DSC291: Advanced Statistical Natural Language Processing

Overview

Zhiting Hu

Lecture 1, March 29, 2022

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Class webpage: http://zhiting.ucsd.edu/teaching/dsc291spring2022

DSC291-Spring2022

Logistics Lectures Homework Project

Advanced Statistical Natural Language Processing

DSC 291 • Spring 2022 • UC San Diego

Logistics

- Lectures
  - **Time:** Tuesday/Thursday 3:30pm-4:50pm
  - **Location:** HSS 1315

- No discussion session as a DSC 291 class
- Instead: Office hours, Piazza, ad-hoc meetings if needed

Logistics

Instructor: Zhiting Hu
Email: zhh019@ucsd.edu
Office hours: Thursday 2:30-3:30pm
Location: SDSC E247

TA: Pushkar Bhuse
Email: pbhuse@ucsd.edu
Office hours: TBA
Location: TBA

- Canvas
- Discussion forum: Piazza
- Homework & writeup submission: Gradescope

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
  - Theory exercises, implementation exercises
  - 3 total late days without penalty
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
  - Each student will give an oral presentation on a research paper
  - 10 mins = 8 mins presentation + 2 mins QA
  - Discuss both strengths and limitations of the paper
  - Sign up in a google sheet (TBA)
  - Starting TBA
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
    - 3 or 4-member team to be formed and sign up in a google sheet (TBA)
    - Designed to be as similar as possible to researching and writing a conference-style paper:
        - Due to tight timeline, fine to use synthetic/toy data for proof-of-concept experiments + explanation of theory/intuition of why your approach is likely to work
    - **Proposal** : 2 pages excluding references (10%) -- Due 04/14
        - Overview of project idea, literature review, potential datasets and evaluation, milestones
    - **Midway Report** : 4-5 pages excluding references (20%)
    - **Presentation** : oral presentation, 15-20mins (20%)
    - **Final Report** : 6-8 pages excluding references (50%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)
  - Contribution to discussion on Piazza
  - Complete mid-quarter evaluation
  - Any constructive suggestions

Advanced Statistical Natural Language Processing

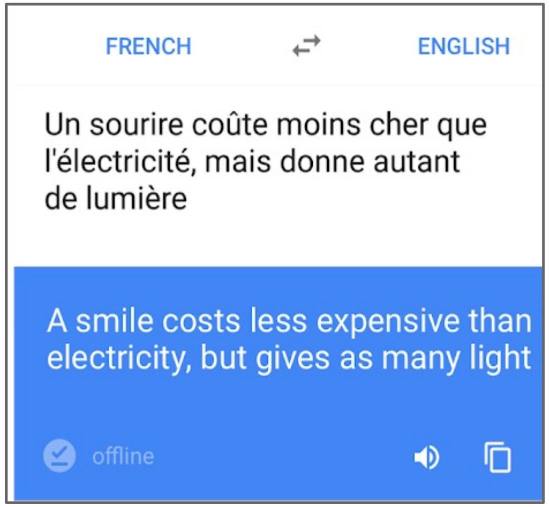
Advanced Statistical Natural Language Processing

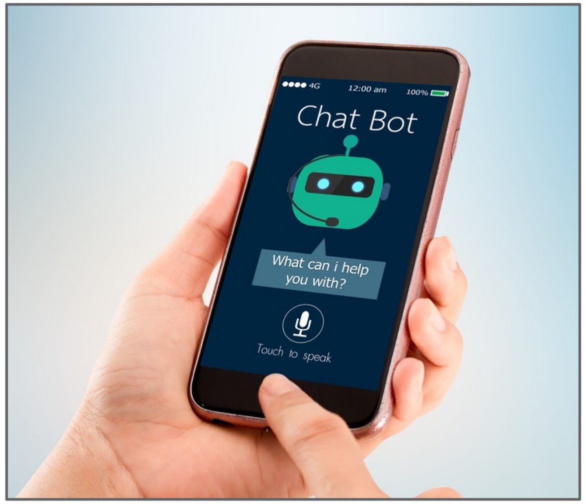
What is NLP?

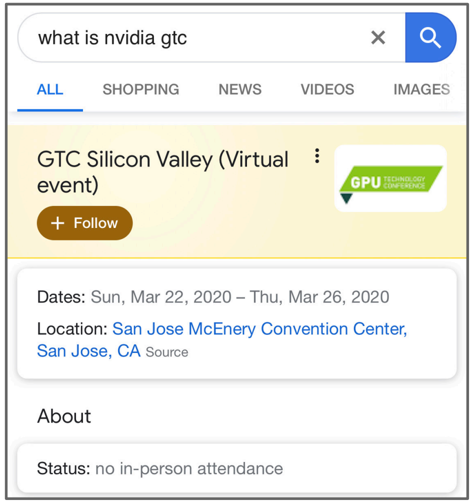
Statistical machine learning (ML) methods

We'll cover only a subset of advanced, latest methods

What is NLP


FRENCH
ENGLISH
Un sourire coûte moins cher que l'électricité, mais donne autant de lumière
A smile costs less expensive than electricity, but gives as many light
offline


Chat Bot
What can i help you with?
Touch to speak


what is nvidia gtc
ALL
SHOPPING
NEWS
VIDEOS
IMAGES
GTC Silicon Valley (Virtual event)
Follow
Dates: Sun, Mar 22, 2020 – Thu, Mar 26, 2020
Location: San Jose McEnery Convention Center, San Jose, CA Source
About
Status: no in-person attendance


God @TheTweetOfGod · 6h
Sentiment →
"The alien who resides with you shall be to you as the citizen among yo
love the alien as yourself, for you were aliens in the land of Egypt: I am
Danielle Panabaker @dpanabaker · 18h
Sentiment →
Happy New Year!
289
588
8.4K

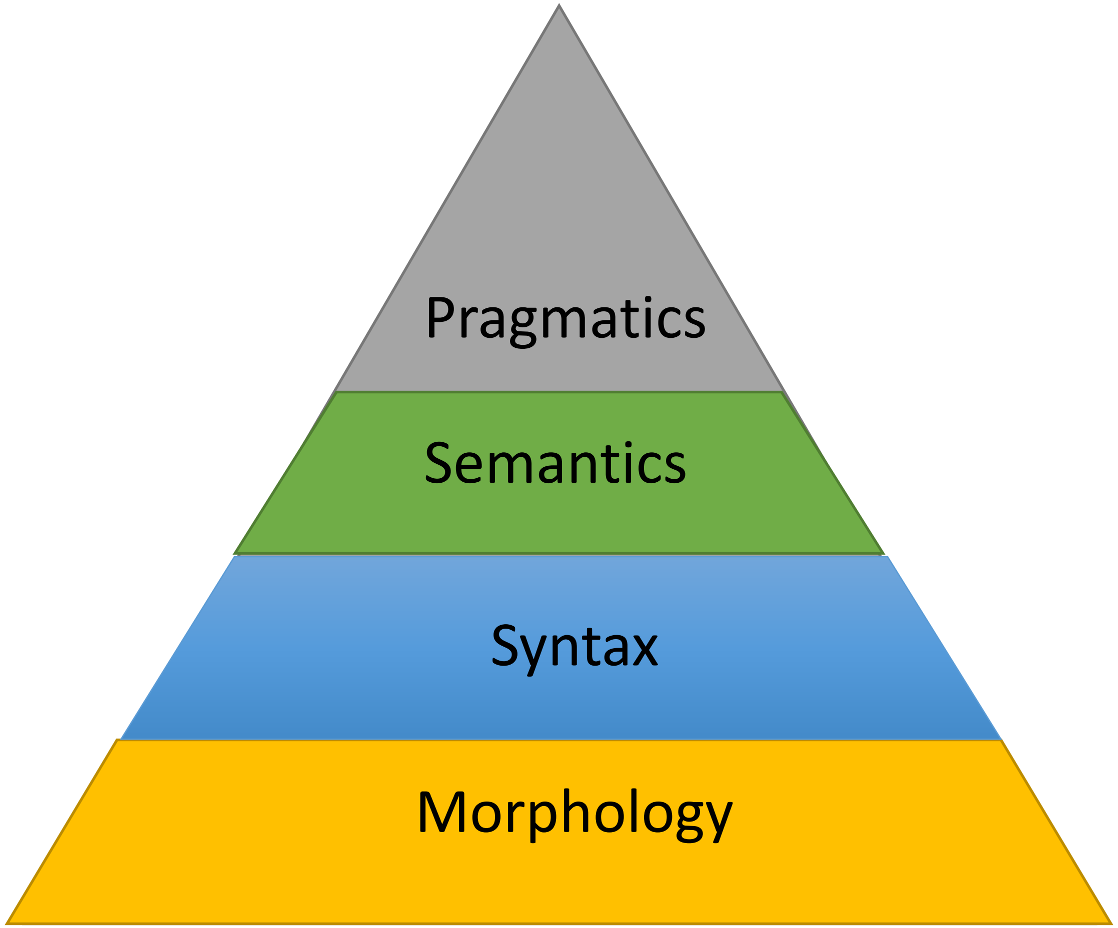
What is NLP

- NL ∈ { English, German, Chinese, Spanish, Hindi, American Sign Language, . . ., Lushootseed }

- Automation of:
  - analysis or “understanding” (to some degree) what a text means
  - generation of fluent, meaningful, context-appropriate text
  - acquisition of these capabilities from knowledge and data

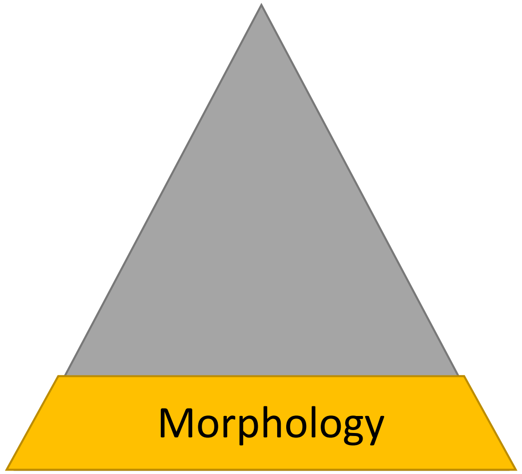
[Courtesy: UW CSE 447 by Noah Smith]

