

TTIC 31210:
Advanced Natural Language Processing

Kevin Gimpel
Spring 2017

Lecture 11:
Inference in Bayesian NLP

- Project proposals due Wednesday, May 10

Bayesian Modeling

- this typically means:
 - defining parameters as random variables and giving them prior distributions
 - marginalizing out random variables whenever possible (like those parameter random variables)
 - aiming to compute posterior over all latent variables conditioned on observations (and maybe some hyperparameters)

Motivation

- a Bayesian approach is more impactful when there are latent variables
- NLP has a lot of models with latent variables
- unsupervised learning in NLP: “consider the unseen output as a latent variable”

Topic Modeling

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden. "We arrived at the 800 number, but coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

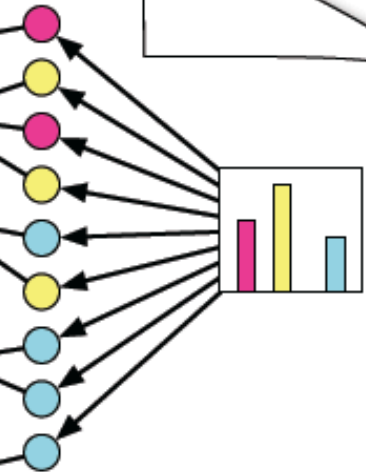


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

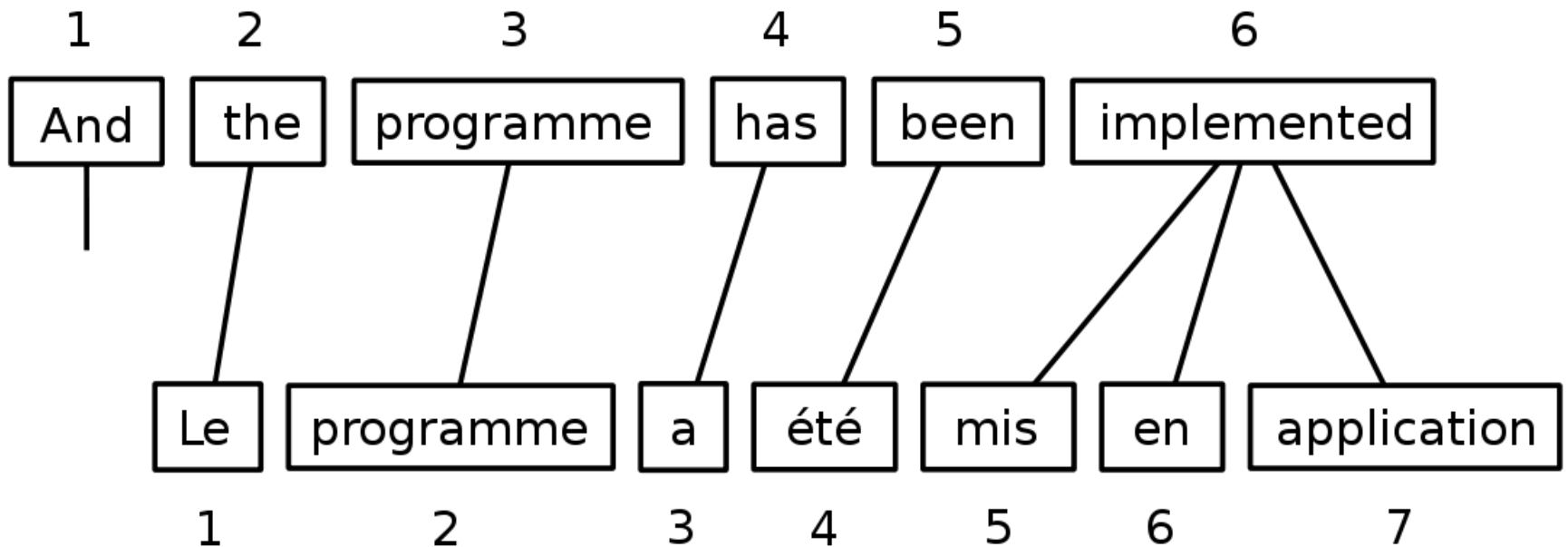
Topic proportions and assignments



Blei et al. (2003)

