Introduction

He He

New York University

September 5, 2020

Logistics


He He


Gaomin Wu


Shivesh Ganju


Gauri Dhawan

- Best way to communicate: Piazza (remember to sign up).
- Lectures and office hours will be online (sadly).
- Let us know if you have accessibility needs.

What you'll be able to do by the end of this course

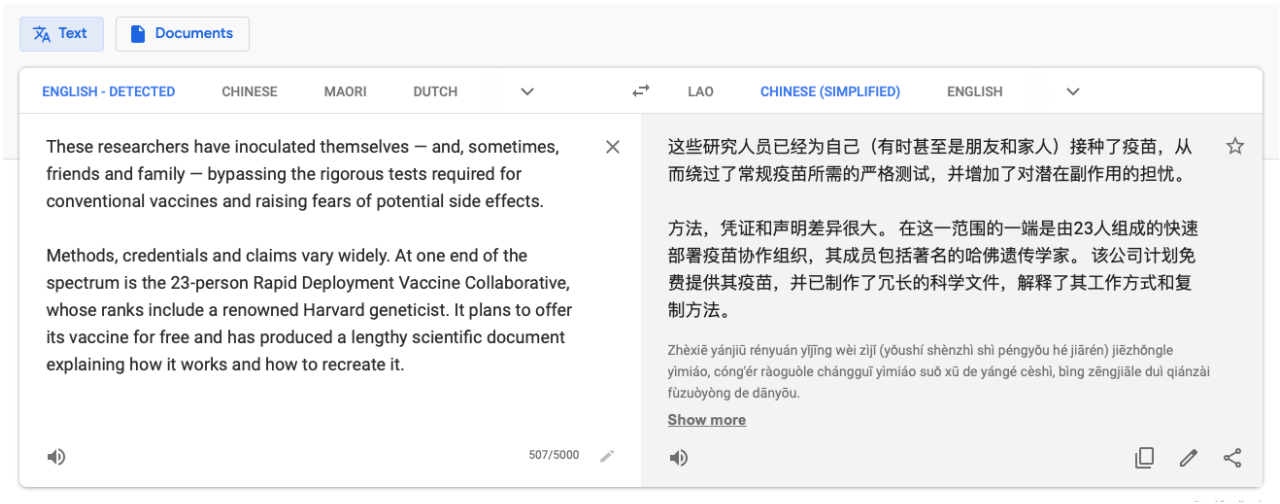
- Understand the core problems and challenges in NLP
- Formalize tasks as statistical learning problems
- Have a toolbox for solving different families of NLP problems
- Gain hands-on experience in building NLP systems

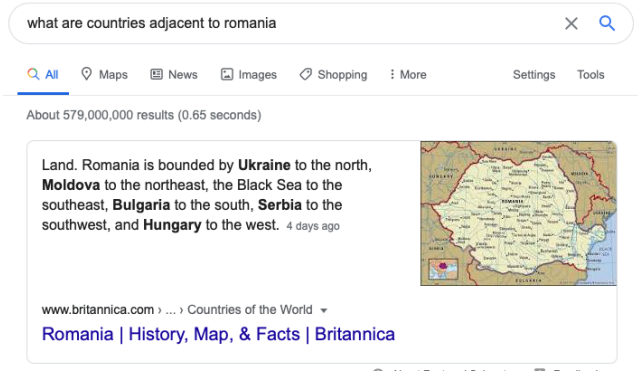
What we expect you to know

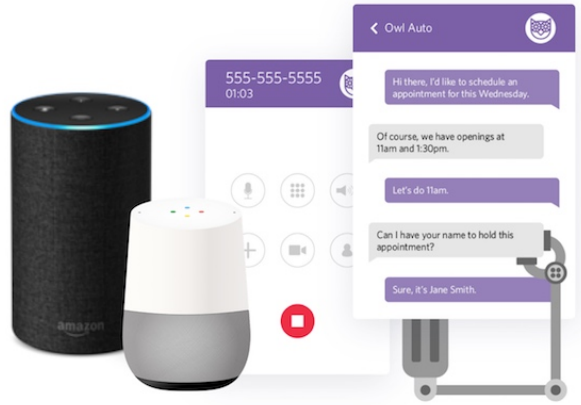
- **Linear algebra**: vector, dot product, gradient computation etc.
- **Probability and statistics**: conditional probability, expectation, Bayes rule etc.
- **Basic machine learning**: loss function, gradient descent etc.
- **Programming**: read and write Python code, use Numpy (and MXNet)