Language Understanding Pyramid

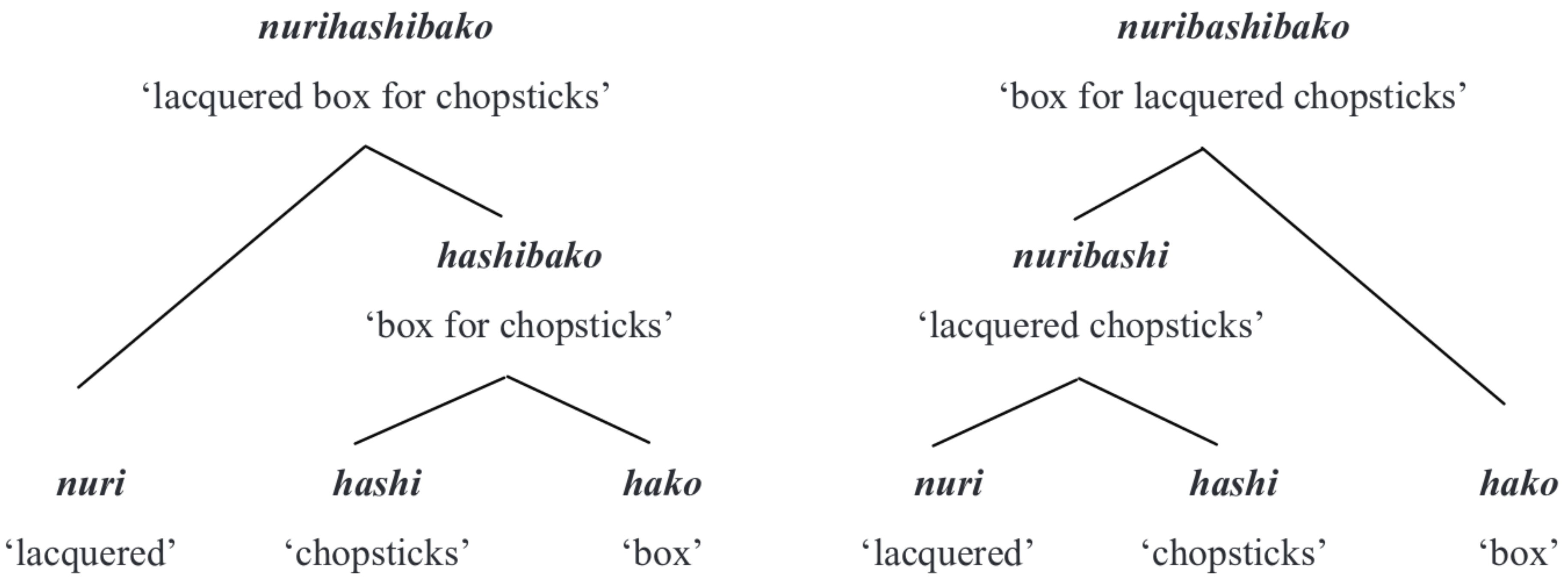


[Courtesy: Hector Liu, Forte talk]

Morphology

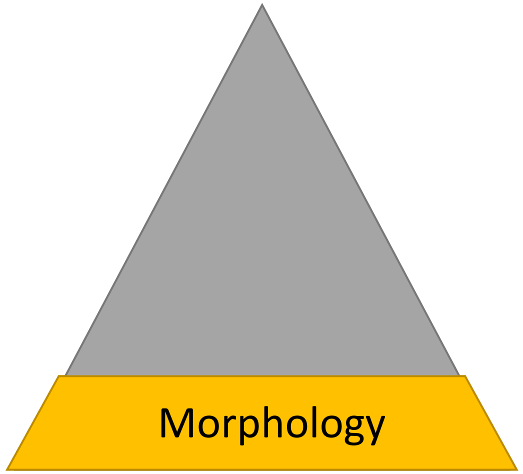


Morphology Analysis



[Courtesy: Hector Liu, Forte talk]

Morphology



Stemming

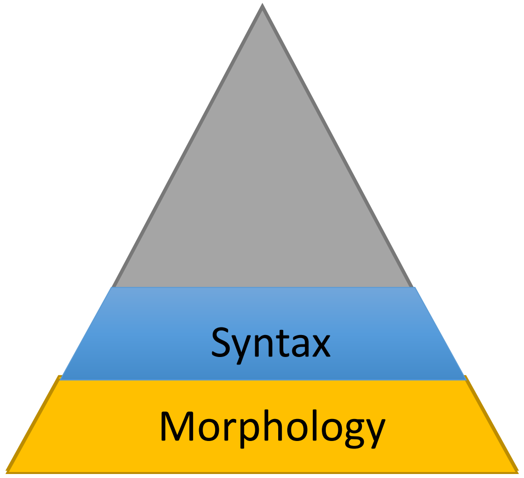
adjustable → adjust_
formality → formaliti
formaliti → formal
airliner → airlin_

Lemmatization

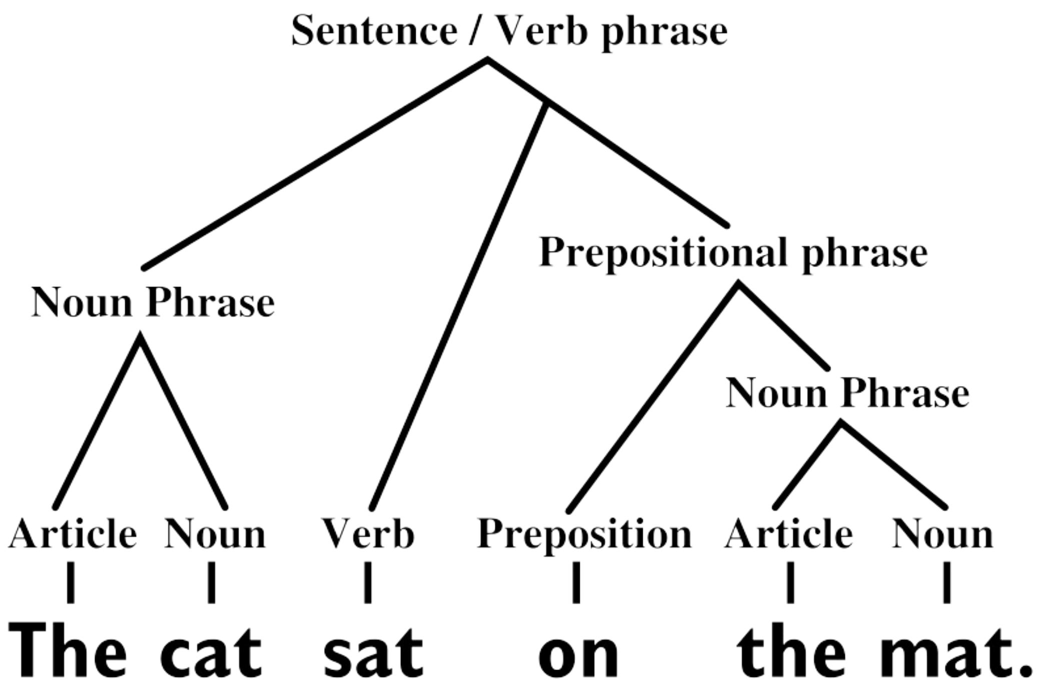
was → (to) be
better → good
meeting → meeting

Stemming and lemmatization: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Syntax

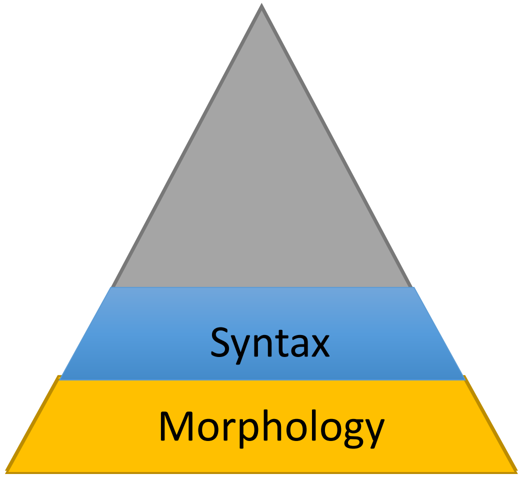


Constituent Parsing

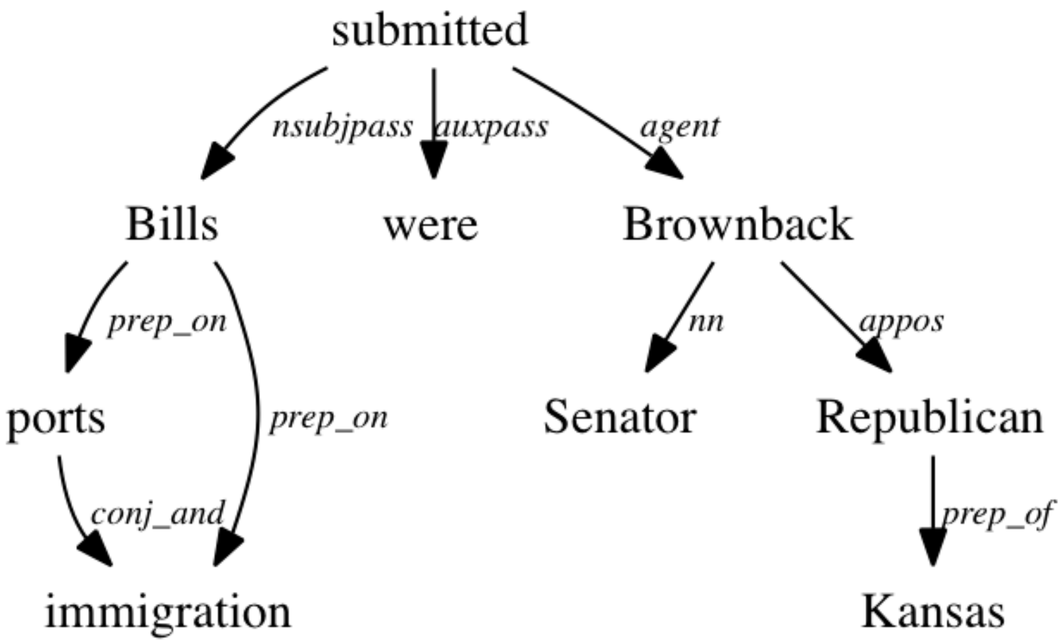


[Courtesy: Hector Liu, Forte talk]

