectures on Human Language Technologies* Anna Atefeh Farzindar, Diana Inkpen ISBN: 9781681738116 | PDF ISBN: 9781681738123 Hardcover ISBN: 9781681738130 Copyright © 2020 | 219 Pages DOI: 10.2200/S00999ED3V01Y202003HLT046	
	Speech and Language Processing, 3rd Edition Draft An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Dan Jurafsky and James H. Martin Free Access → https://web.stanford.edu/~jurafsky/slp3/	
Hands-on Resources	NLP with Python – Analyzing Text with the Natural Language Toolkit https://www.nltk.org/book/ NLP with PyTorch: Build Intelligent Language Applications Using Deep Learning. https://github.com/joosthub/PyTorchNLPBook	

Marking Scheme	***Research* Project** - Problem Definition - Literature Review - Presentation I: Proposal - Proposed Solution - Experiment - Presentation II: Demonstration - Conference Submission (Pending Instr. 's Approval)	65% - 05% - 10% - 05% - 10% - 20% - 05% - 10%
	Assignments (+peer review)	15%
	Midterm Exam	10%
	Final Exam	10%
Remarks	The written reports will be assessed not only on their academic merit but also on the student's communication skills as exhibited through the reports. To achieve a passing grade, the students must achieve at least 70% of the entire marking scheme. The fraction mark is rounded to the ceiling. The students earn final course grades as per the Senate policy for Grading and Calculation of Averages.	

OFFICE

Tuesday-Thursday 5:30 PM-6:30 PM

NATURAL LANGUAGE

Engaging in Natural Language Behavior

- 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 & intentions of the speaker
- Discourse
  knowledge about linguistic units larger than a single utterance

Stephen Hawking talks about technology and ACAT (Assistive Context Aware Tool-kit)
https://www.youtube.com/watch?v=txy8ikfekzs
https://01.org/acat/
image: Ted S. Warren/AP

Engaging in Natural Language Behavior

- Phonetics and Phonology

    knowledge about linguistic sounds
    how words are pronounced in terms of sequences of sounds
    how each of these sounds is realized acoustically

Engaging in Natural Language Behavior

- Phonetics and Phonology

knowledge about linguistic sounds
how words are pronounced in terms of sequences of sounds
how each of these sounds is realized acoustically

Speech Recognition (SR):

recognize words from an audio signal like in assistants: Alexa, Homepod

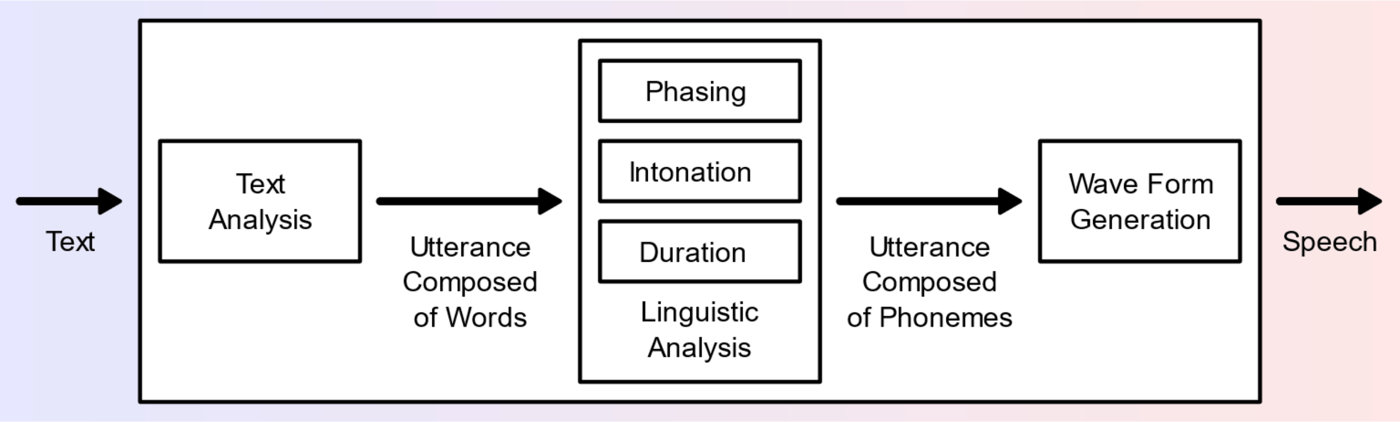
Speech Synthesis (Synthesizers):

generate an audio signal from a sequence of words like in Automatic Announcement
Automatic Answering Machine