Table of Contents

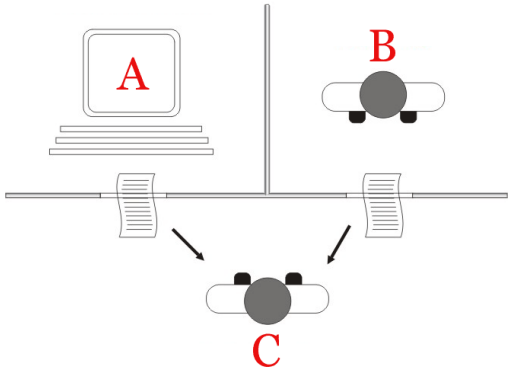
Products powered by NLP technologies







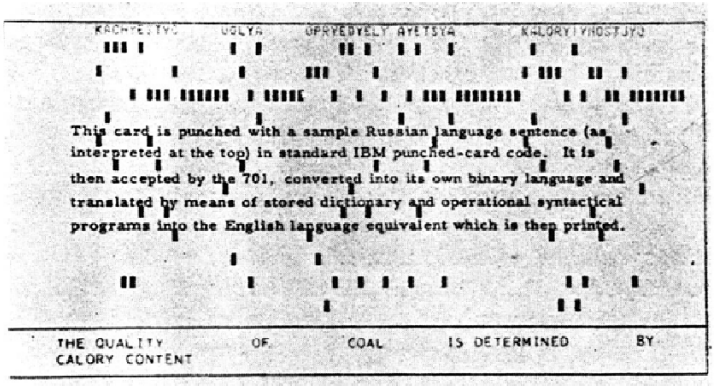
The imitation game



"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9*, to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning."* Turing (1950)

The Georgetown-IBM experiment

- The program:



- 250 words
- 6 grammar rules, e.g.
  "If first code is 110, is third code associated with preceding complete word equal to 21? If so, reverse order of appearance of words in output (i.e., word carrying 21 should follow that carrying 110)—otherwise, retain order."

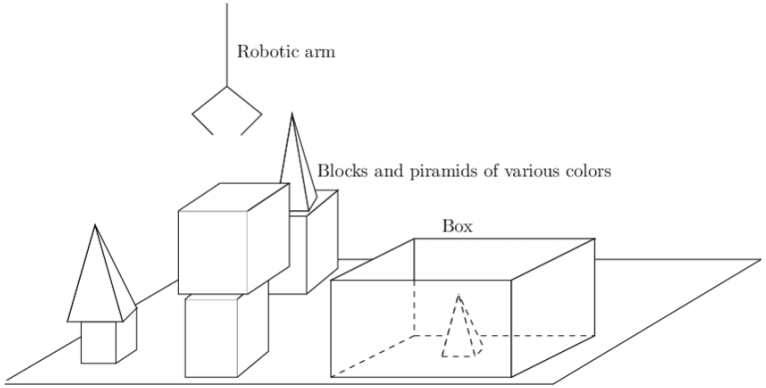
ELIZA

- Built by Joseph Weizenbaum at MIT in 1964 to demonstrate the superficiality of human-machine communication.
- Surprisinly, people were convinced that ELIZA had human intelligence.

Human: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
Human: He says I'm depressed much of the time.
ELIZA: I am sorry to hear you are depressed.
Human: It's true. I'm unhappy.
ELIZA: Do you think coming here will help you not to be unhappy?

SHRDLU

- Built by Terry Winograd at MIT in 1968.



Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Limitations of early systems

- Optimism in the 50's and 60's
  "Within the very near future—much less than twenty-five years—we shall have the technical capability of substituting machines for any and all human functions in organizations."

- Disappointing results due to
  - **Limited computation**: hardware limitation
  - **Combinatorial explosion**: algorithms are intractable in realistic settings
  - **Underestimated complexity**: ambiguity, commonsense knowledge etc.

The rise of statistical learning in the 80's

- Notable progress in MT from IBM (neglected knowlege of linguistics).
- HMMs widely used for speech recognition.
  "Every time I fire a linguist, the performance of the speech recognizer goes up."—Frederick Jelinek.
- Machine learning is the main driving force of NLP today.

