4CSLL5 'Advanced Computational Linguistics' Machine Learning Techniques in Machine Translation, Speech Recognition and Topic Modelling

Martin Emms

March 25, 2014

Unsupervised Learning from Data

Machine Translation

Unsupervised Learning from Data

Machine Translation

Speech Recognition

Unsupervised Learning from Data

Machine Translation

Speech Recognition

Topic Modelling

Unsupervised Learning from Data

Machine Translation

Speech Recognition

Topic Modelling

Use Unsupervised Machine Learning

Unsupervised Learning from Data

Machine Translation

Speech Recognition

Topic Modelling

Use Unsupervised Machine Learning

Possibly other attempts to infer from Big Data will use **Unsupervised Machine Learning?**

In Machine Translation

Unsupervised methods are used to go from sentence pairs to word pairs

In Machine Translation

Unsupervised methods are used to go from sentence pairs to word pairs

la maison est grand	the house is big
c'est un haricot vert	its a green bean
je le lui donne	I give it to him
⋮	⋮
⋮	⋮

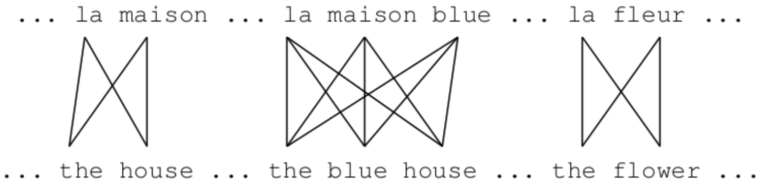
$\Rightarrow$

$p(la\|the)$	0.453
$p(le\|the)$	0.334
$p(maison\|house)$	0.876
$p(bleu\|blue)$	0.563

Done using EM Algorithm

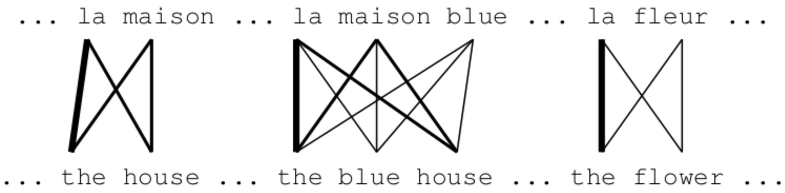
- Incomplete data
  - if we had *complete data*, would could estimate *model*
  - if we had *model*, we could fill in the *gaps in the data*
- Expectation Maximization (EM) in a nutshell
  1. initialize model parameters (e.g. uniform)
  2. assign probabilities to the missing data
  3. estimate model parameters from completed data
  4. iterate steps 2–3 until convergence

Learning Word Translations



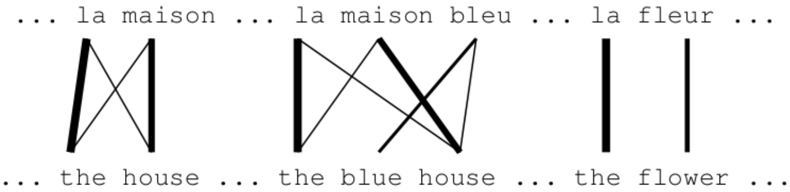
- Initial step: all alignments equally likely

Learning Word Translations



- After one iteration
- Alignments, e.g., between la and the are more likely

Learning Word Translations



- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

Learning Word Translations

- eventually converges, inherent hidden structure found by EM
- **Learns** translation probabilities

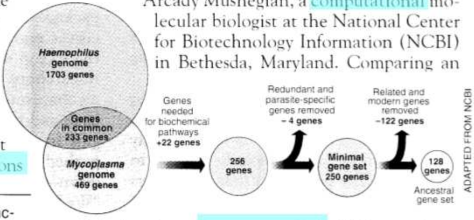
$p(la \mid the)$	0.453
$p(le \mid the)$	0.334
$p(maison \mid house)$	0.876
$p(bleu \mid blue)$	0.563

Topics in 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, who 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