Engaging in Natural Language Behavior

- Phonetics and Phonology

knowledge about linguistic sounds
how words are pronounced in terms of sequences of sounds
how each of these sounds is realized acoustically



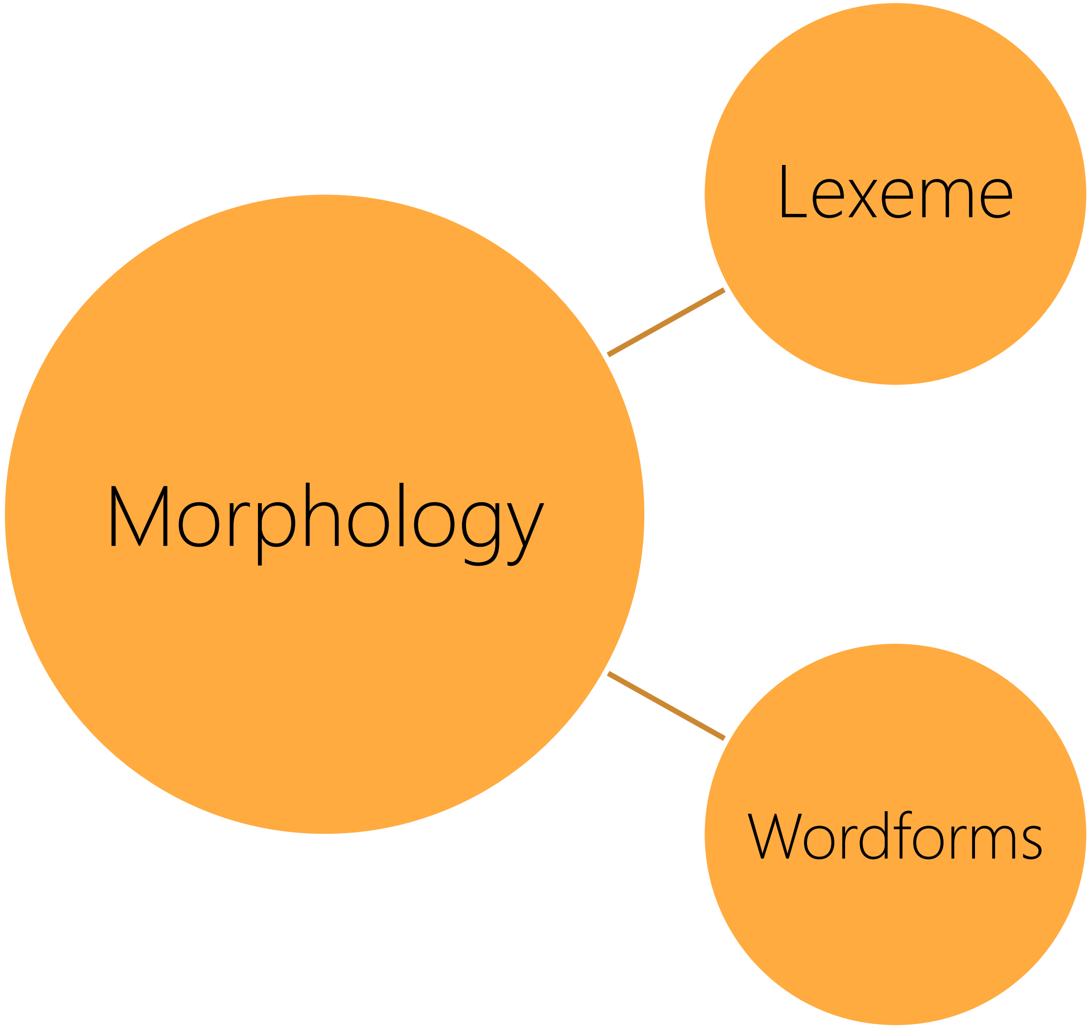
Text-To-Speech System (TTS)

Phone [fəʊn]→Diphones [fə], [əʊ], [ʊn]→much more natural than combining simple phones