Table of Contents

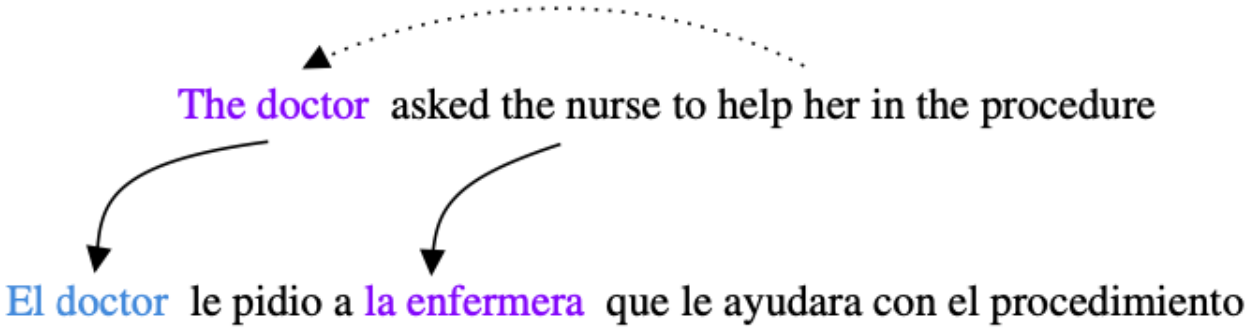
Are we there yet?

Predictions are not robust to benign perturbations [Ribeiro+ 2020].

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT***	Labels: negative, positive, neutral		
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
		Failure rate = 76.4%	

Are we there yet?

MT systems are prone to gender-biased translation errors [Stanovsky+ 2019].



Are we there yet?

QA models are easily distracted by irrelevant sentences [Jia+ 2017].

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Why is language hard?

sparsity

ambiguity

context-dependent

Why is language hard?

- **Discrete**
  - How to define metrics?
    I work at NYU. vs I work for NYU.
    This is good. vs This is actually good.
  - How to define transformations?
    The food is okay. → The food is awesome!
    They made a brief return to Cambridge to drop the book. → They returned.

Why is language hard?

- **Compositional**
  - How to generalize when we don't see all possible combinations?

    Vocabulary (from [Lake+ 2018]):

    {jump, walk, turn, once, twice, left, right, before, after, and}

    Sentences:

    jump
    jump left
    jump left and walk right
    jump left after walk right once before turn left twice
    ...

Why is language hard?

- **Sparse**
  - How to handle the long tail?

    BoA's financial assistant Erica:

    > The bank "learned [that] there are over 2,000 different ways to ask us to move money."

    Zipf's law:

    > word frequency $\propto \frac{1}{\text{rank}}$

Why is language hard?

- **Ambiguous**
  - How to interpret meaning in context?

