Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 4

Administrativia

- Is everybody getting my emails? There are a few addresses that have emails bounce back.

Last class

Bayesian inference: $p(\theta)$ prior $p(w_1 \ldots w_n | \theta) = \prod_{i=1}^{n} p(w_i | \theta)$

$$p(\theta | w_1, \ldots w_n) = \frac{p(\theta)\, p(w_1 \ldots w_n | \theta)}{p(w_1 \ldots w_n)}$$

$$p(w_1 \ldots w_n) = \int_{\theta} p(w_1 \ldots w_n | \theta)\, p(\theta)\, d\theta$$

MAP - maximum a posteriori estimate

$$\theta^* = \arg\max_{\theta} p(\theta | w_1 \ldots w_n).$$

MAP and posteriors

In general,

- Priors are especially important when the amount of data is small
- As there is more data, the prior becomes less influential on the posterior
- Under some mild conditions, the posterior is a distribution concentrated around the MLE

Conjugacy of prior and likelihood

similar "Beta"

$$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\beta} \qquad\qquad p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$$

Prior is “hyperparametrised”. What is the posterior?

$$p(\theta|w_1 \ldots w_n) \propto \theta^{\underbrace{\alpha+a}_{\alpha'}}(1-\theta)^{\underbrace{\beta+b}_{\beta'}}$$

$$p(\theta|w_1 \ldots w_n) \propto \theta^{\alpha'_1}(1-\theta)^{\beta'}$$

Definition of Conjugacy

$[0,1]$ $\{(\alpha, \beta)\}$

Let P be a set of priors hyperparametised by a set $\mathcal{A}$, for a parameter space Θ. Therefore, each $p \in P$ is a probability distribution $p(\theta \mid \alpha)$. Let M be a model over Ω such that each $p \in M$ is a probability distribution $p(w \mid \theta)$. We say, P **is conjugate to** M, if for any choice of $\alpha \in \mathcal{A}$ and data $w_1, \ldots, w_n$ it holds that $p(\theta \mid w_1, \ldots, w_n, \alpha) \in P$.

Definition of Conjugacy

Let P be a set of priors hyperparametised by a set $\mathcal{A}$, for a parameter space Θ. Therefore, each $p \in P$ is a probability distribution $p(\theta \mid \alpha)$. Let M be a model over Ω such that each $p \in M$ is a probability distribution $p(w \mid \theta)$. We say, P **is conjugate to** M, if for any choice of $\alpha \in \mathcal{A}$ and data $w_1, \ldots, w_n$ it holds that $p(\theta \mid w_1, \ldots, w_n, \alpha) \in P$.

Previous example (argh-blah example):

$M =$ $\{ p(w \mid \theta) \mid \theta \in [0,1] \}$

$P =$ $\{ \theta^{\alpha}(1-\theta)^{\beta} = p(\theta) \mid \alpha, \beta \geq 0 \}$

Posterior new hyperparameters:

Conjugacy – always useful?

Trivial non-useful example of conjugacy

$P = \{$set of all distributions $p(\theta)\}$

$M =$ some model

Conjugacy – always useful?

Another trivial non-useful example of conjugacy

choose some θ_0

$\mathcal{P} = \{ p(\theta) \}$ such that

$$p(\theta) = \begin{cases} 1 & \theta = \theta_0 \\ 0 & o/w \end{cases}$$

$\mathcal{M}$ $p(w \mid \theta)$

$$p(\theta \mid w) = \frac{p(\theta)\, p(w \mid \theta)}{p(w)} = \begin{cases} 1 & \theta = \theta_0 \\ 0 & o/w \end{cases} = \in \mathcal{P}$$

Conjugacy: summary

Conjugacy is useful when:

- The prior is not too poor
- It is easy to calculate the posterior hyperparameters

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1, \ldots, w_n)$? # bits required to encode

θ given $w_1 \ldots w_n$

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$? see previous slide

What is $-\log_2 p(\theta)$?