Engaging in Natural Language Behavior

- Morphology
  - knowledge of the meaningful components of words
  - producing and recognizing variations of individual words
  - the way words break down into component parts that carry different meanings
  - study of words, how they are formed, their relationships in the same language

Engaging in Natural Language Behavior



Lexeme

- AKA Lemma or citation form
- Refer to a same entity or concept or …
- Change is called Inflection
- Inflection Rules:

  Singular vs. Plural: index → indexes, indices
  Contractions: cannot → can't
  Tenses: do → did, done, does

Wordforms

- Form new lexemes
- Refer to different entities or concepts or …
- Wordformation Rules:

  Compounding: [Dog][catch][er], [Dish][wash][er]
  Lemma(Dishwashers) = Dishwashers
  Lemma(Dishwashers) ≠ Dishwash ≠ Dish ≠ Wash

Engaging in Natural Language Behavior

- Syntax

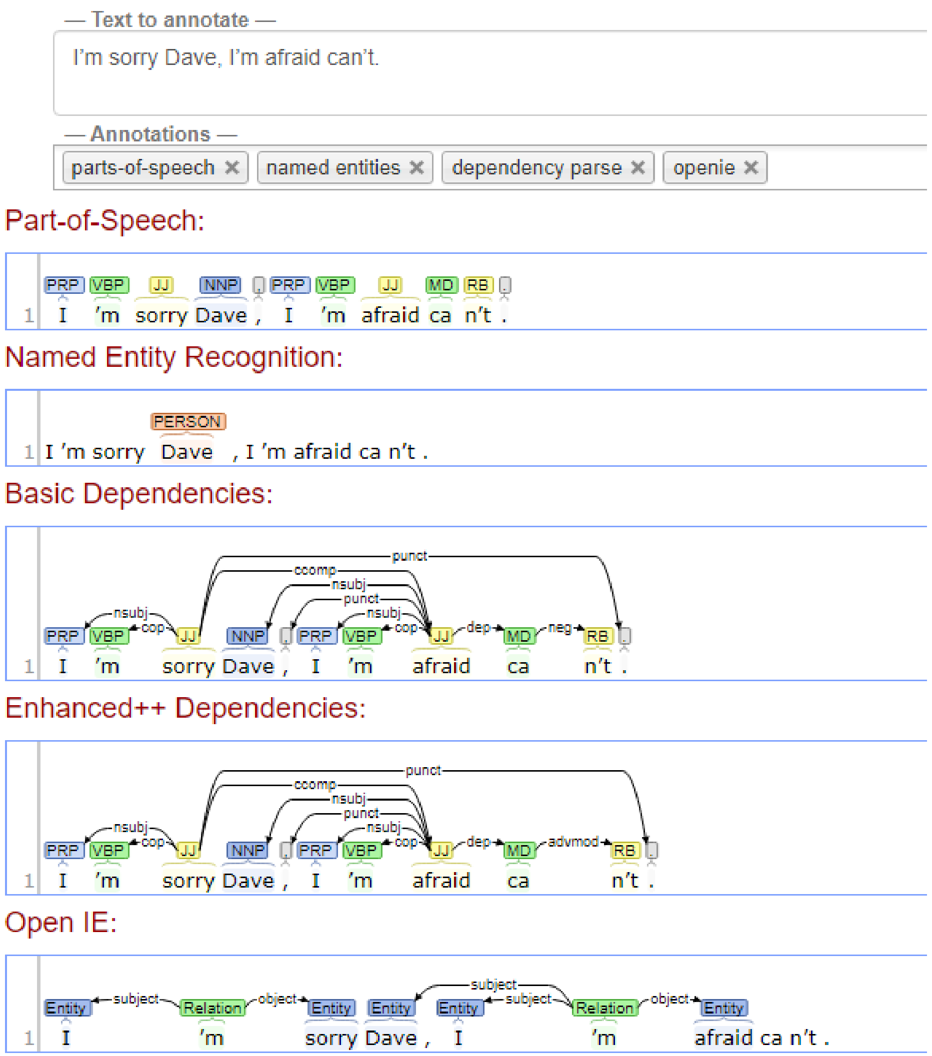
  knowledge of the structural relationships between words
  knowledge needed to stream (order) words
  moving beyond individual words

HAL:
I’m I do, sorry that afraid Dave I’m can’t.
I’m sorry Dave, I’m afraid can’t.