Word Alignment

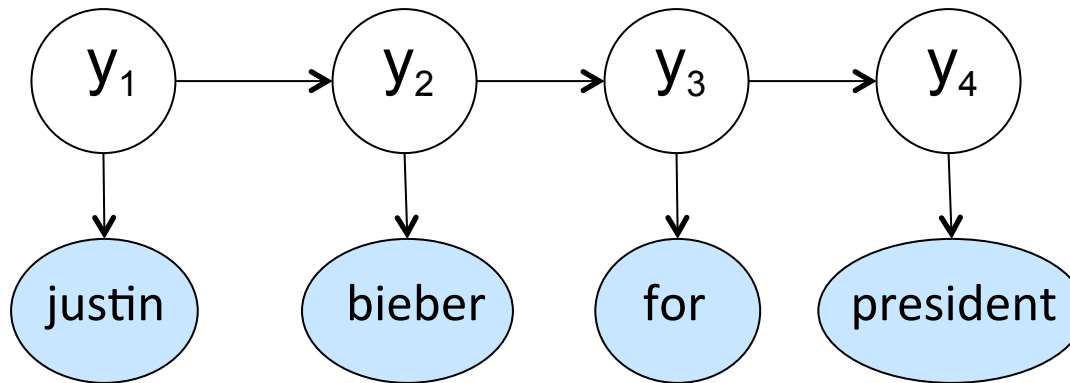
parallel sentences are observed, word alignments are latent variables:



Brown et al. (1990)

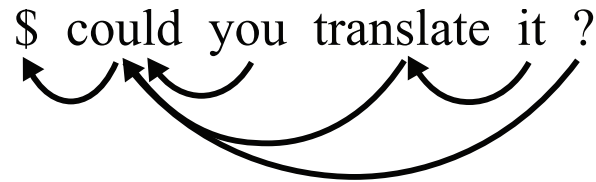
Unsupervised Part-of-Speech Tagging

sentences are observed, part-of-speech tags are treated as latent variables:



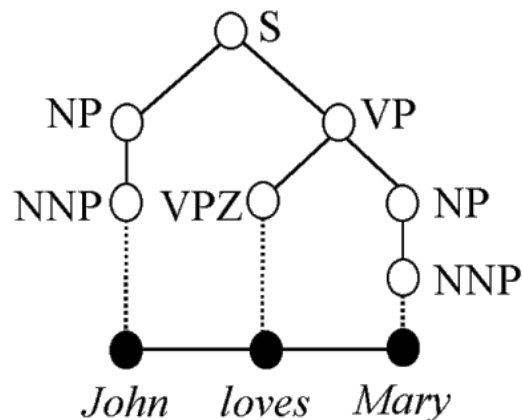
Unsupervised Dependency Parsing

sentences are observed, dependency parse trees are treated as latent variables:



Latent Syntactic Categories for Parsing

- split Penn Treebank syntactic categories into finer subcategories



NNP

NNP-0	Jr.	Goldman	INC.
NNP-1	Bush	Noriega	Peters
NNP-2	J.	E.	L.
NNP-3	York	Francisco	Street
NNP-4	Inc	Exchange	Co
NNP-5	Inc.	Corp.	Co.

RB

RB-0	recently	previously	still
RB-1	here	back	now
RB-2	very	highly	relatively
RB-3	so	too	as
RB-4	also	now	still
RB-5	however	Now	However

Petrov (2009)