Syntax

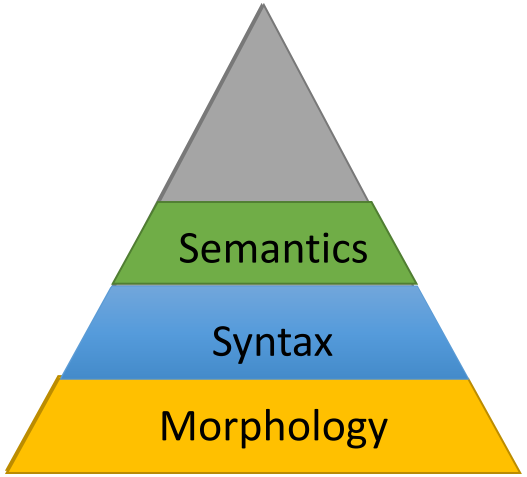


Dependency Parsing

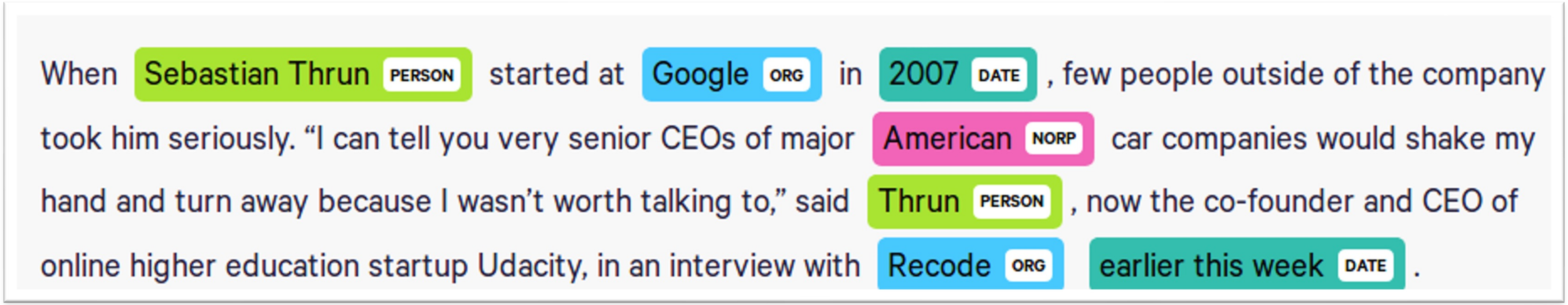


[Courtesy: Hector Liu, Forte talk]

Semantics

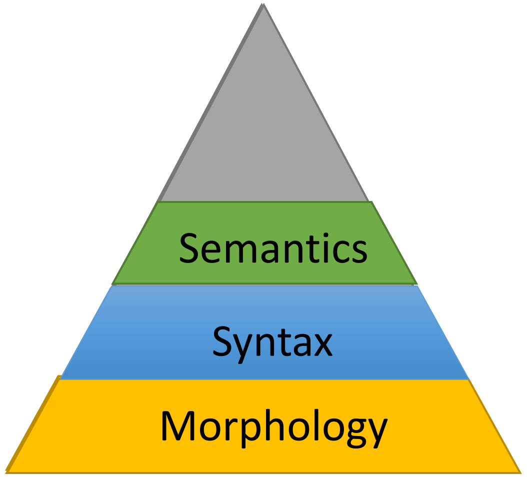


Named Entity Recognition

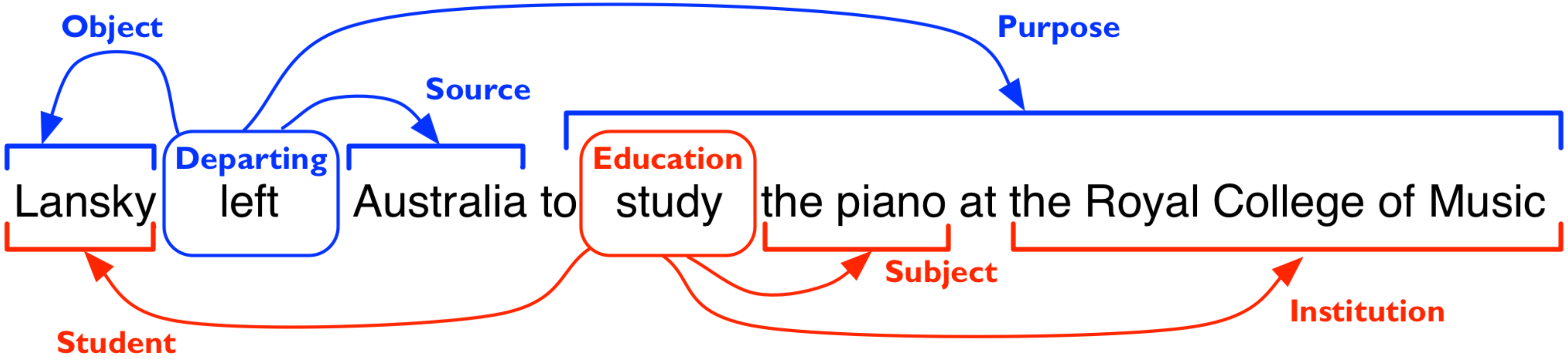


[Courtesy: Hector Liu, Forte talk]

Semantics

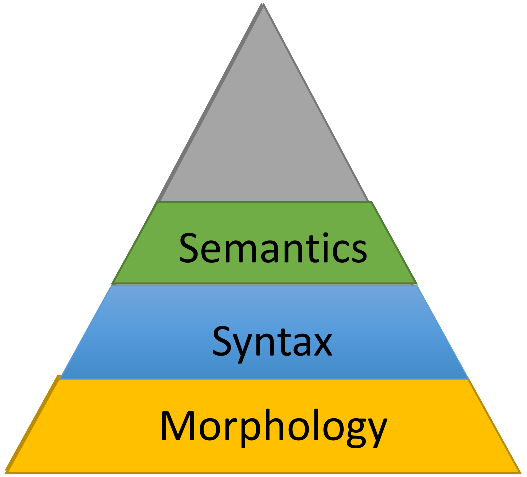


Semantic Roles

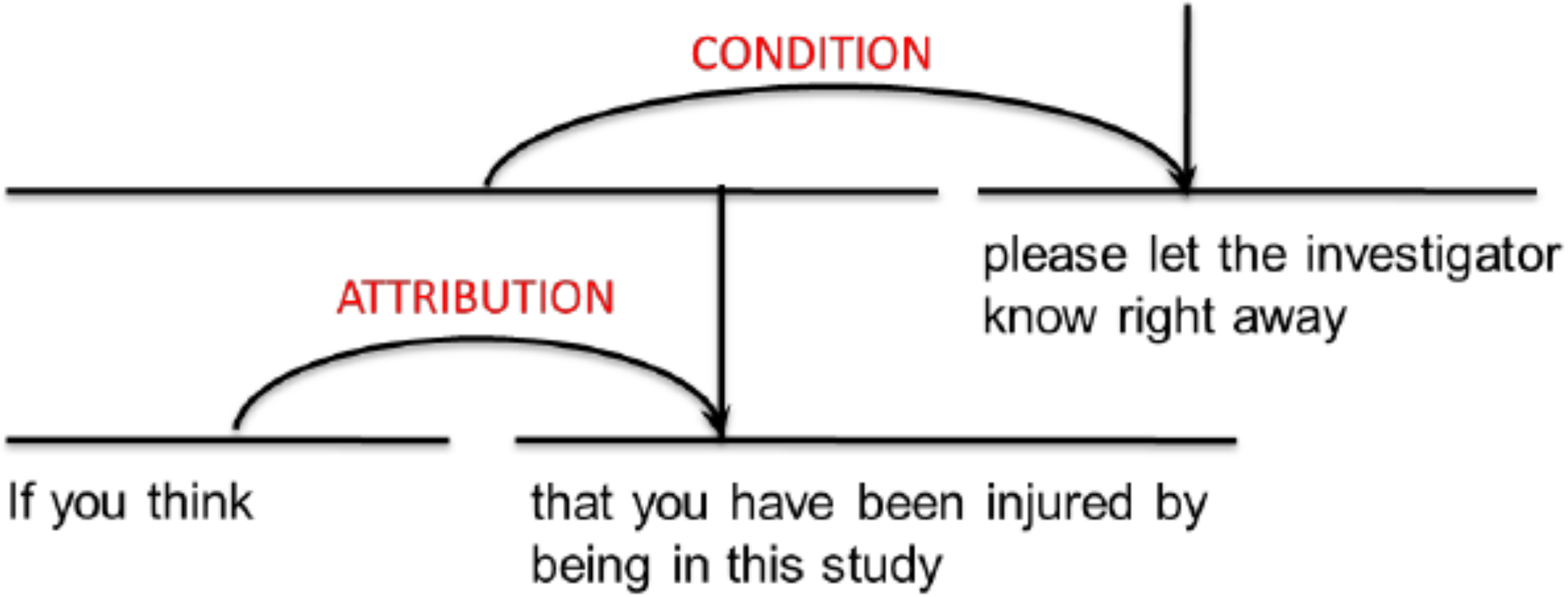


[Courtesy: Hector Liu, Forte talk]