Engaging in Natural Language Behavior

- Syntax

Stanford's CoreNLP



Tagging

```
I/PRP  'm/VBP  sorry/JJ  Dave/NNP  ,/,  I/PRP  'm/VBP  afraid/JJ  ca/MD  n't/VB  ./.
```

Parse

```
(ROOT
  (S
    (NP (PRP I))
    (VP (VBP 'm)
      (ADJP (JJ sorry))
      (NP (NNP Dave))
      (, ,)
      (SBAR
        (S
          (NP (PRP I))
          (VP (VBP 'm)
            (ADJP (JJ afraid)
              (SBAR
                (S
                  (VP (MD ca)
                    (VP (VB n't))))))))))
    (. .)))
```

Universal dependencies

```
nsubj(sorry-3, I-1)
cop(sorry-3, 'm-2)
root(ROOT-0, sorry-3)
dep(sorry-3, Dave-4)
nsubj(afraid-8, I-6)
cop(afraid-8, 'm-7)
dep(sorry-3, afraid-8)
aux(n't-10, ca-9)
ccomp(afraid-8, n't-10)
```

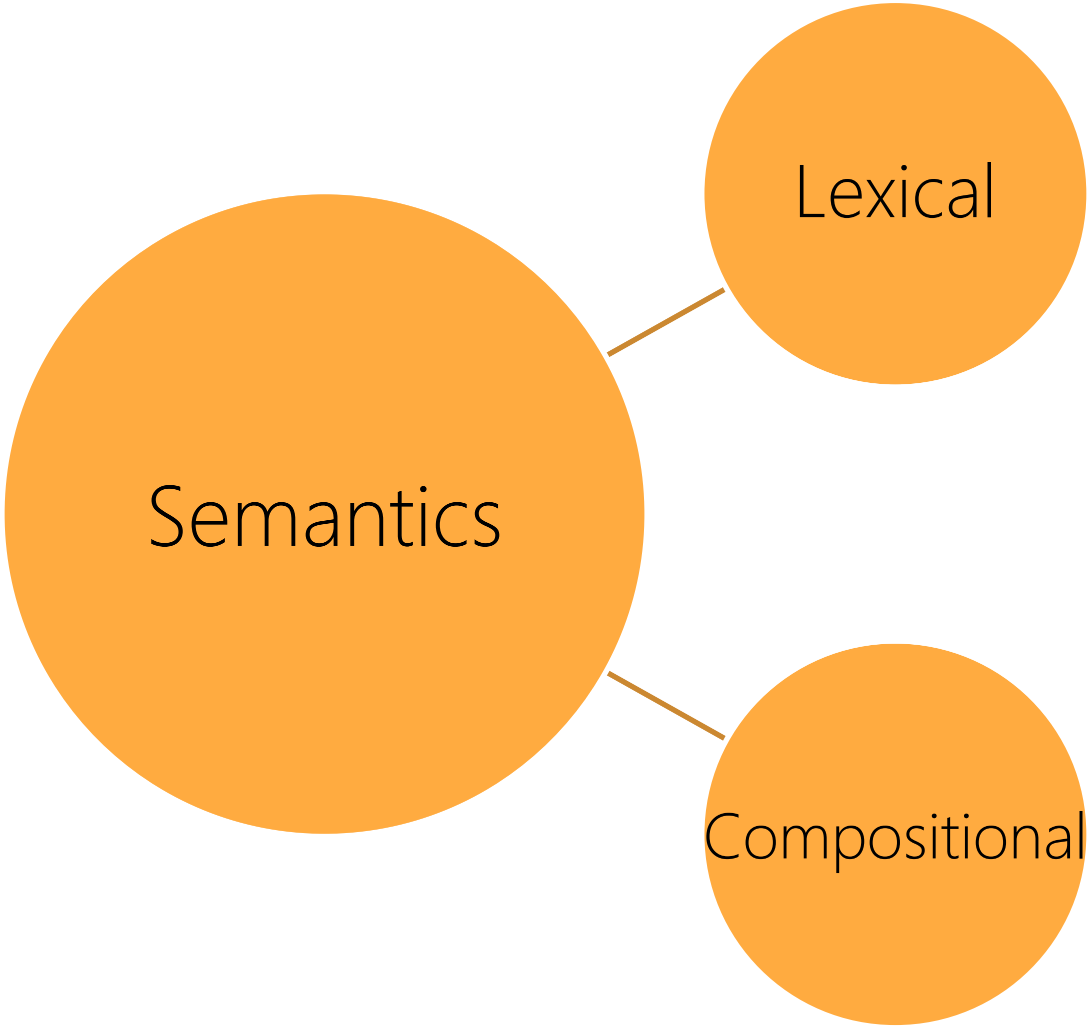
Universal dependencies, enhanced

```
nsubj(sorry-3, I-1)
cop(sorry-3, 'm-2)
root(ROOT-0, sorry-3)
dep(sorry-3, Dave-4)
nsubj(afraid-8, I-6)
cop(afraid-8, 'm-7)
dep(sorry-3, afraid-8)
aux(n't-10, ca-9)
ccomp(afraid-8, n't-10)
```

