Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 1

Topics in NLP

- We will cover the basic methodology in NLP
- There will be a focus on statistical learning
- Even more so, *structured prediction*

Topics in NLP

Prerequisites:

- Some familiarity with machine learning and probability
- If something is unclear, ask!

Things to Do:

- Student presentations (20%)
- Brief paper responses (25%)
- Essay (55%)

Office hours: By appointment

NLP in the Old Days

1950s-1980s: handwritten rules

IBM'S WATSON (right) AND FRIENDS:† For a mathematical wizard . . .

NLP Now

late 1980s until now: statistical learning

Learning

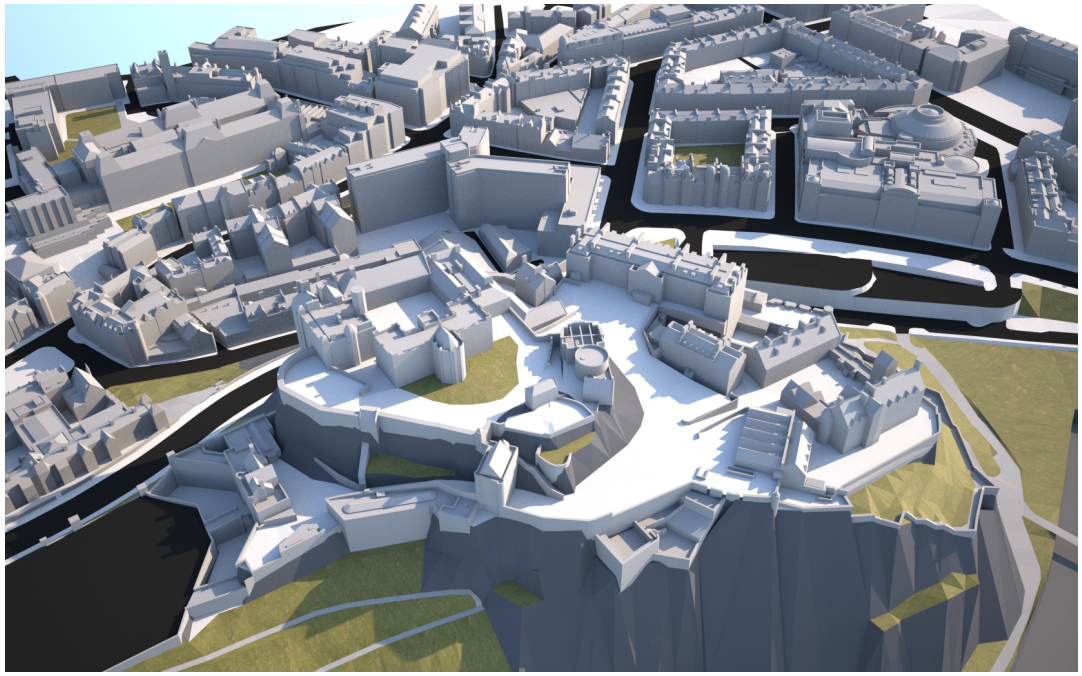
Learning is:

- Experience translated into expertise/knowledge
- Memorisation with generalisation

Machine learning and NLP:

- Experience = Training data
- Knowledge = Decoder or Prediction Model
- Used to either **mimic** humans or **transcend** their abilities

What is a Model?



From Merriam-Webster:

- a usually small copy of something
- a set of ideas and numbers that describe the past, present, or future state of something (such as an economy or a business)

When is a model a good model?

What is a Statistical Model?

Predict the future. Probabilistically.

Probability and Statistics: Reminder

Probability distribution? Example: unigram model

$\Omega = \{$ the, cat, chased, on, a, ... $\}$ sample space

$p(w) \geq 0 \qquad \sum_w p(w) = 1$

Random variables

Random variable:

$X(w) = \Omega_2$

$X(\text{jumping}) = \text{ing}$

$X(\text{jumped}) = \text{ed}$

$\vdots$

$X(w)$ returns the last two letters

$$p(X(w) = ed) = \sum_{\substack{w \text{ ends} \\ \text{in ed}}} p(w)$$

$$p(X(w) = 1) = \sum_{\substack{X(w)=1 \\ w \text{ is related} \\ \text{to a movie}}} p(w) = \sum_{w} p(w) X(w) = E[X]$$

$X(w)$ is 1 if the word is related to movies

$X(w)$ is 0 o/w

Model Family

A set of probability distributions (unigram example):

$$\{ p_1(w), p_2(w), p_3(w), \ldots \}$$

Parameters

A set of parameters: Θ

$p(w|\theta)$ = probability that depends on θ

$\{ p(w|\theta) \mid \theta \in \Theta \}$

$p_1(w) \ldots p(w|\theta_1)$

$p_2(w) \quad p(w|\theta_2) \ldots$

$|\Omega|$

$\theta \in \mathbb{R}$ $\qquad \theta_i \in (0,1) \qquad \sum_i \theta_i = 1$

Another Parametrisation

Rely on properties of the words:

We can build parameterisations that treat words as "groups"

Estimation

What is training data?