Semantics

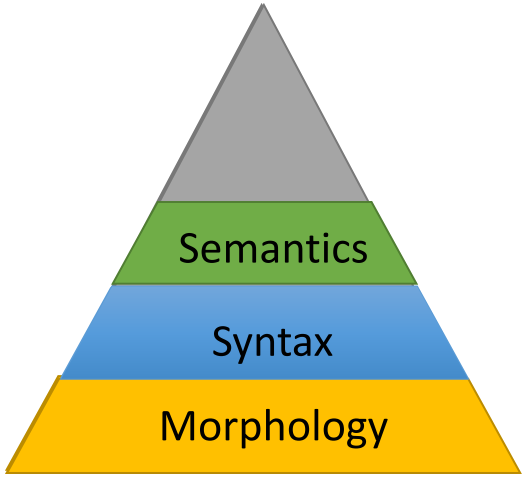


Discourse Parsing



[Courtesy: Hector Liu, Forte talk]

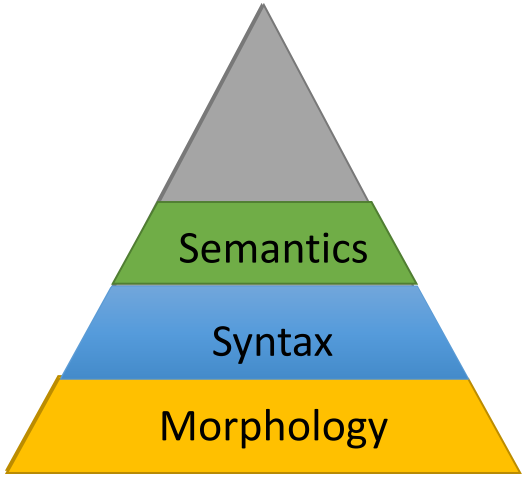
Semantics



Coreference

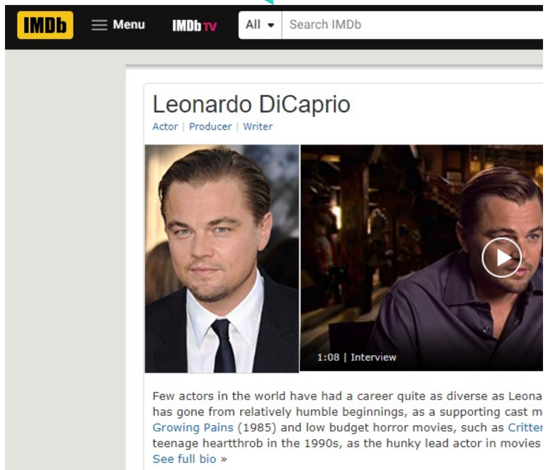
"I voted for Nader because he was most aligned with my values," she said.

Semantics



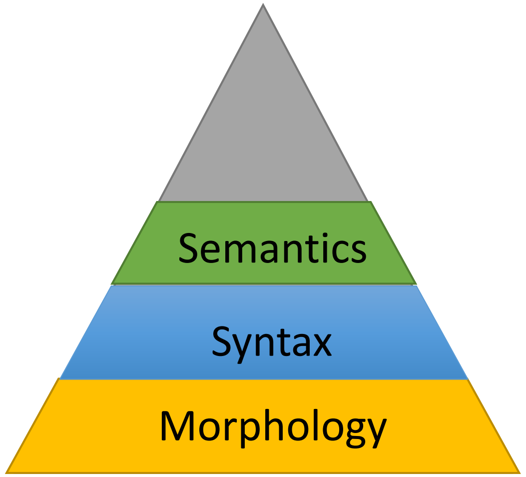
Entity Linking

Kate Winslet and Leonardo Dicaprio have definitely created a timeless classic.



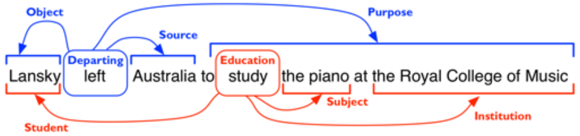
[Courtesy: Hector Liu, Forte talk]

Language Understanding Pyramid

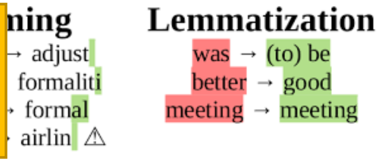


Morphology

Semantic Parsing

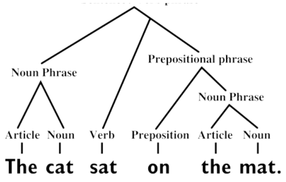


Lemmatize

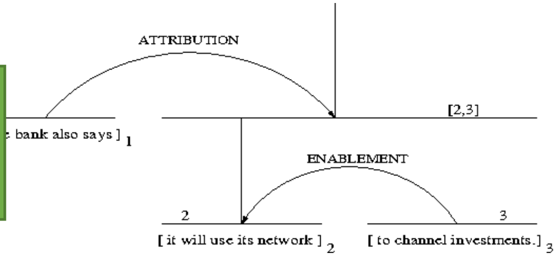


NER

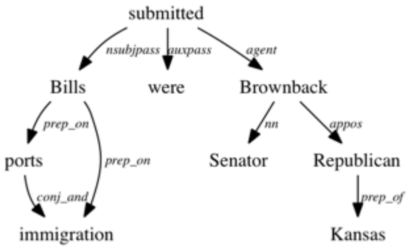
Constituent Parse



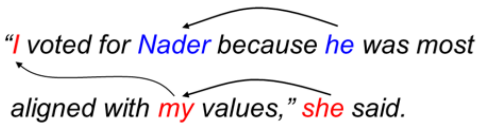
Discourse Parsing



Dependency Parse



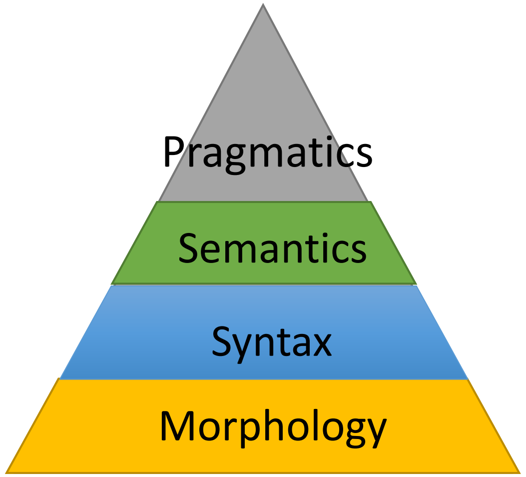
Coreference



Entity Linking

[Courtesy: Hector Liu, Forte talk]

Pragmatics



[Courtesy: Hector Liu, Forte talk]

Why NLP is Hard

- Ambiguity
  - A string may have many possible interpretations in different contexts, and resolving ambiguity correctly may rely on knowing a lot about the world.

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- Who or what is wrapped in paper?
- An event of perception, or an assault?

Why NLP is Hard

- Ambiguity
  - A string may have many possible interpretations in different contexts, and resolving ambiguity correctly may rely on knowing a lot about the world.
  - Richness: any meaning may be expressed many ways, and there are immeasurably many meanings.
  - Linguistic diversity across languages, dialects, genres, styles, …
- Appropriateness of a representation depends on the application
- Typically, representation of language is a theorized construct, not directly observable, or it is encoded numerically (vectors, matrices, tensors) and inscrutable
- There are many sources of variation and noise in linguistic input

[Courtesy: UW CSE 447 by Noah Smith]

Advanced Statistical Natural Language Processing

What is NLP?

Statistical machine learning (ML) methods

We'll cover only a subset of advanced, latest methods

Machine Learning

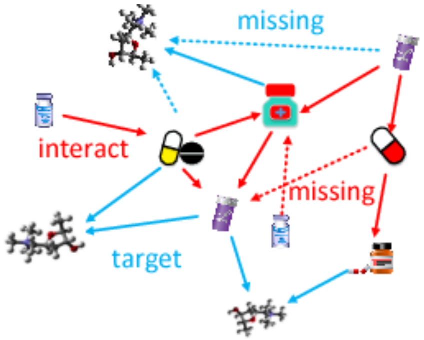
- Computational methods that enable machines to learn concepts and improve performance from **experiences**.