Engaging in Natural Language Behavior

- Semantics

knowledge of meaning



- The meaning of the words

Afraid → Scared
Afraid → Politely apologetic

- The meaning of the sentences

Sorry, I'm afraid I can't
I'm afraid of being sorry

Engaging in Natural Language Behavior

- Pragmatics

knowledge of the relationship of meaning to the goals & intentions of the speaker

REQUEST: HAL, open the pod bay door.
REQUEST: HAL, open the pod bay door, please!
REQUEST: HAL, open the pod bay door, please, please!

STATEMENT: HAL, the pod bay door is open.

QUESTION: HAL, is the pod bay door open?

REFUSE: Dave, I won’t.
REFUSE: Dave, I’m afraid, I can’t.

Engaging in Natural Language Behavior

- Discourse

  knowledge about linguistic units larger than a single utterance
  coreference resolution for what pronouns like it or she refers to
  another kind of pragmatic knowledge

REQUEST: HAL, open the pod *after*.
QUESTION: HAL, is *he* doing well?

Need to examine the discourse (context)

AllenNLP

- Answer a question
  - Reading Comprehension
  - Visual Question Answering
- Annotate a sentence
  - Named Entity Recognition
  - Open Information Extraction
  - Sentiment Analysis
  - Dependency Parsing
  - Constituency Parsing
  - Semantic Role Labeling
- Annotate a passage
  - Coreference Resolution
- Semantic parsing
  - WikiTables Semantic Parsing
  - Cornell NLVR Semantic Parsing

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for many higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.

End-to-end Neural Coreference Resolution (Lee et al, 2017) is a neural model which considers all possible spans in the document as potential mentions and learns distributions over possible anteceedents for each span, using aggressive pruning strategies to retain computational efficiency. It achieved state-of-the-art accuracies on on the Ontonotes 5.0 datasetin early 2017. The model here is based on that paper, but we have substituted the GloVe embeddings that it uses with SpanBERT embeddings. On Ontonotes this model achieves an F1 score of 78.87% on the test set.

Contributed by: Zhaofeng Wu

Demo Usage

Enter text or Paul Allen was born on January 21, 1953, in Seattle...

Document

Paul Allen was born on January 21, 1953, in Seattle, Washington, to Kenneth Sam Allen and Edna Faye Allen. Allen attended Lakeside School, a private school in Seattle, where he befriended Bill Gates, two years younger, with whom he shared an enthusiasm for computers. Paul and Bill used a teletype terminal at their high school, Lakeside, to develop their programming skills on several time-sharing computer systems.

Run >

0 Paul Allen was born on January 21 , 1953 , in 1 Seattle , Washington , to Kenneth Sam Allen and Edna Faye Allen . 0 Allen attended 4 Lakeside School , a private school in 1 Seattle , where 0 he befriended 2 Bill Gates , two years younger , with whom 0 he shared an enthusiasm for computers . 3 0 Paul and 2 Bill used a teletype terminal at 4 3 their high school , Lakeside , to develop 3 their programming skills on several time - sharing computer systems .

AMBIGUITY

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 a waterfowl.