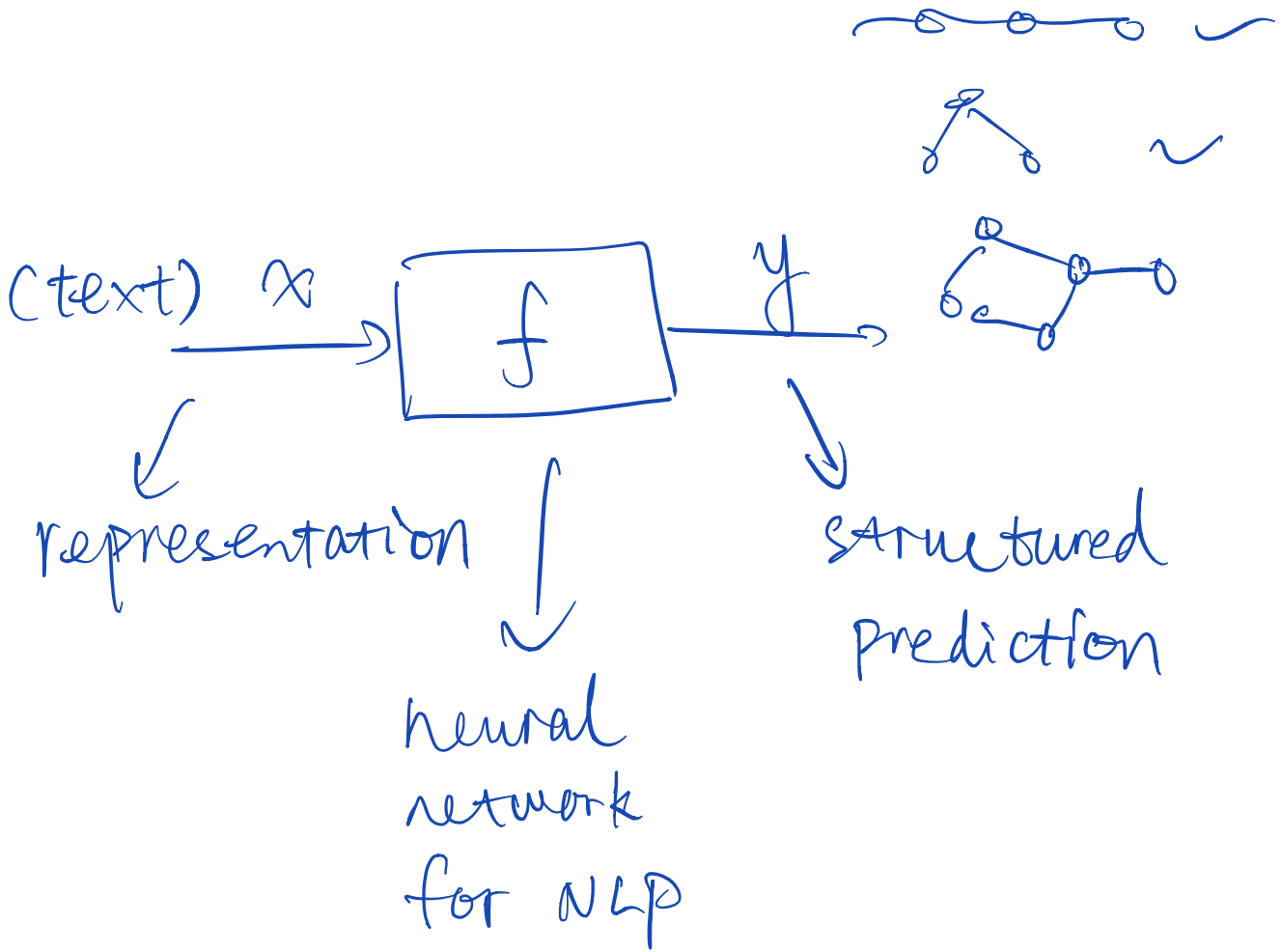
    Bass: fish? guitar? frequency?

    I shot an elephant in my pajamas: who is in the pajamas?

    The spirit is willing but the flesh is weak.
    $\rightarrow$ The vodka is strong but the meat is rotten.

Table of Contents

Overview


(text) x
f
y
representation
neural network for NLP
structured prediction

Representation of text

Symbolic: a set of objects (concepts)

$$\text{a white kitten} = \{\text{white}, \text{a}, \text{kitten}, \text{noun phrase}\}$$

Distributed: a list of components (properties)

$$\text{a white kitten} = [\text{COLOR=white}, \text{SIZE=small}$$
$$\text{OBJECT=kitten}, \text{NUMBER=1}]$$

Pros and cons?

Structured prediction: sequences

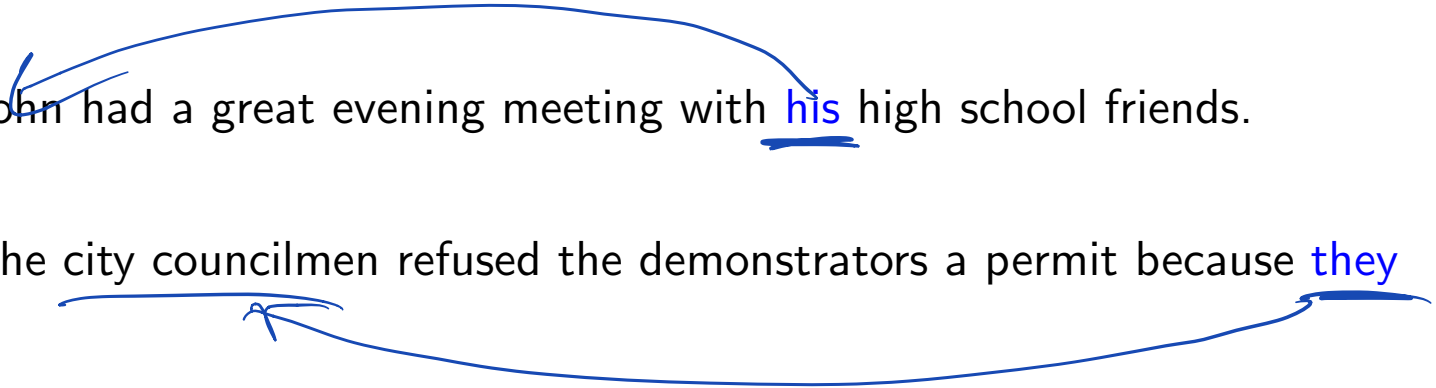
- Named entity recognition

New York University is a private research university based in New York City. It is founded in 1831 by Albert Gallatin.