Morphological Segmentation

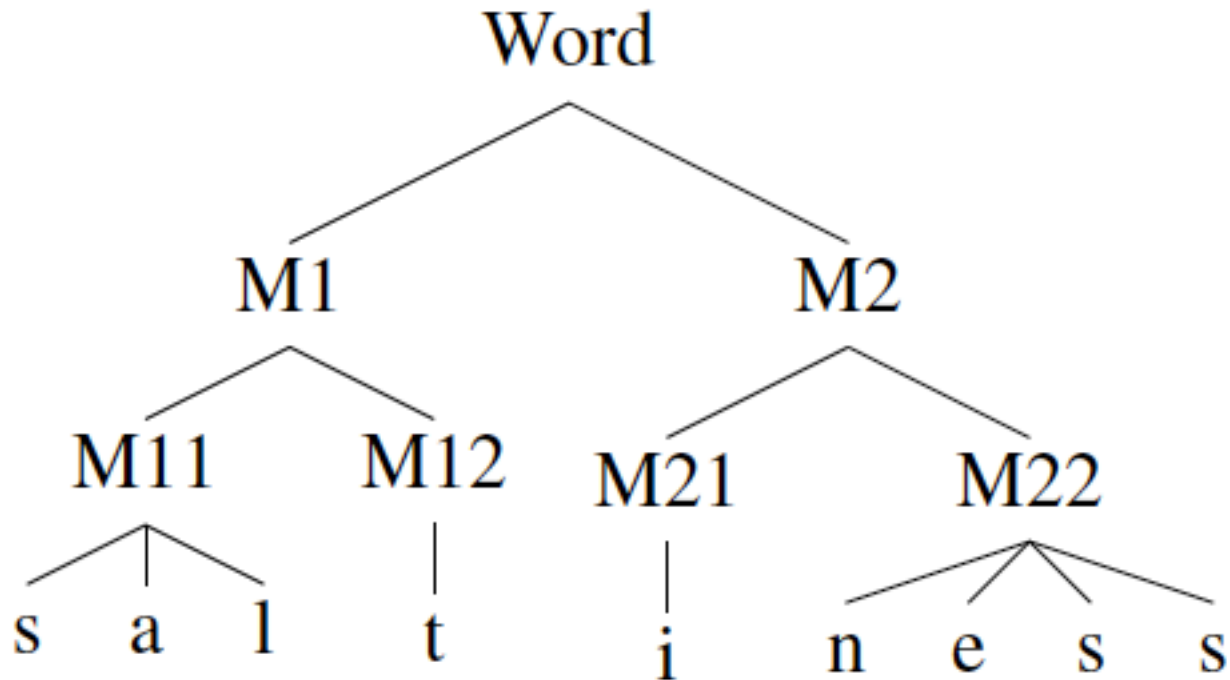


Figure 1: The parse tree generated by the metagrammar of depth 2 for the word *saltiness*.

Morphological Segmentations, POS, and Syntactic Trees



Snyder & Barzilay

Generative Stories

- we hypothesize latent variables through which data are generated
- define “generative story” that describes how latent variables are generated, then how data is generated using latent variables
- we parameterize the distributions & add parameter generation to generative story

Generative Story Template

- 1: Draw a set of parameters θ from $p(\Theta)$
- 2: Draw a latent structure z from $p(Z | \theta)$
- 3: Draw the observed data x from $p(X | z, \theta)$

$$p(x, z, \theta) = p(\theta)p(z | \theta)p(x | z, \theta)$$

Inference

- in general, inference roughly means “calculate statistical quantities of interest”
- examples:
 - compute the mode of some random variables when conditioning on some and marginalizing out others
 - compute marginals of some random variables (variable posteriors when marginalizing out everything else)
 - compute posterior distribution over some subset of random variables

Learning?

- in Bayesian NLP, there's often no "learning"
- there is only "inference"
- just define model and do inference to calculate what we want to calculate
 - no parameters are being estimated from data*
 - we are not optimizing any loss function*
 - there is no gradient descent*

* typically

Generative Story Template

- 1: Draw a set of parameters θ from $p(\theta \mid \alpha)$
- 2: Draw a latent structure z from $p(z \mid \theta)$
- 3: Draw the observed data x from $p(x \mid z, \theta)$

$$p(x, z, \theta \mid \alpha) = p(\theta \mid \alpha) p(z \mid \theta) p(x \mid z, \theta)$$