I made her duck

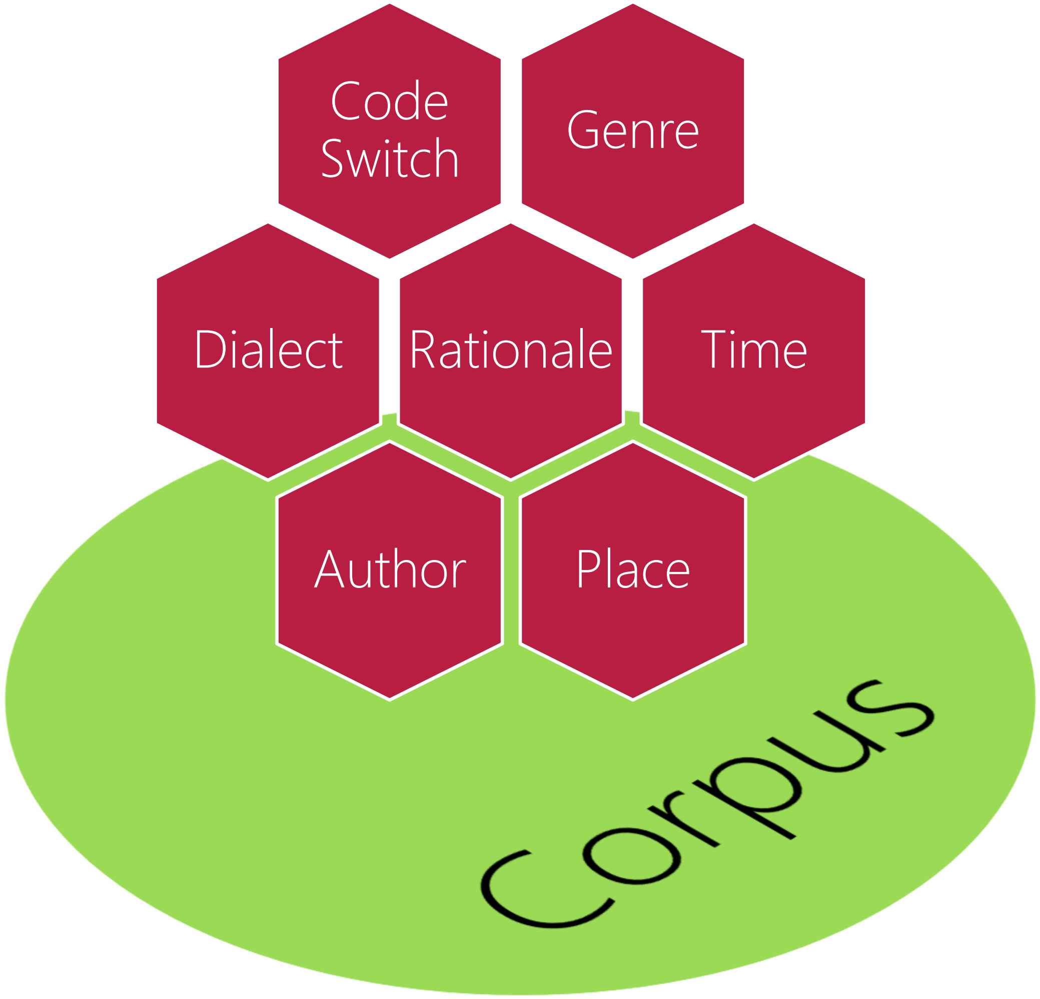
- Phonetics and Phonology
  [I][made][her duck] vs. [I][made her] [duck]
- Morphology
  [duck]: 'NOUN' vs. 'VERB'
  [her]: 'PRP' (object pronoun) vs. 'PRP$' (possessive pronoun)
- Syntax
  [her duck]: direct object
  [her][duck]: direct object, indirect object
- Semantics
  Polysemy: [made]: create vs. cook vs. cause
- Pragmatics
  [I][made][her duck] vs. [I][made her] [duck]
- Discourse
  Who is [her]?