$w_1, w_2, w_3, \ldots, w_n$

Estimation

What is the fit of the data to the model?

$f(w_1, \ldots, w_n, \theta)$ tells us whether θ predicts well $w_1, \ldots, w_n$

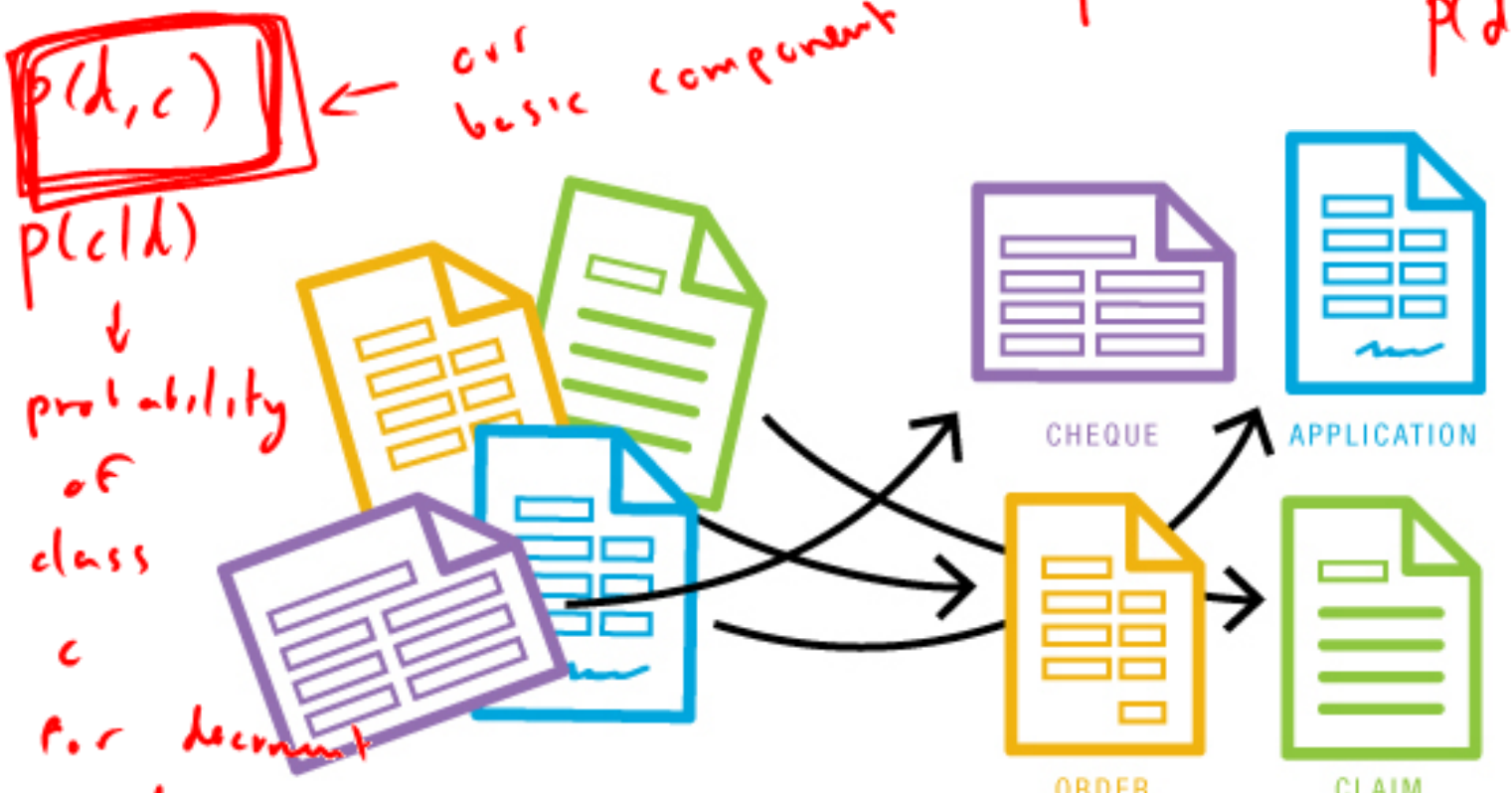
$$\theta^* = \arg\max_{\theta} f(w_1 \ldots w_n, \theta)$$