Experiences of all kinds

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

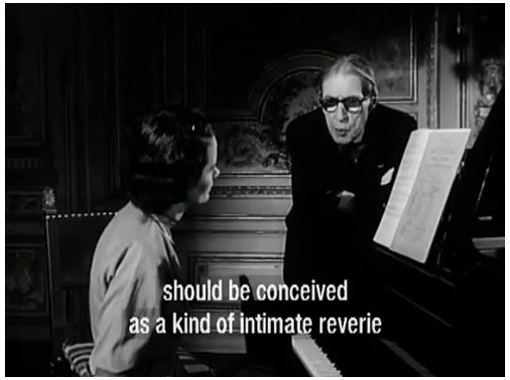


Knowledge graphs

Rewards

Auxiliary agents

Adversaries

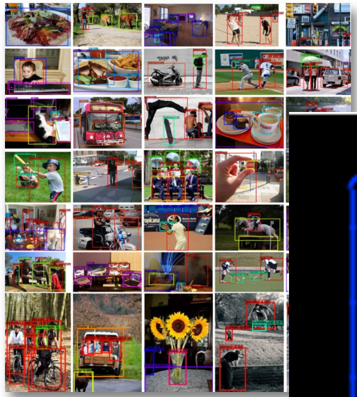


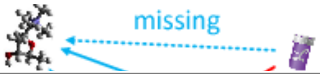
Teachers

…

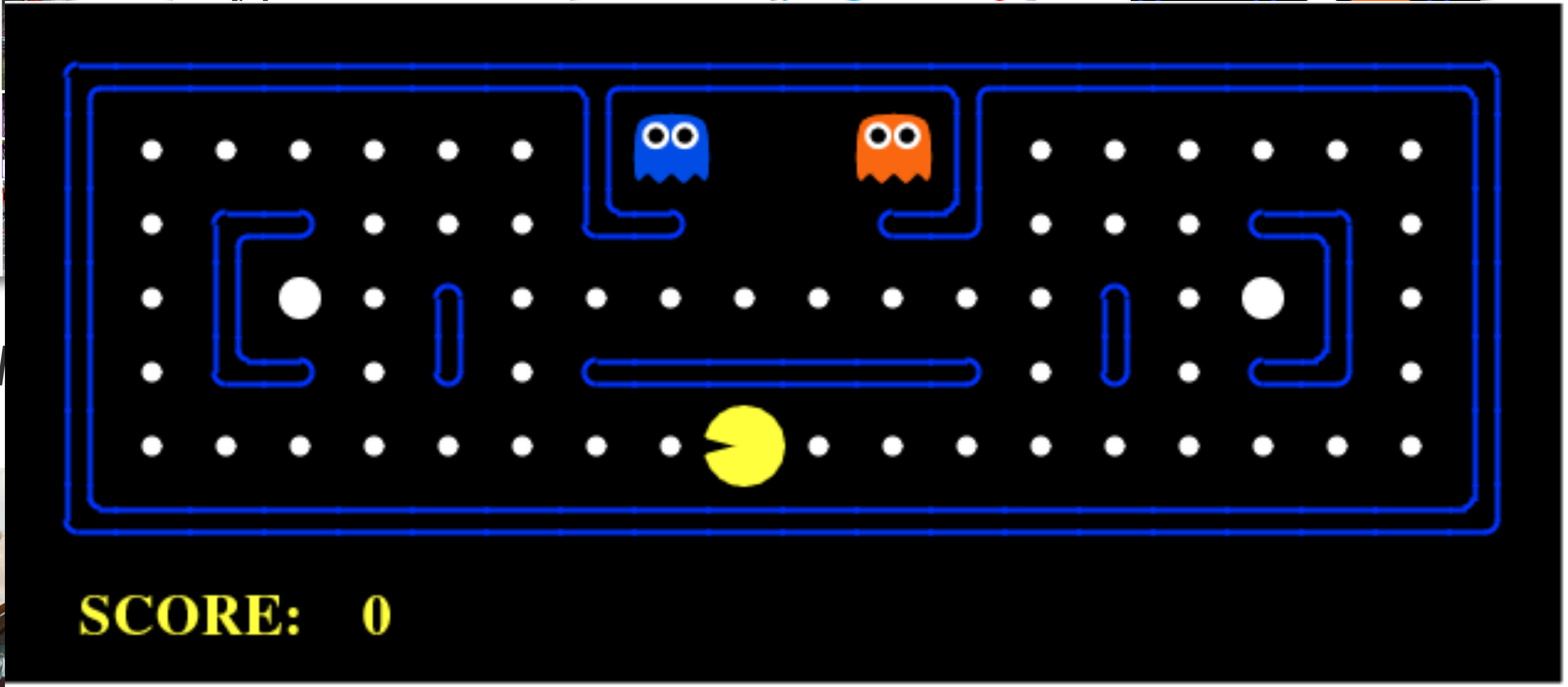
And all combinations thereof

Experiences of all kinds





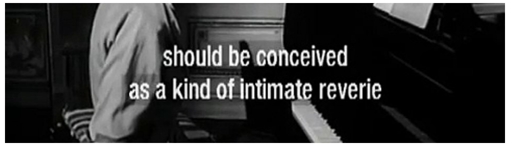
Data exampl



Auxiliary agents

ations thereof

Adversaries



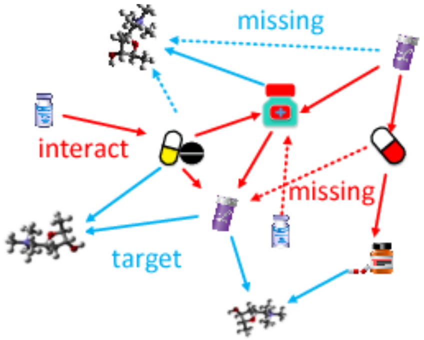
Master classes

Experiences of all kinds

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

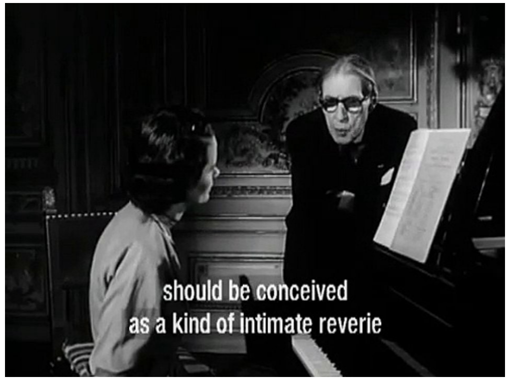


Knowledge graphs

Rewards

Auxiliary agents

Adversaries



Master classes

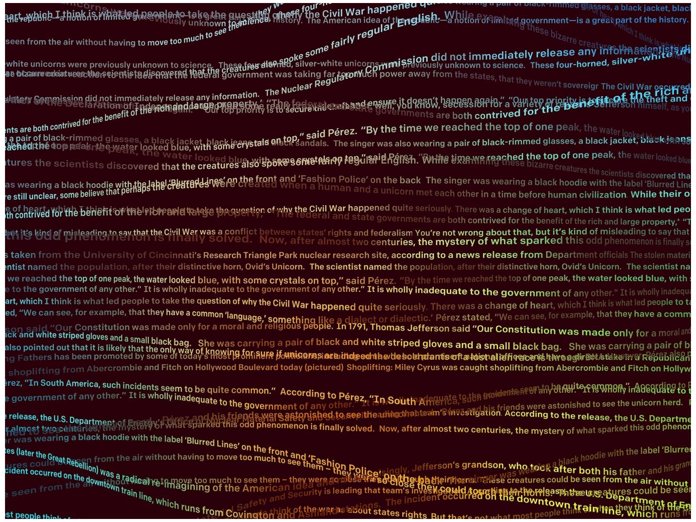
…

And all combinations thereof

Experiences: (massive) data examples

Image classification

Machine translation



Language modeling
(BERT, GPT-2, **GPT-3**, …)

45TB of text data: CommonCrawl, WebText, Wikipedia, corpus of books, …

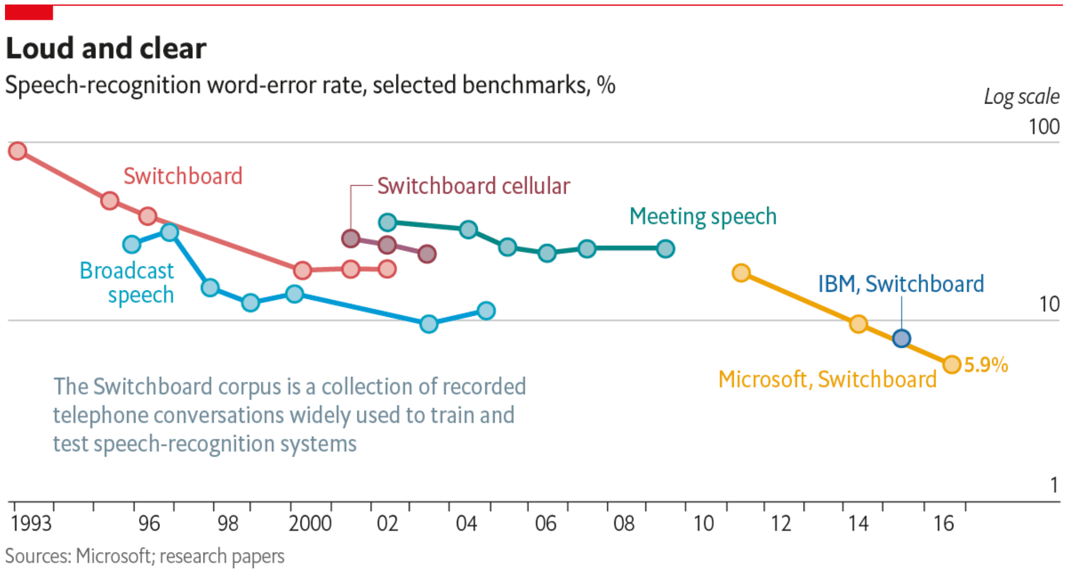
Experiences: (massive) data examples

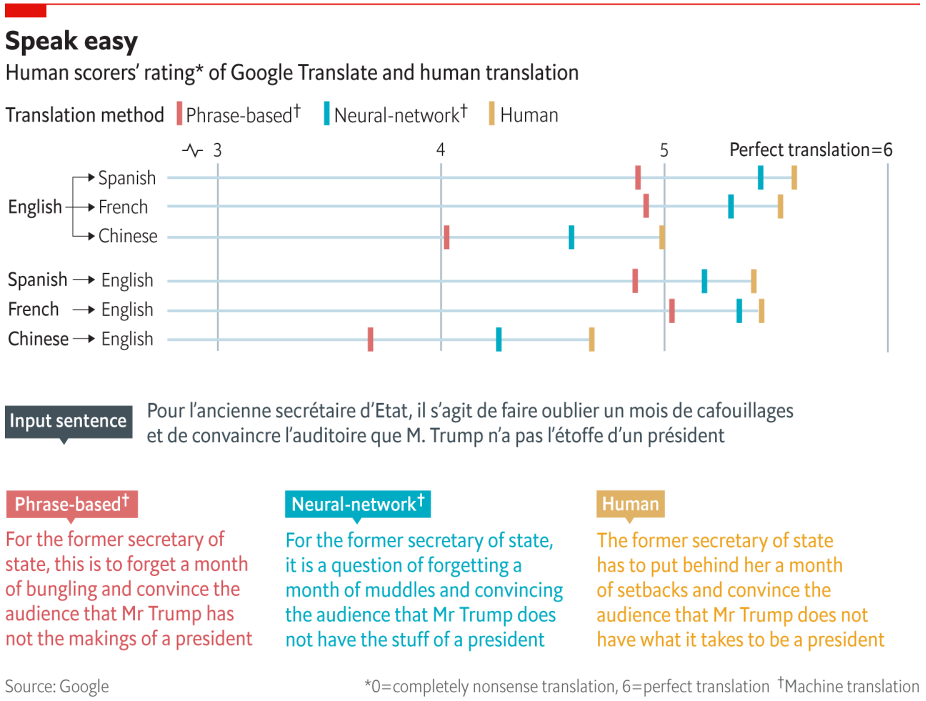
TECH ARTIFICIAL INTELLIGENCE

OpenAI's text-generating system GPT-3 is now spewing out 4.5 billion words a day

Robot-generated writing looks set to be the next big thing

By James Vincent | Mar 29, 2021, 8:24am EDT





[The Economist]