CT of the maxillofacial area showed no facial bone fracture. CT of the brain showed no acute changes.

Structured prediction: sequences

- Anaphora resolution



Structured prediction: sequences

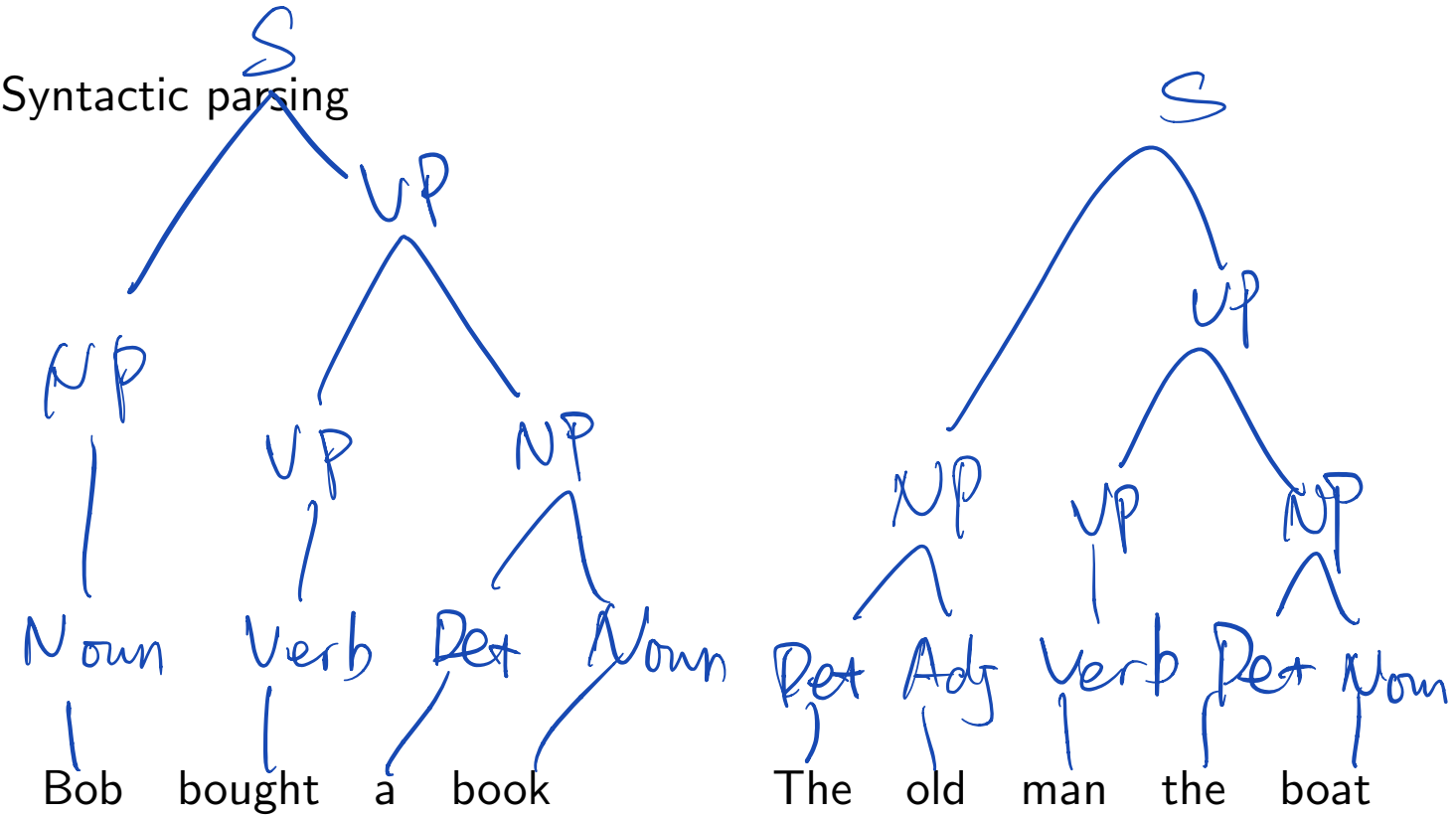
- Semantic role labeling (slot filling)

  I would like to book a ticket from New York to San Francisco on Christmas eve.