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?

What is $-\log_2 p(\theta)$? # bits required to encode θ according to the prior

What is $-\log_2 p(w_1,\ldots,w_n|\theta)$? # bits required to encode the data for a given θ

Minimum Description Length and MAP

What is $-\log_2 p(\theta | w_1, \ldots, w_n)$?

What is $-\log_2 p(\theta)$?

What is $-\log_2 p(w_1, \ldots, w_n | \theta)$?

MAP: $\theta^* = \arg\max_\theta \log_2 p(\theta) + \log_2 p(w_1, \ldots, w_n | \theta)$

Encoding θ^* requires separately:

- Encoding the hypothesis according to the prior
- Encoding the data according to the hypothesis

That's the "minimum description length" criterion

$\theta^* = \arg\max p(\theta | w_1 \ldots w_n) =$

$= \arg\max \dfrac{p(w_1 \ldots w_n | \theta)\, p(\theta)}{p(w_1 \ldots w_n)} =$

$= \arg\max p(w_1 \ldots w_n | \theta)\, p(\theta) =$

$= \arg\max \log_2 p(w_1 \ldots w_n | \theta) + \log_2 p(\theta)$

$= \arg\min_\theta -\log_2 p(w_1 \ldots w_n | \theta) - \log_2 p(\theta)$

$\boxed{\text{MDL}}$

Summary

Bayesian analysis:

- Only uses Bayes' rule to do inference
- Posterior is a *distribution* over parameters
- Can summarise the posterior, e.g. MAP, to get a point estimate
- Need to be careful about choice of prior
- Especially important with small amounts of data
- MAP has a connection to minimum description length (MDL)

Today’s class

What is our Ω?

parse tree → dep → constit

sequence

pos tags

FSA

Today's class

What is our Ω?

Examples:

- Finite sets of symbols (such as a set of words)
- Sequences
- Trees - "dependency" and others
- Graphs and hypergraphs
- Miscellaneous - tailored to a specific problem

Bag of words

$\Omega = \{ \text{documents} \}$ $\quad d = (w : c)_{w \in V}$

- Does not have much structure
- Still, a very useful way to decompose the space of documents
- Especially when interested in "content" and not "syntax"
- We will re-visit this model later

Segmentation

$\Omega =$ { sentences } { words } ... $\Rightarrow$ { strings } $\times$ { segmentations for the string }

$|w_1 + \ldots | w_n|$

Useful for:

- Segmentation of languages such as Chinese
- Identifying co-locations (New York)
- Tokenisation
- Sentence segmentation (a "solved" problem)
- Morphological segmentation (for example, Turkish)

$\boxed{\text{Mr.}}$ Mouse ate the cheese$\boxed{.}$

Sequence labelling

$\Omega =$ $V^* \times T^*$ T - set of labels V - vocabulary

When is it useful?

- Part-of-speech tagging
  - POS tagging using majority vote: 90%
  - POS tagging using sequence labelling: 97%
- Whenever context is needed to decipher an observation

Chunking

$\Omega = \{ \text{strings} \} \times \{ \text{substrings} \}$

Mr. Mouse ate the cheese
C C O O O
B I O O O

When is it useful?

B-I-O

- Shallow parsing (or as a precursor to full parsing)
- Identifying named entities
- Connection to sequence labelling?

training data → strings, → chunks ⟹ sequence labelling (changing representation.)