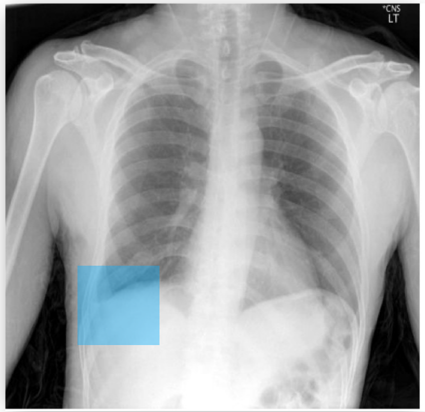
Problems with few data (labels)

- Privacy, security issues

Assistive diagnosis

Normal findings

Abnormal findings



``The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring …''

Problems with few data (labels)

- Expensive to collect/annotate

  Controllable content generation

Controlling sentiment

Pos	*The film is full of imagination!*
Neg	*The film is strictly routine!*

Controlling writing style

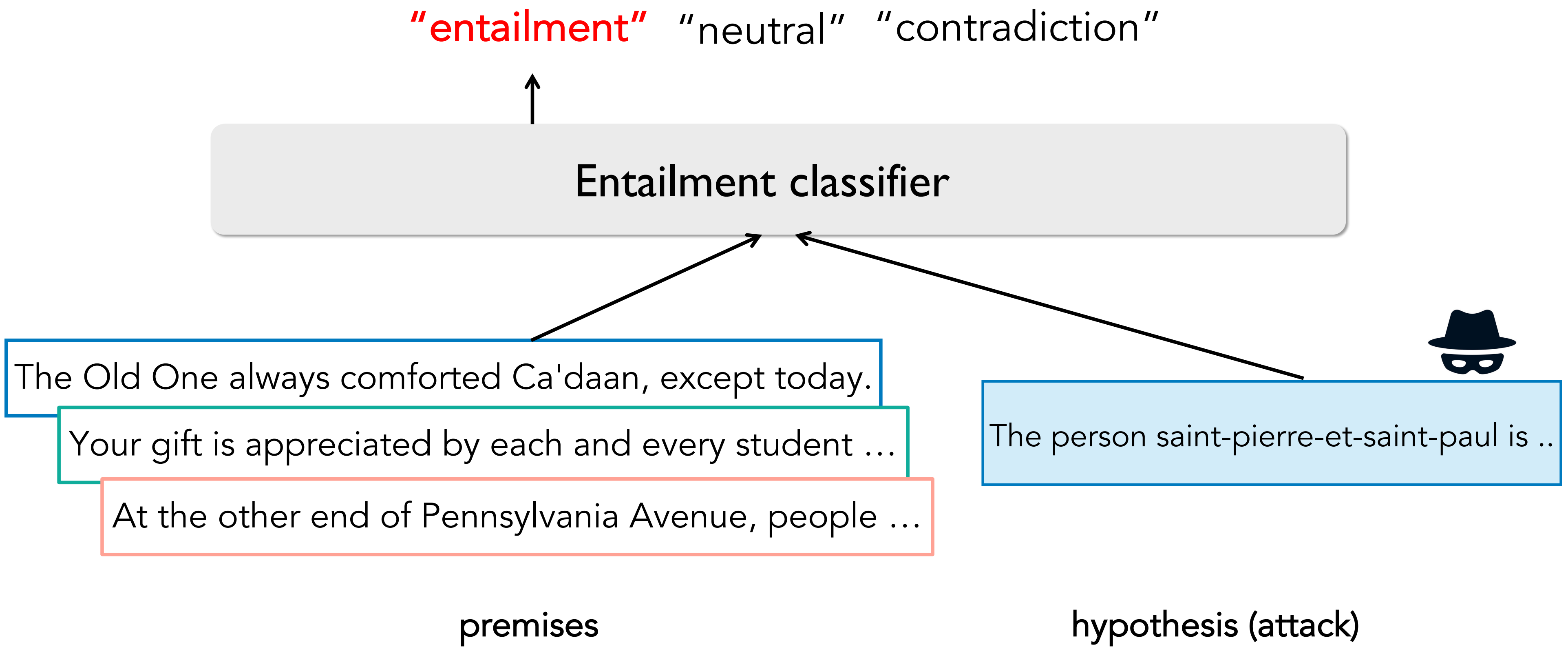
Plain	*LeBron James contributed 26 points, 8 rebounds, 7 assists.*
Elaborate	*LeBron James rounded out the box score with an all around impressive performance, scoring 26 points, grabbing 8 rebounds and dishing out 7 assists.*

Applications: personalized chatbot, live sports commentary production

Problems with few data (labels)

- Difficult / expertise-demanding to annotate

Adversarial attack

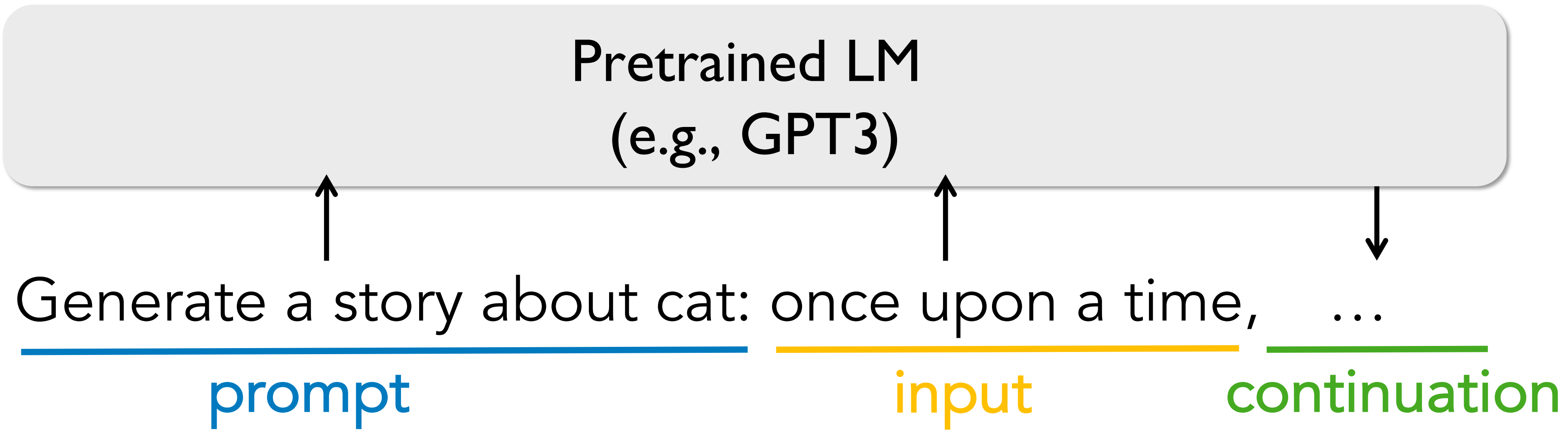


Applications: test model robustness

Problems with few data (labels)

- Difficult / expertise-demanding to annotate

Prompt generation: automatically generating prompts to steer pretrained LMs



Problems with few data (labels)

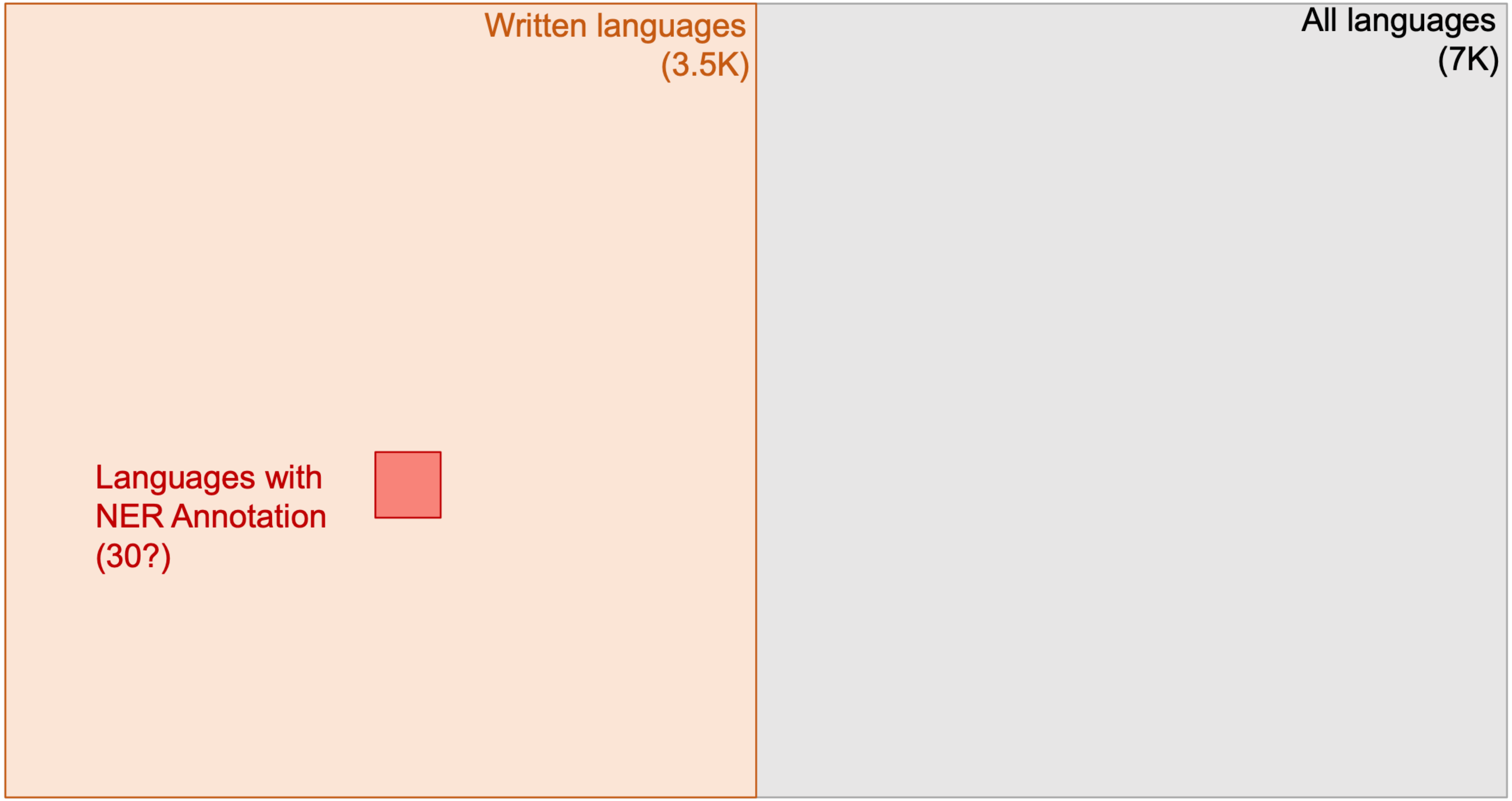
- Specific domain

Low-resource languages

~7K languages in the world

Problems with few data (labels)