DISAMBIGUATION

I made her duck

- Phonetics and Phonology → Speech Act Interpretation
  Speech Signal
- Morphology → Part-of-Speech Tagging
  [duck]: ‘NOUN’ vs. ‘VERB’
  [her]: ‘PRP’ (object pronoun) vs. ‘PRP$’ (possessive pronoun)
- Syntax → Syntactic Disambiguation
  [her duck]: direct object
  [her][duck]: direct object, indirect object
- Semantics → Word Sense Disambiguation
  [made]: create vs. cook vs. cause
- Pragmatics
- Discourse → Coreference
  Who is [her]?

Language is *so* situated!


Code Switch
Genre
Dialect
Rationale
Time
Author
Place
Corpus

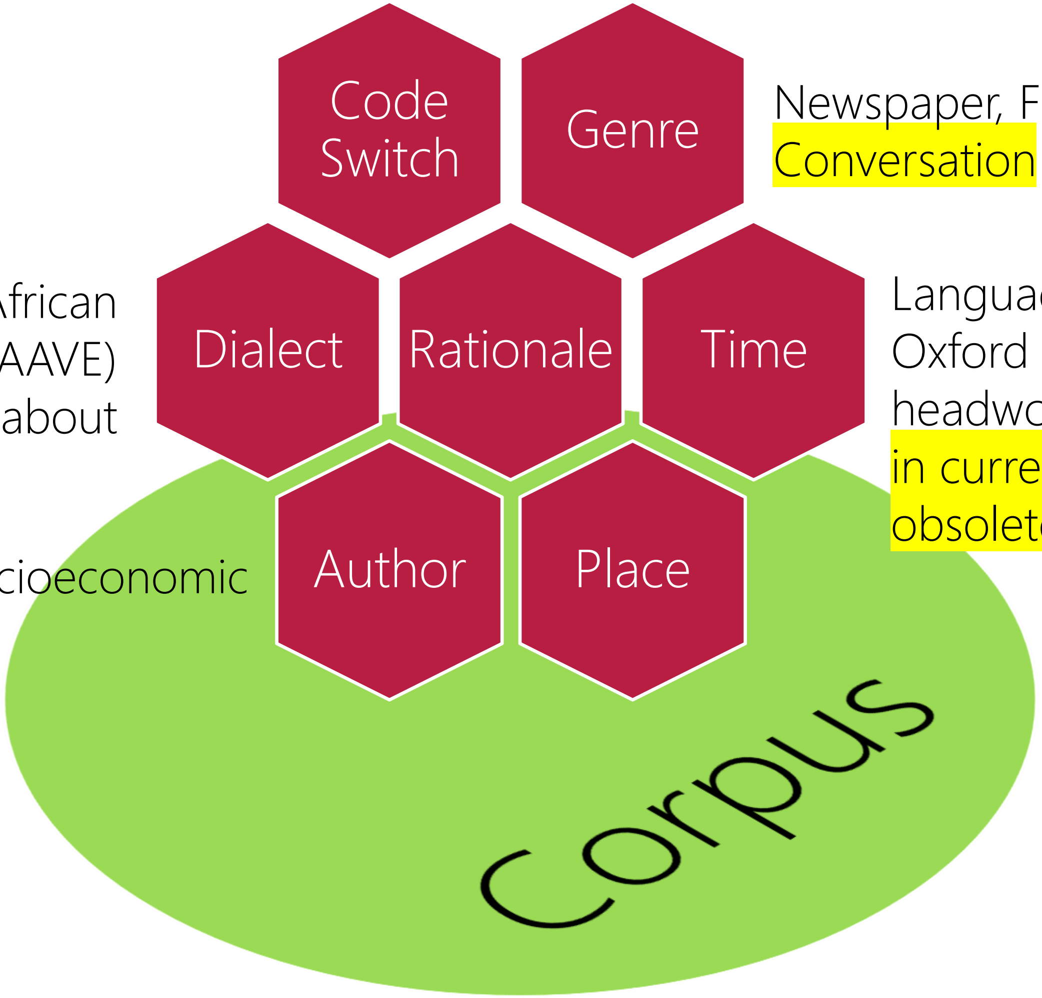
dost tha or ra- hega ... dont wory ... but dherya rakhe
[he was and will remain a friend ... don’t worry ... but have faith]

Newspaper, Fiction, non-Fiction, Academic, Social, Conversation → Disfluencies

Standard American English (SAE) vs. African American Vernacular English (AAVE)
iont: I don’t, talmebout: talking about

Language changes over time
Oxford Dictionary has 273,000 headwords; 171,476 of them being in current use, 47,156 being obsolete words

Age, Gender, Race, Socioeconomic



Corpus (*plural* Copra) Samples

Brown University

English

newspaper, fiction, non-fiction, academic, etc.

