Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 8

Learning from Incomplete Data

- Semi-supervised learning
- Latent variable learning
- Unsupervised learning

How to estimate a PCFG?

We learned how to estimate a PCFG from treebank

Reminder:

Unsupervised learning: PCFGs

How to estimate a PCFG from strings?

General case: Viterbi (or “hard”) EM

Model:

Observed Data:

Step 0:

Step 1:

Step 2:

Repeat steps 1–2

Maximum likelihood estimation

General principle: write down the likelihood of **whatever** you observe, and then maximise with respect to parameters

Model: $p(x, y \mid \theta)$

Observed: $x_1, \ldots, x_n$

Likelihood:

$L(x_1, \ldots, x_n \mid \theta) =$

The EM Algorithm

- A softer version of hard EM
- Instead of identifying a single tree per sentence, identify a distribution over trees (E-step)
- Then re-estimate the parameters, with each tree for each sentence “voting” according to its probability (M-step)
- Semiring parsing: instead of CKY use the inside algorithm

EM: Main Disadvantage

Sensitivity to initialisation (finds local maximum)

Global log-likelihood optimisation in general is “hard”

Latent-variable learning

"Structure" is present

Some information is missing from model

Model: $p(x, y, h \mid \theta)$

Observed: $(x_1, y_1), \ldots, (x_n, y_n)$

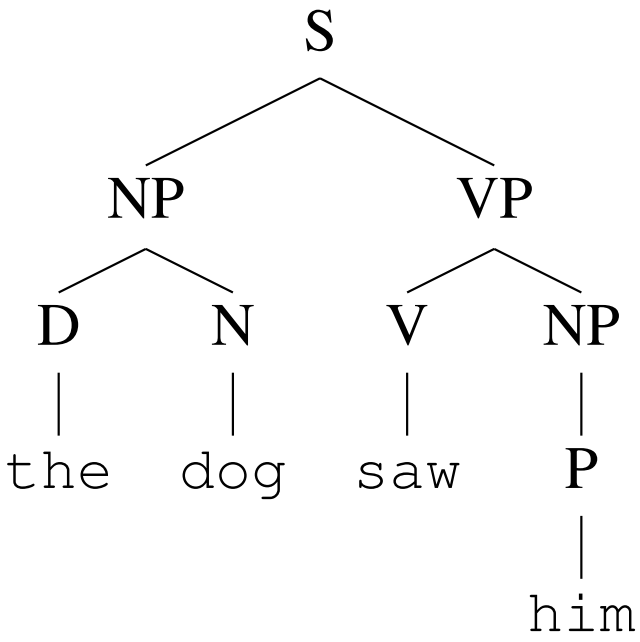
Log-likelihood:

$L(x_1, \ldots, x_n, y_1, \ldots, y_n|\theta) =$

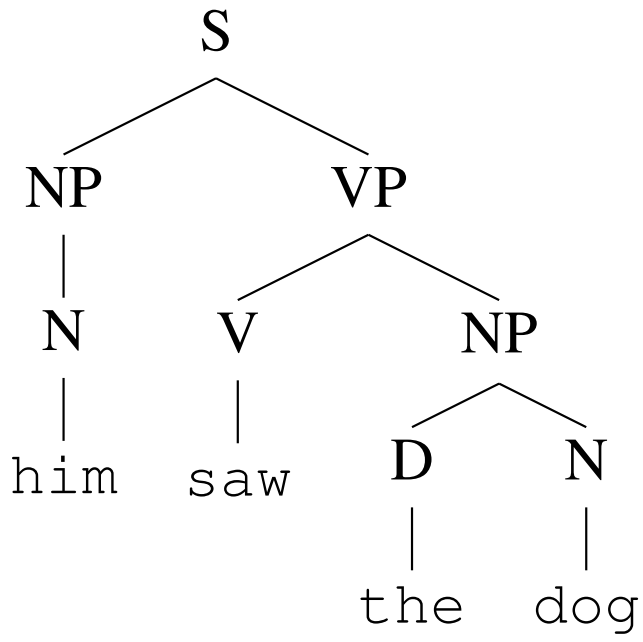
Example of Latent-Variable Use in PCFGs

"Context-freeness" can lead to over-generalisation:

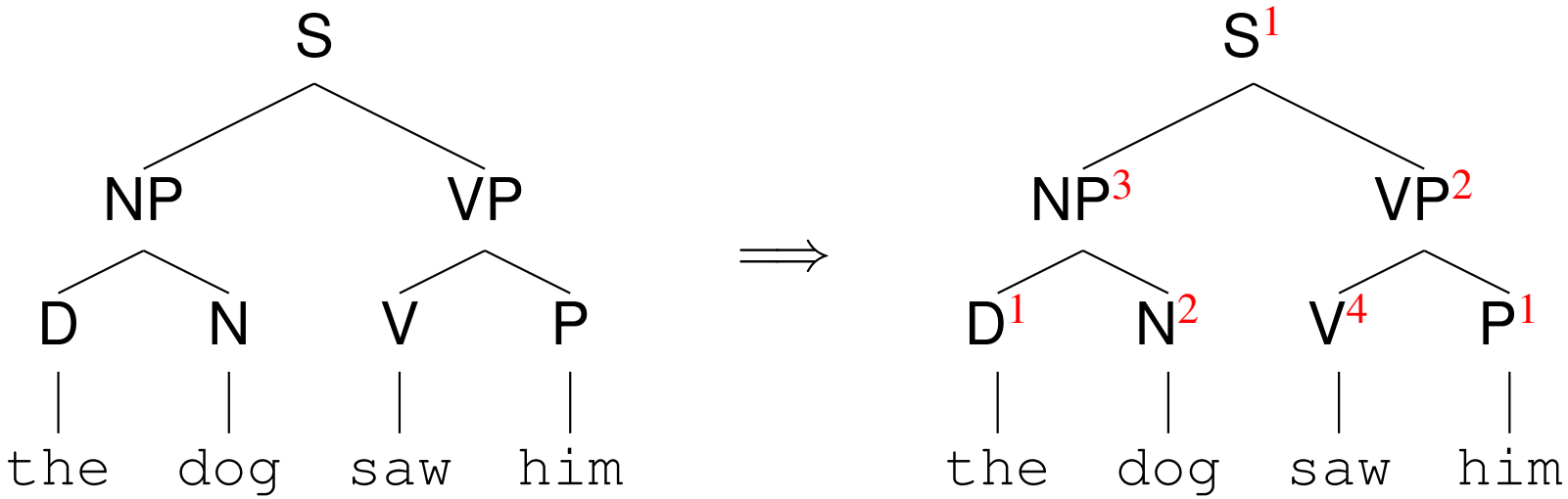
Seen in data:



Unseen in data (ungrammatical):



Latent-Variable PCFGs



The **latent states** for each node are never observed

How to learn with latent variables?

- Expectation-Maximisation (EM)
- Spectral learning
- Neural networks
- Other methods

Semi-supervised Learning

Main idea: use a relatively small amount of annotated data, and exploit also large amounts of unannotated data

The term itself is used in various ways with various methodologies

Example: Word Clusters and Embeddings

- Learn clusters of words or embed them in Euclidean space using large amounts of text
- Use these clusters/embeddings as features in a discriminative model

Semi-supervised Learning: Example 2

Combine the log-likelihood for labelled data with the log-likelihood for unlabelled data

$$L(x_1, y_1, \ldots, x_n, y_n, x'_1, \ldots, x'_m|\theta) =$$

Semi-supervised Learning: Example 3

Self-training

Semi-supervised Learning: Example 3

Self-training

Step 1:

Step 2:

Step 3:

Potentially, repeat step 2

Summary

- Learning from incomplete data alleviates the need to annotate data
- Three ways to use incomplete data: unsupervised learning, semi-supervised learning and learning with latent-variables