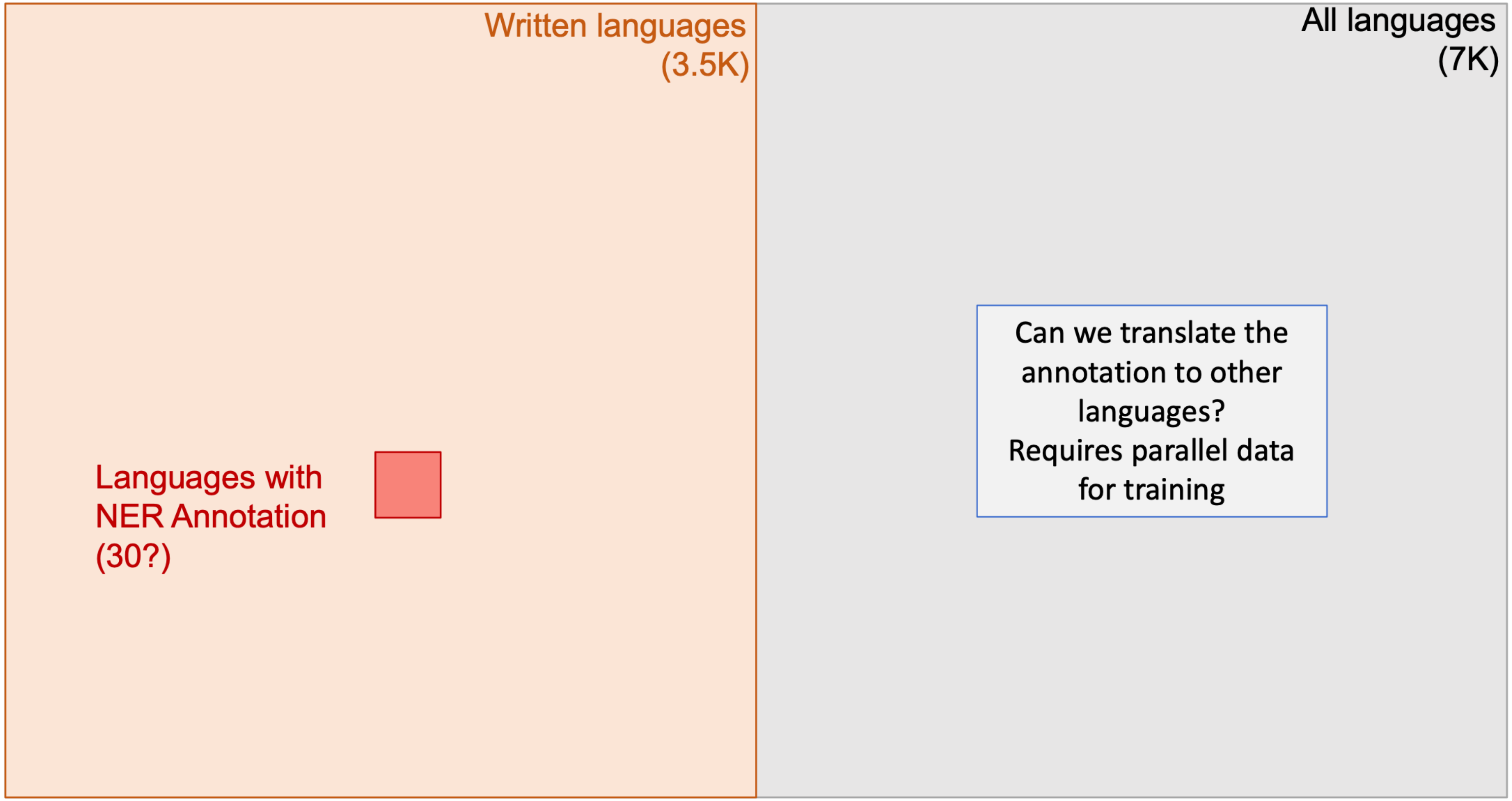
- Specific domain   Low-resource languages



[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

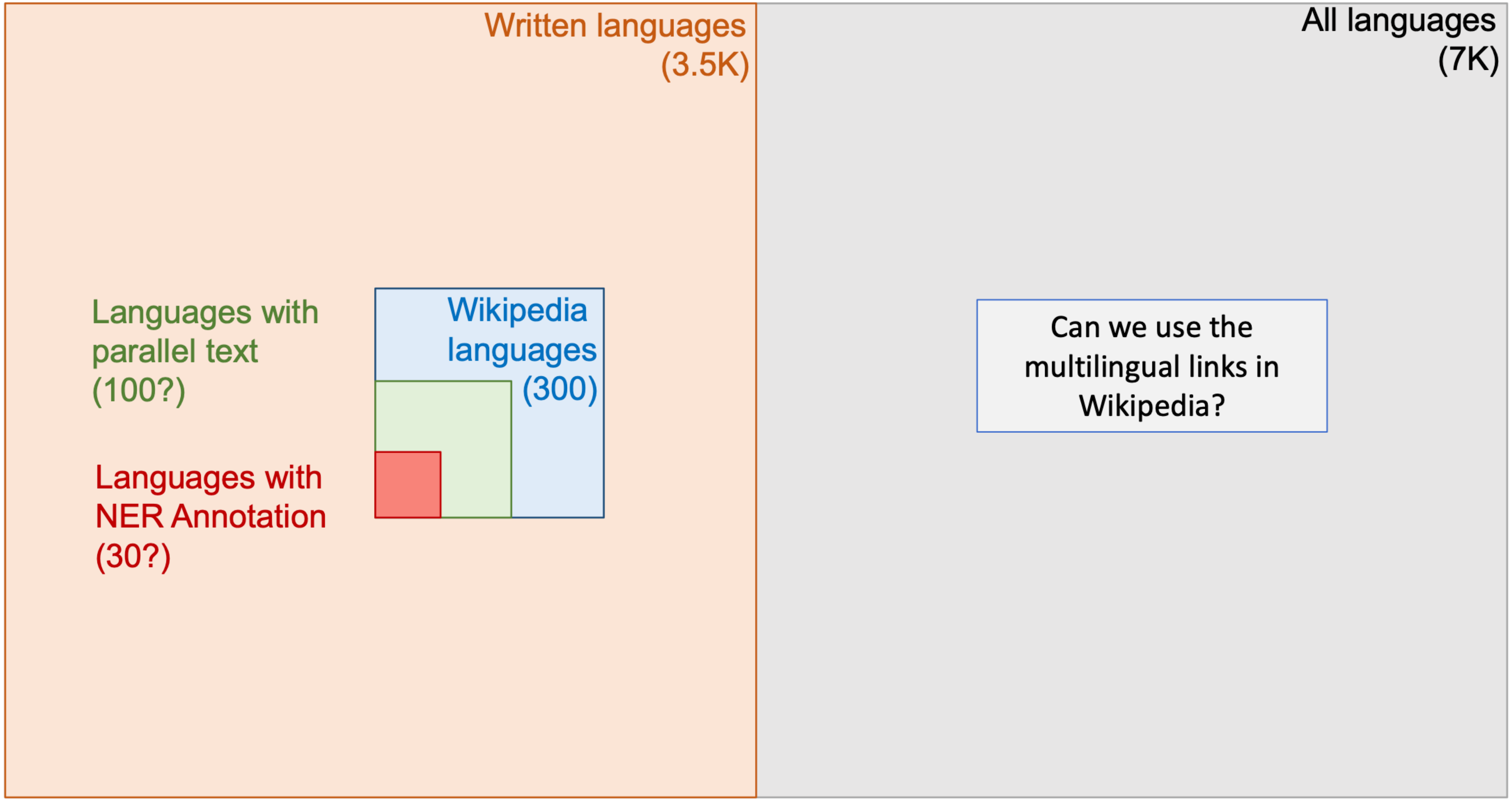
- Specific domain   Low-resource languages