Our data is a set of samples: $x^{(1)}, x^{(2)}, \dots, x^{(n)}$

$$\begin{aligned} \text{joint: } & p(x^{(1)}, \dots, x^{(n)}, z^{(1)}, \dots, z^{(n)}, \theta \mid \alpha) \\ & = p(\theta \mid \alpha) \left(\prod_{i=1}^n p(z^{(i)} \mid \theta) p(x^{(i)} \mid z^{(i)}, \theta) \right) \end{aligned}$$

Key Quantities

$$p(x, z, \theta \mid \alpha) = p(\theta \mid \alpha) p(z \mid \theta) p(x \mid z, \theta)$$

Our data is a set of samples: $x^{(1)}, x^{(2)}, \dots, x^{(n)}$

$$\begin{aligned} \text{joint: } & p(x^{(1)}, \dots, x^{(n)}, z^{(1)}, \dots, z^{(n)}, \theta \mid \alpha) \\ & = p(\theta \mid \alpha) \left(\prod_{i=1}^n p(z^{(i)} \mid \theta) p(x^{(i)} \mid z^{(i)}, \theta) \right) \end{aligned}$$

$$\text{posterior: } p(z^{(1)}, \dots, z^{(n)}, \theta \mid x^{(1)}, \dots, x^{(n)}, \alpha)$$

$$\text{collapsed posterior: } p(z^{(1)}, \dots, z^{(n)} \mid x^{(1)}, \dots, x^{(n)}, \alpha)$$

Markov Chain Monte Carlo (MCMC)

- MCMC algorithms are widely used in Bayesian modeling but also useful more generally
- can be used to generate samples from distributions that are hard to sample from
- samples can be used to estimate quantities of interest
- these estimates are unbiased

Gibbs Sampling

- Gibbs sampling is the simplest and most widely-used MCMC algorithm (at least in NLP)

Gibbs Sampling Template

$U_1, \dots, U_p =$ latent variables

$U_{-i} =$ all latent variables other than U_i

$\mathbf{X} =$ all observed data and hyperparameters

Gibbs sampling:

initialize all U_i to values u_i

repeat until convergence:

sample u from $p(U_i \mid u_{-i}, \mathbf{X})$

set $U_i \leftarrow u$

Gibbs Sampling Template

Gibbs sampling:

initialize all U_i to values u_i

repeat until convergence:

sample u from $p(U_i | u_{-i}, \mathbf{X})$

set $U_i \leftarrow u$

At convergence, each time we update any value of any random variable in U_1, \dots, U_p , we have another sample from the posterior

these samples can be used to estimate any quantity of interest while offering some nice theoretical properties

Disadvantages of Gibbs Sampling?

Gibbs sampling:

initialize all U_i to values u_i

repeat until convergence:

sample u from $p(U_i | u_{-i}, \mathbf{X})$

set $U_i \leftarrow u$

nearby samples are not necessarily uncorrelated, so it can take many samples for good estimates, especially of rare events

guarantees are at convergence, “burn-in” time can be hard to estimate & depends on initialization

LDA Inference Cheat Sheet

- 1: For each topic, draw a multinomial word distribution $\beta_k \sim \text{Dirichlet}(\psi)$
- 2: For each document
 - a: Draw a multinomial topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
 - b: For each position in document
 - i: Draw a topic $z \sim \text{Multinomial}(\theta)$
 - ii: Draw a word $w \sim \text{Multinomial}(\beta_z)$

$$a \sim \text{Multinomial}(\theta)$$

$$p(A = a_i | \theta) = \theta_i$$

$$\theta \sim \text{Dirichlet}(\alpha)$$

$$p(\Theta = \theta | \alpha) = \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i - 1}$$

Gibbs sampling:

initialize all U_i to values u_i

repeat until convergence:

sample u from $p(U_i | u_{-i}, \mathbf{X})$

set $U_i \leftarrow u$

$K = \#$ topics

$N = \#$ documents

$M = \#$ words in each document

$V = \#$ words in vocabulary