Mr. Mouse ate the cheese

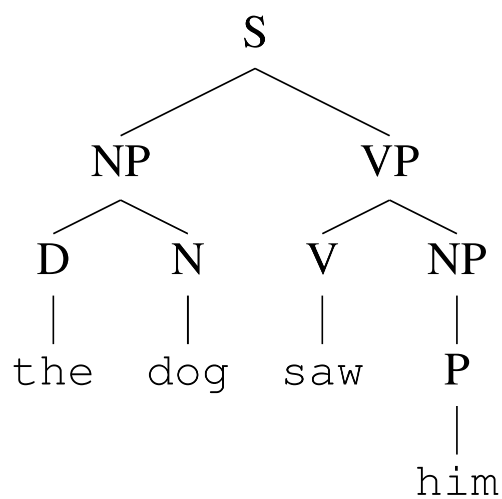
$\boxed{\text{Mr. Mouse}}$

Parsing

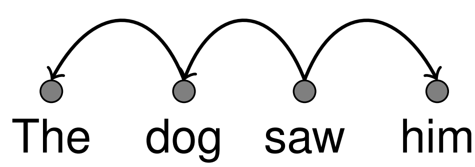
$\Omega =$

Two main types of parsing structures:

- Constituency

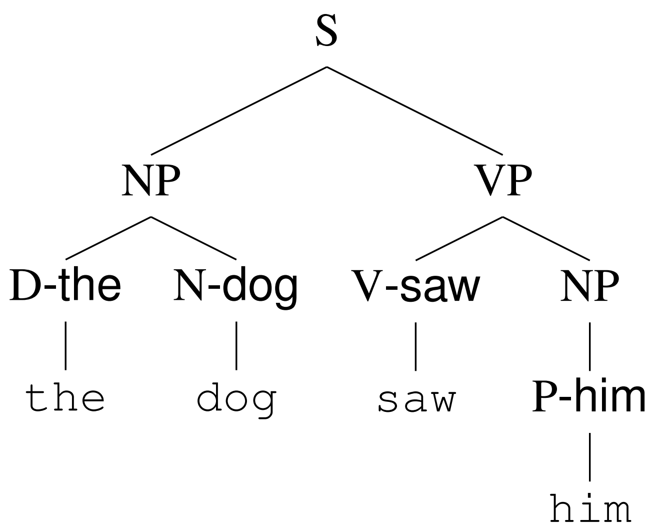


- Dependency

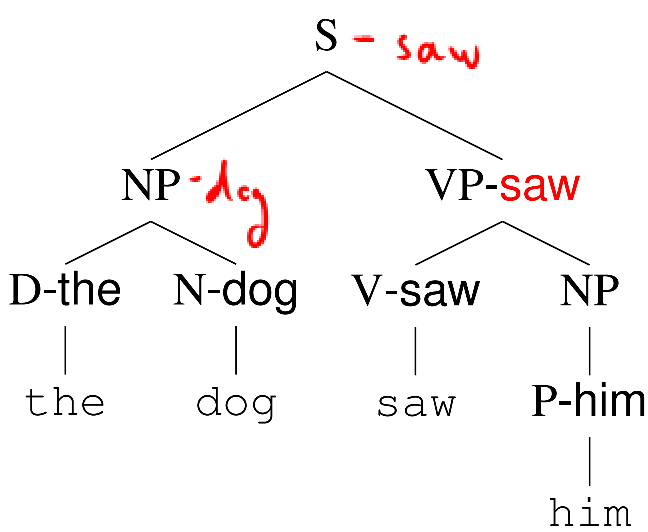


encode relations directly
(head-modifier)

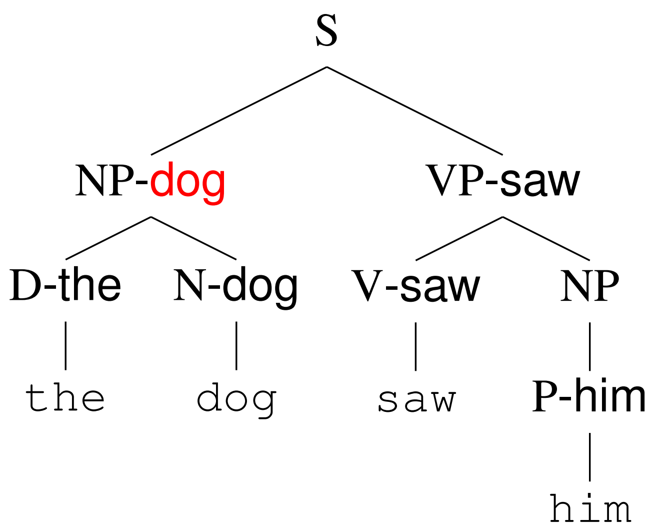
Conversion of constituency to dependency