[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

- Specific domain   Low-resource languages

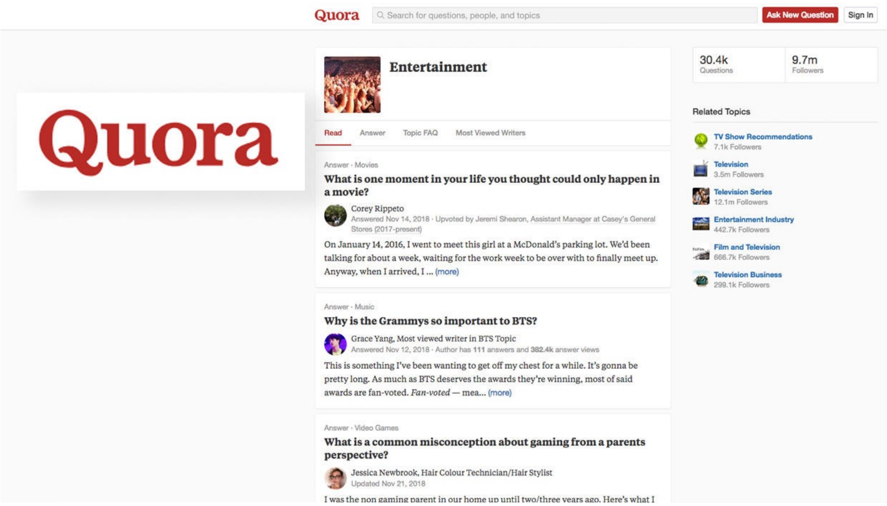


[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

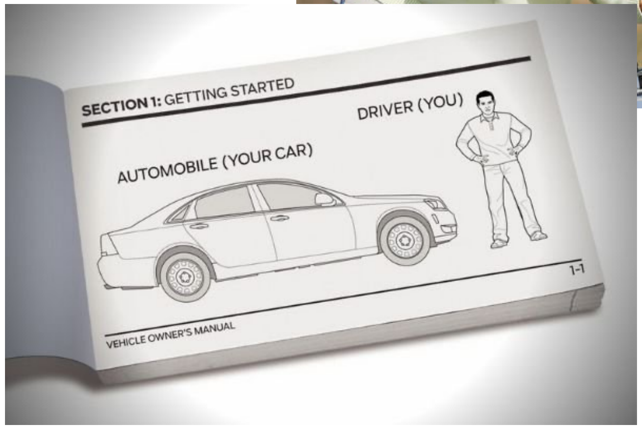
- Specific domain

Question answering



reddit

QA based on car manual?



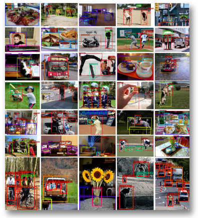
Problems with few data (labels)

- Privacy, security issues
- Expensive to collect/annotate
- Difficult / expertise-demanding to annotate
- Specific domain

Machine learning solutions given few data (labels)

- How can we make more efficient use of the data?
  - Clean but small-size
  - Noisy
  - Out-of-domain

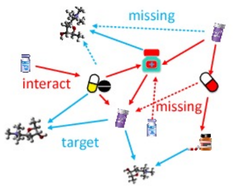
- Can we incorporate other types of experiences in learning?



Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints



Knowledge graphs

Rewards

Auxiliary agents

Adversaries

Master classes

…

And all combinations thereof

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

$$\min_{\theta} \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

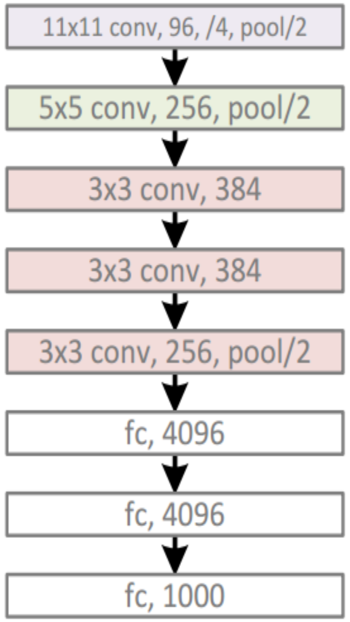
  - Neural networks
  - Graphical models
  - Compositional architectures

Components of a ML solution (roughly)

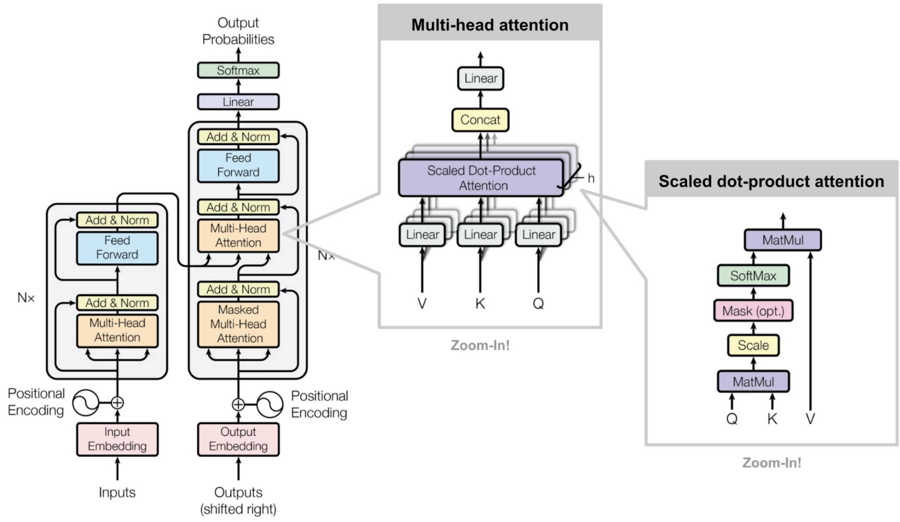
- Loss
- Experience
- Optimization solver
- Model architecture

Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

- Neural networks
- Graphical models
- Compositional architectures



Convolutional networks



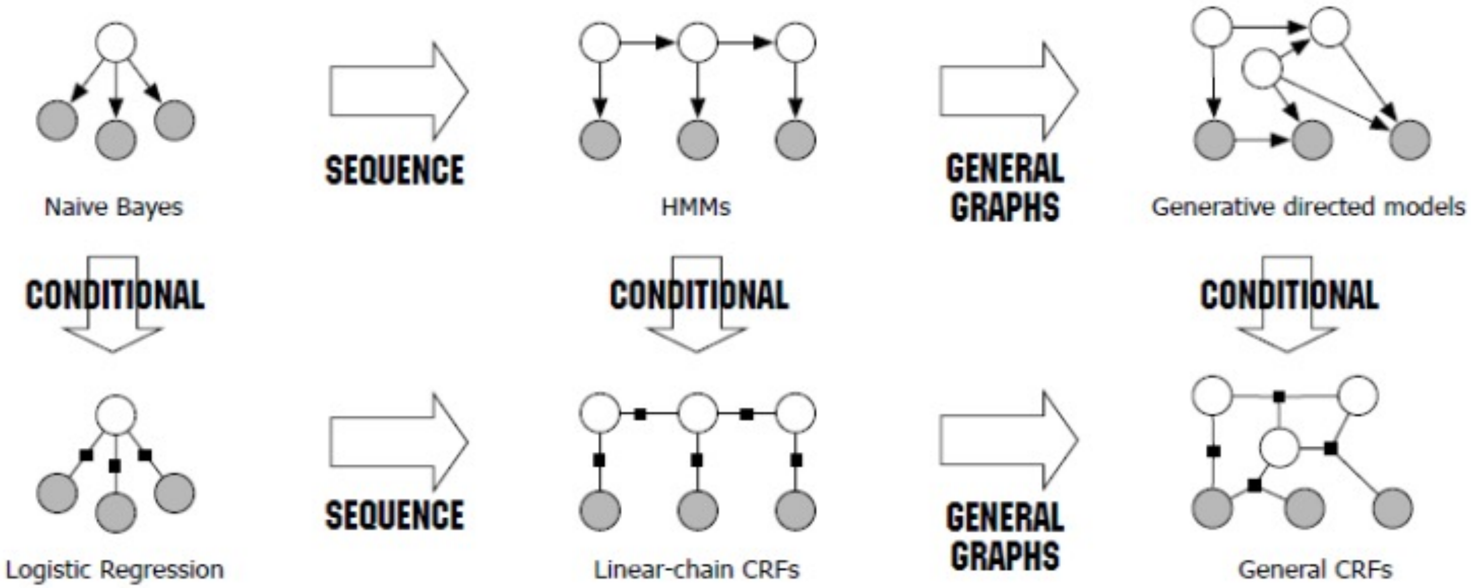
Transformers

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

- Neural networks
- Graphical models
- Compositional architectures



Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

Assuming you know basic procedures:

- (Stochastic) gradient descent
- Backpropagation
- Lagrange multiplier
- …

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

Core of most learning algorithms

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Machine learning solutions

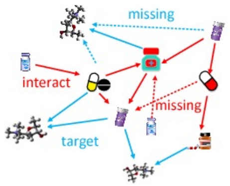
- (1) How can we make more efficient use of the data?
  - Clean but small-size, Noisy, Out-of-domain

- (2) Can we incorporate other types of experiences in learning?

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints



Knowledge graphs

Rewards

Auxiliary agents

Adversaries

Master classes

...

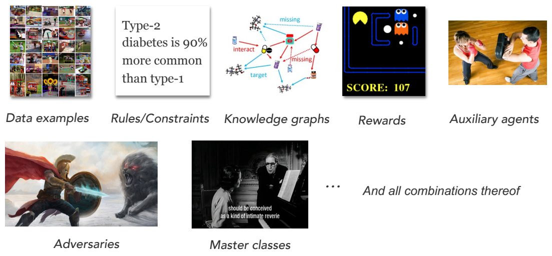
And all combinations thereof

Machine learning solutions

- (1) How can we make more efficient use of the data?
  - Clean but small-size, Noisy, Out-of-domain, …
- Algorithms
  - **Supervised learning**: MLE, maximum entropy principle
  - **Unsupervised learning**: EM, variational inference, VAEs
  - **Self-supervised learning**: successful instances, e.g., BERT, GPT-3, contrastive learning, applications to downstream tasks
  - **Distant/weakly supervised learning**: successful instances
  - **Data manipulation:** augmentation, re-weighting, curriculum learning, …
  - **Meta-learning**

Machine learning solutions

- (2) Can we incorporate other types of experiences in learning?



  - Learning from auxiliary models, e.g., adversarial models:
    - Generative adversarial learning (GANs and variants), co-training, …
  - Learning from structured knowledge
    - Posterior regularization, constraint-driven learning, …
  - Learning from rewards
    - Reinforcement learning: model-free vs model-based, policy-based vs value-based, on-policy vs off-policy, extrinsic reward vs intrinsic reward, …

  - Learning in dynamic environment
    - Online learning, lifelong/continual learning, …

Questions?