  action=`book_ticket`
  departure city=
  destination city=
  date=
  time=

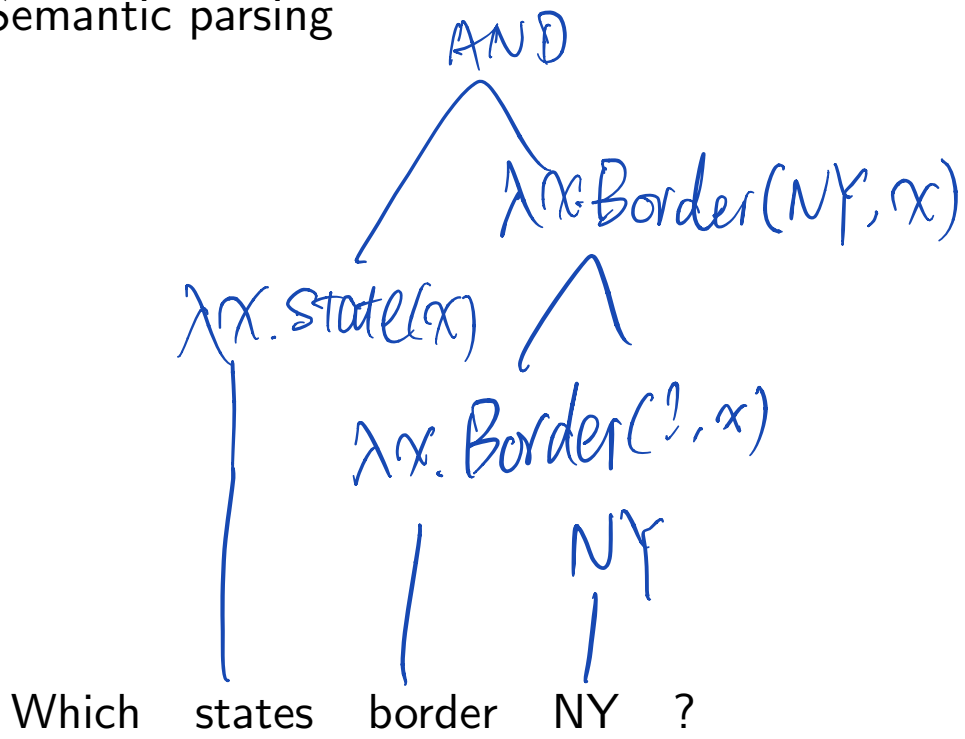
Structured prediction: trees

- Syntactic parsing



Structured prediction: trees

- Semantic parsing



Text generation

- Machine translation

  爱屋及乌 $\rightarrow$ Love me, love my dog

- Data-to-text

Date	min	max
tomorrow	21°C	29°C

$\rightarrow$ Tomorrow's temperature will be between 21 and 29 degrees.

- Summarization

The Justice Department plans to bring an antitrust case against Google as soon as this month, after Attorney General William P. Barr overruled career lawyers who said they needed more time to build a strong case against one of the world's wealthiest, most formidable technology companies, according to five people briefed on internal department conversations.

$\rightarrow$ Justice Dept. plans to file antitrust charges against Google in coming weeks.

Predict structures

- Modeling

  How to model iteractions among substructures?

- Learning

  How to efficiently learn the model parameters?