1963–64

#Documents = Size = 500

#Tokens = 1 M

#Vocab = Unique Tokens = Types = 38 K

Switchboard

American English

Telephone Conversations between strangers

Early 1990s

#Conversations = Size = 2430

#Tokens = 2.4 M

#Vocab = Unique Tokens = Types = 20 K

Google N-grams

English

Google Books

#Tokens = 1 G

#Vocab = Unique Tokens = Types = 13 M

Herdan’s Law or Heaps’ Law

Herdan, G. (1960). Type-token mathematics. The Hague, Mouton.
Heaps, H. S. (1978). Information retrieval. Computational and theoretical aspects. Academic Press.

$$|V| = kN^{\beta};\ 0 < \beta < 1$$

k and β are positive constants. The value of β depends on the corpus size and the genre, but at least for the large corpora β ranges from 0.67 to 0.75.

Datasheets (Data Statements) for Datasets

Emily M. Bender, Batya Friedman, ACL (2018)
Gebru, Timnit, et al. (2018)

direct stakeholders. For example, Speer (2017) found that a sentiment analysis system rated reviews of Mexican restaurants as more negative than other types of food with similar star ratings, because of associations between the word *Mexican* and words with negative sentiment in the larger corpus on which the word embeddings were trained. (See also Kiritchenko and Mohammad,

NLP is AI-hard (AI-Complete)
Let’s do it!

TEXT SEGMENTATION

dividing written text into meaningful units, such as words, sentences, or topics

Word Segmentation: Tokenization

- Whitespace (default, natural word delimiter)
- Exceptions

    New York
    rock 'n' roll
    Contractions: I'm
    Japanese | Chinese | Thai don't have spaces between words
    Emoticons: :)
    Hashtags: #nlproc.

Word Boundaries: Tokenization: Space

- Split()
- Regular Expressions (RE): Finite State Automata
  - Alphabetical: [a-zA-Z]*
  - Alpha-numerical: [a-zA-Z0-9]*
  - Punctuations: Ph.D., AT&T, cap'n
  - Special Chars

    Currency $45.55

    Dates (01/02/06)

    URLs http://www.stanford.edu

    Twitter hashtags #nlproc

    Email hfani@uwindsor.ca

What should be considered as word?

- Disfluencies in *utterances*
  - Fragments: broken-off repeated words: miss- misspelled, you- yourself
  - Fillers: non-lexical: huh, uh, erm, um, well, so, like, hmm
- Punctuations , . : ; ? !
  - part-of-speech tagging
  - parsing
  - speech synthesis
- Morphemes:
  - smallest meaning-bearing unit of a language
  - ‘unlikeliest’ : morphemes [un-], [likely], [-est]

What should be considered as word?

- Chinese

As Chen et al. (2017) point out, this could be treated as 3 words ('Chinese Treebank' segmentation):

(2.5) 姚明 进入 总决赛
YaoMing reaches finals

or as 5 words ('Peking University' segmentation):

(2.6) 姚 明 进入 总 决赛
Yao Ming reaches overall finals

Finally, it is possible in Chinese simply to ignore words altogether and use characters as the basic elements, treating the sentence as a series of 7 characters:

(2.7) 姚 明 进 入 总 决 赛
Yao Ming enter enter overall decision game

What should be considered as word?

- Chinese

characters are at a reasonable semantic level for most applications
most word standards result in a huge vocabulary with large numbers of very rare words
Take characters as words

As Chen et al. (2017) point out, this could be treated as 3 words ('Chinese Treebank' segmentation):

(2.5) 姚明 进入 总决赛
YaoMing reaches finals

or as 5 words ('Peking University' segmentation):

(2.6) 姚 明 进入 总 决赛
Yao Ming reaches overall finals

Finally, it is possible in Chinese simply to ignore words altogether and use characters as the basic elements, treating the sentence as a series of 7 characters:

(2.7) 姚 明 进 入 总 决 赛
Yao Ming enter enter overall decision game