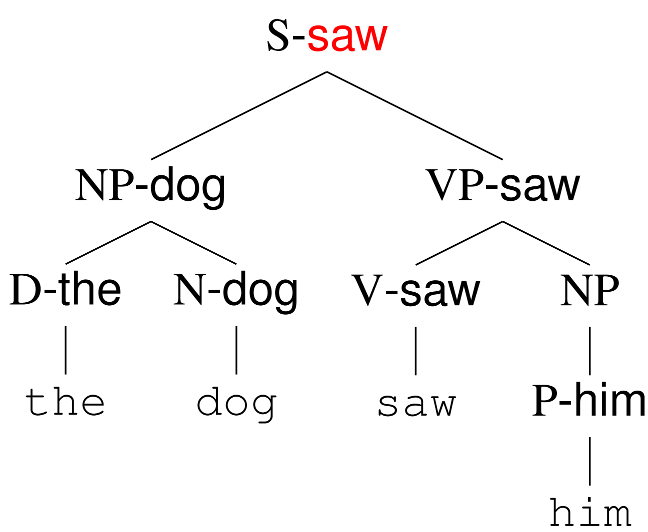
Conversion of constituency to dependency



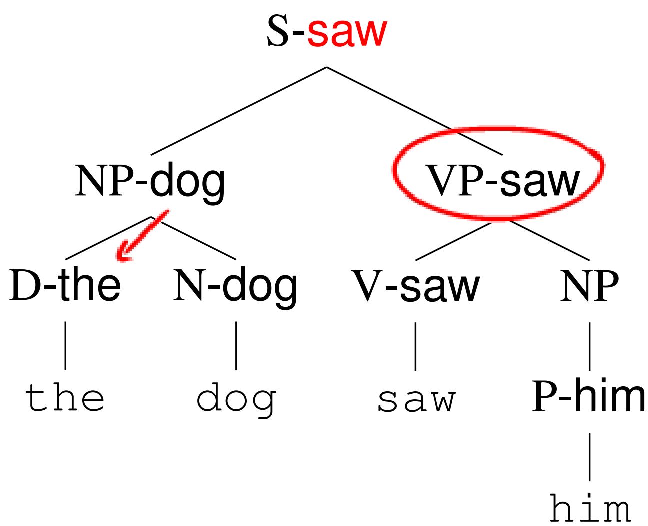
Conversion of constituency to dependency



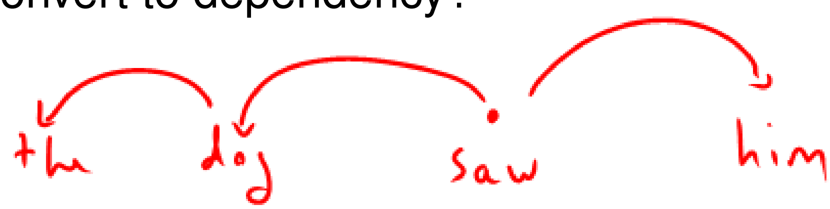
Conversion of constituency to dependency



Conversion of constituency to dependency



How to convert to dependency?



Conversion of dependency to constituency

Not trivial

Some information is lost (syntactic categories)

But at least can recover the spans of the “constituents”

Projective vs. non-projective parsing

Projective trees:

We are never going to have crossing edges if we draw the edges above the sentence.

Non-projective trees:

We saw a house on Tuesday that we liked