  $score(x, y)$

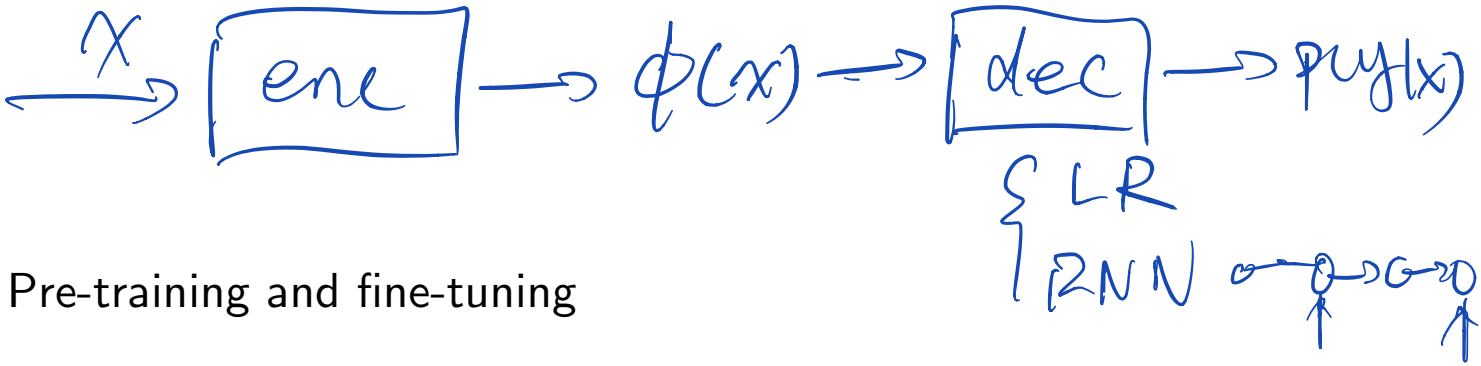
- Inference

  How to efficiently find the best structure given a learned model?

  $\arg\max_{y \in Y} score(x, y)$

Neural networks for NLP

- Encoder-decoder models



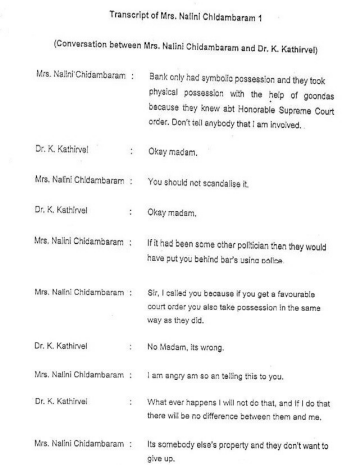
- Pre-training and fine-tuning

Beyond individual sentences: discourse

What makes a collection of sentences **coherent**?

John took a train from Paris to Istanbul. He likes spinach.

Jane took a train from Paris to Istanbul. She had to attend a conference.



Transcript of Mrs. Nalini Chidambaram 1

(Conversation between Mrs. Nalini Chidambaram and Dr. K. Kathirvel)

Mrs. Nalini Chidambaram : Bank only had symbolic possession and they took physical possession with the help of goondas because they knew abt Honorable Supreme Court order. Don't tell anybody that I am involved.

Dr. K. Kathirvel : Okay madam.

Mrs. Nalini Chidambaram : You should not scandalise it.

Dr. K. Kathirvel : Okay madam.

Mrs. Nalini Chidambaram : If it had been some other politician then they would have put you behind bar's using police.

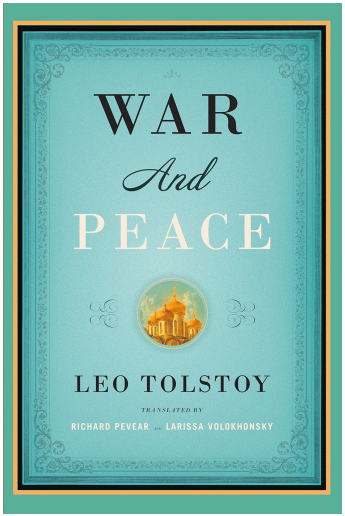
Mrs. Nalini Chidambaram : Sir, I called you because if you get a favourable court order you also take possession in the same way as they did.

Dr. K. Kathirvel : No Madam, its wrong.

Mrs. Nalini Chidambaram : I am angry am so an telling this to you.

Dr. K. Kathirvel : What ever happens I will not do that, and if I do that there will be no difference between them and me.

Mrs. Nalini Chidambaram : Its somebody else's property and they don't want to give up.



OPEN ACCESS Freely available online PLOS BIOLOGY

A Role for Parasites in Stabilising the Fig-Pollinator Mutualism

Derek W. Dunn[1,2,3], Simon T. Segar[1,2], Jo Ridley[4], Ruth Chan[1], Ross H. Crozier[4], Douglas W. Yu[5], James M. Cook[1,2,5]*

1 Division of Biology, Imperial College London, Ascot, United Kingdom, 2 School of Biological Sciences, University of Reading, Reading, United Kingdom, 3 School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, 4 School of Marine and Tropical Biology, James Cook University, Townsville, Queensland, Australia, 5 Natural Environment Research Council (NERC) Centre for Population Biology, Imperial College London, Ascot, United Kingdom