θ^* over final model

NLP Problem Example: Document Classification

sentiment analysis, document topic, ...

$$p(c|d) = \frac{p(d,c)}{p(d)}$$



$$p(d) = \sum_c p(d,c)$$

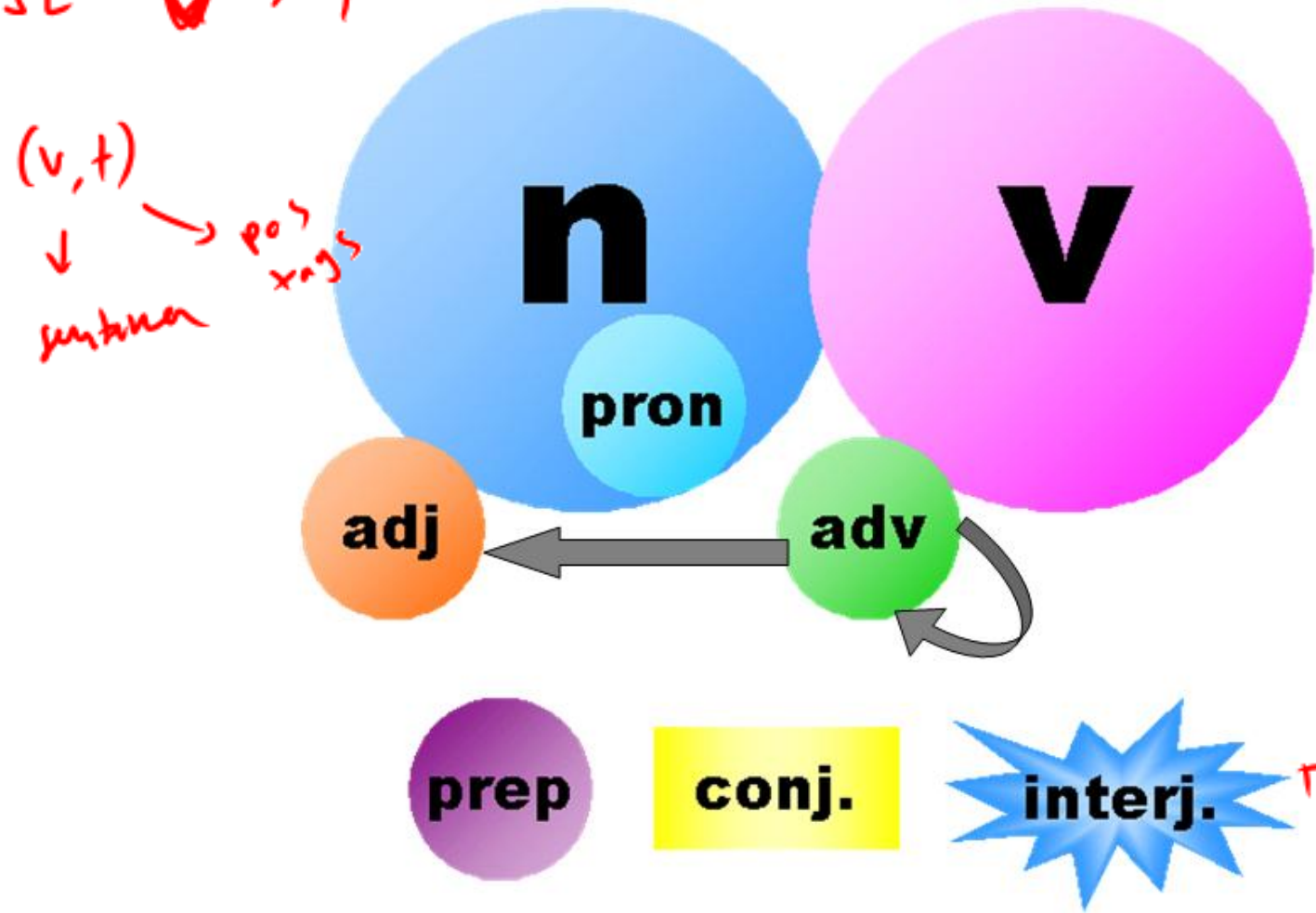
$$= \sum_c p(d|c)\, p(c)$$

$$\Omega = D \times C$$

$(d,c) \in \Omega$ d is the document c-th class

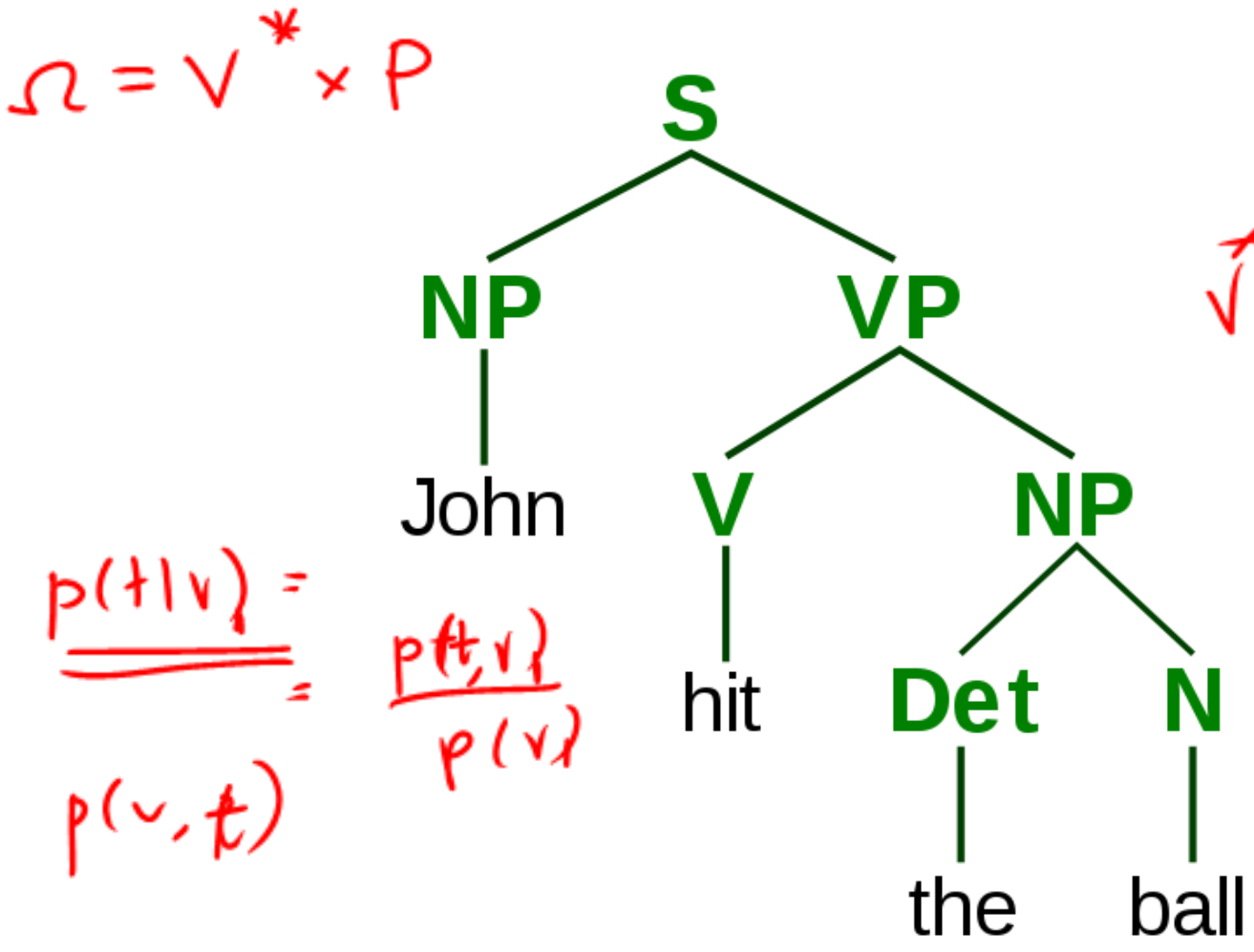
NLP Problem Example: POS Tagging