Mutualisms are interspecific interactions in which both players benefit. Explaining their maintenance is problematic, because cheaters should outcompete cooperative conspecifics, leading to mutualism instability. Monoecious figs (*Ficus*) are pollinated by host-specific wasps (Agaonidae), whose larvae gall ovules in their "fruits" (syconia). Female pollinating wasps oviposit directly into *Ficus* ovules from inside the receptive syconium. Across *Ficus* species, there is a widely documented segregation of pollinator galls in inner ovules and seeds in outer ovules. This pattern suggests that wasps avoid, or are prevented from ovipositing into, outer ovules, and this results in mutualism stability. However, the mechanisms preventing wasps from exploiting outer ovules remain unknown. We report that in *Ficus rubiginosa*, offspring in outer ovules are vulnerable to attack by parasitic wasps that oviposit from outside the syconium. Parasitism risk decreases towards the centre of the syconium, where inner ovules provide enemy-free space for pollinator offspring. We suggest that the resulting gradient in offspring viability is likely to contribute to selection on pollinators to avoid outer ovules, and by forcing wasps to focus on a subset of ovules, reduces their galling rates. This previously unidentified mechanism may therefore contribute to mutualism persistence independent of additional factors that invoke plant defences against pollinator oviposition, or physiological constraints on pollinators that prevent oviposition in all available ovules.

Citation: Dunn DW, Segar ST, Ridley J, Chan R, Crozier RH, et al. (2008) A role for parasites in stabilising the fig-pollinator mutualism. PLoS Biol 6(3): e59. doi:10.1371/journal.pbio.0060059

Introduction

In a biosphere driven by selection at the level of the individual gene [1], explaining the existence of cooperation, such as mutualism, is a major scientific challenge. Mutualisms are interspecific ecological interactions characterised by reciprocal benefits to both partners [2] that usually involve costly investments by each. What factors thus prevent one partner from imposing unsustainable costs onto the other to enable mutualism stability [3–7]? In some mutualisms, the larger, more sessile partner, manipulates the other by directing benefits to cooperative individuals and costs to cheaters [4–7]. However, a general consensus on mutualism persistence has only recently been formulated, and this clearly shows that a high benefit-to-cost ratio of cooperating is one important factor [8,9].

Fig trees (*Ficus*) and their host-specific agaonid pollinator wasps are a classic example of an obligate mutualism [10,11]. The wasps pollinate the trees, and the trees provide resources for wasp offspring. In monoecious *Ficus*, female wasps push their way through a specialised entrance into receptive syconia (colloquially, "figs"), which are enclosed inflorescences. The wasps then pollinate the tree while depositing their eggs individually into ovules. Thus, each egg laid costs the tree one seed, but upon emergence, the female wasp offspring disperse that tree's pollen. Trees need to produce both wasps and seeds for the mutualism to persist, but natural selection should favour wasps that exploit the maximum number of fig ovules in the short term, resulting in a conflict of interest between wasp and tree. However, the mutualism has persisted for at least 60 million years and has radiated into more than 750 species pairs [12]. The mechanisms preventing wasps from overexploiting figs remain unknown, despite intensive study over four decades.

Within receptive syconia, the lengths of floral styles are highly variable [13,14], and ovipositing pollinators (long-dressers) favour flowers with shorter styles for their offspring [15–18]. Style and pedicel lengths of flowers are negatively correlated. Short-styled ovules develop into seeds or galls (when a wasp is present) near the syconium inner cavity, while most long-styled ovules develop into seeds near the outer wall [19,20] (Figure 1). These patterns have been shown to reflect the oviposition preferences of foundresses, and are unlikely to be the result of greater elongation of pedicels containing eggs during syconial maturation, because in receptive syconia, pollinators' eggs are mainly present in short-styled inner ovules [16]. These widespread observations have been tied to four, not necessarily mutually exclusive, mechanisms that have been proposed to stabilise the fig-pollinator mutualism: (1) Unbeatable seeds—outer ovules may be defended biochemically or physically against oviposition or larval development [21]. However, no mechanism has yet been identified. (2) Short ovipositors—pollinators' ovipositors may be too short to fully penetrate the long styles of

Academic Editor: Anurag A. Agrawal, Cornell University, United States of America

Received September 14, 2007; **Accepted** January 22, 2008; **Published** March 11, 2008

Abbreviation: s.e., standard error

* To whom correspondence should be addressed. E-mail: james.cook@reading.ac.uk

Beyond individual sentences: grounding

Connect language to the world

"Can you bring me an apple?"

Project

- Related to NLP (doesn't have to be in the scope of this course)
- New algorithms or models for existing problems
- Novel applications of NLP techniques
- Analysis of well-known approaches that leads to new insight
- Overall rule: should increase our knowledge in some way