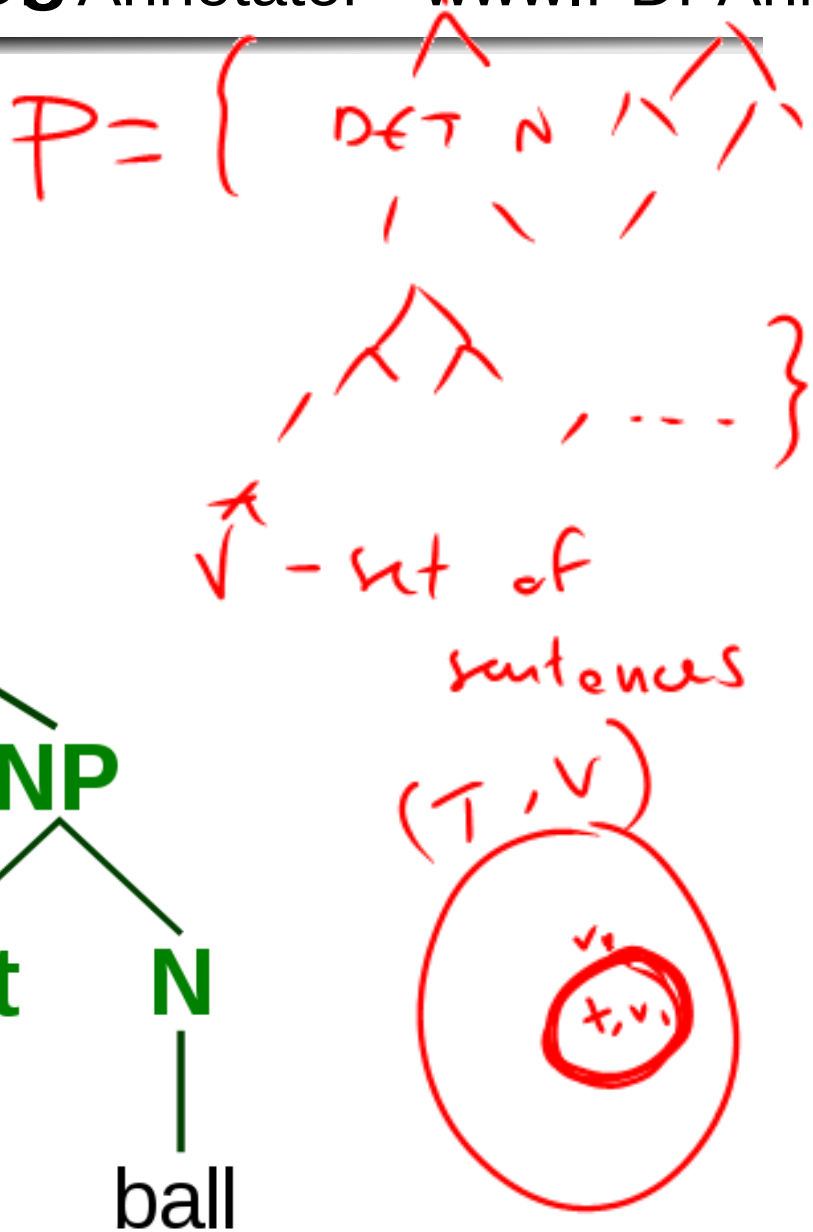
map words to their part-of-speech tags



NLP Problem Example: Parsing

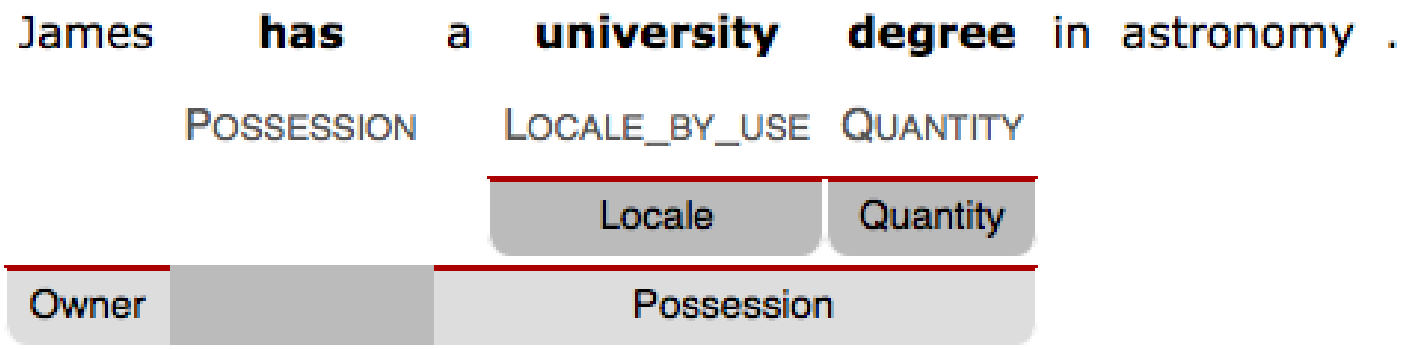
map sentences to their syntax





NLP Problem Example: FrameNet Parsing

find predicate-argument structure



Back to Modelling

What if the space to model is complex? Modelling documents.

Modelling a Problem

- Define a sample space
- Define the structure of the sample space
- Decide on a parametrisation

Then one can proceed with data collection and learning

Modelling - Tradeoffs

- “Exact copy”, detailed
- Not too many parameters
- Efficient to work with

Next class

Paradigms in statistical learning

- Frequentist approaches
- Bayesian approaches
- “Computer science approaches?”