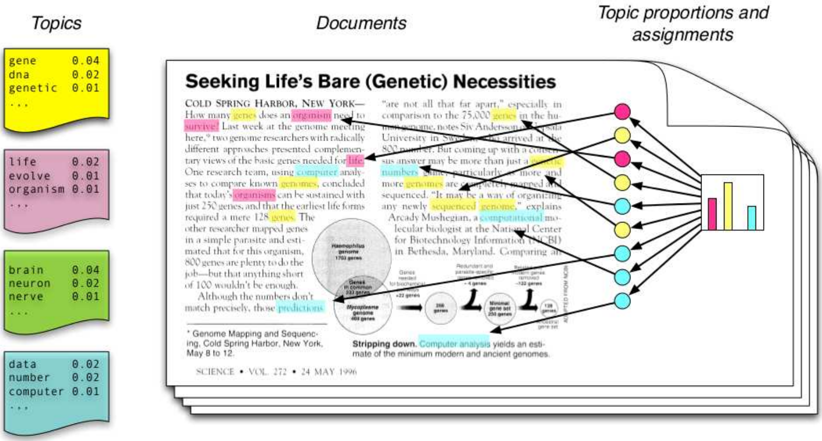


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

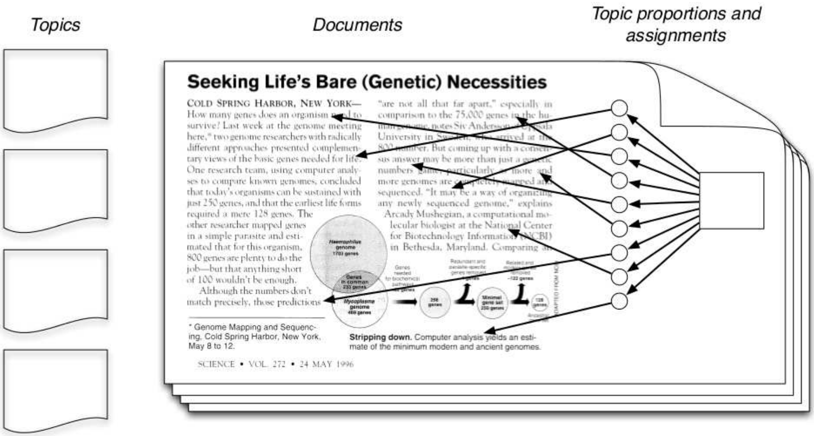
Documents exhibit multiple *topics* **biology** **genetics** **computation** . . .

Topics in Documents



- **topic** = distribution over words
- **document** = distrubution over topics

Topics in Documents



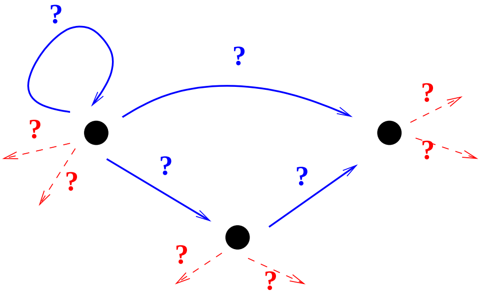
- can only directly see words
- but can learn the rest by **unsupervised methods**

Example learned words-for-topics

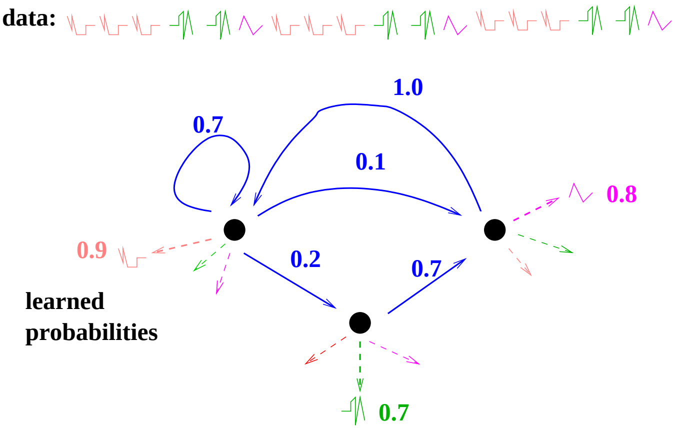
GENETICS	EVOL. BIOL.	COMPUTING
human	evolution	computer
genome	evolutionary	models
dna	species	information
genetic	organisms	data
genes	life	computers
sequence	origin	system
gene	biology	network
molecular	groups	systems
sequencing	phylogenetic	model
map	living	parallel
information	diversity	methods
genetics	group	networks
mapping	new	software
project	two	new
sequences	common	simulations

Hidden Markov Models for Speech or Actions

- words/action visible as an *evidence* sequence
- can be modeled with a HMM
  with transitions between states and visible evidence for a state



Learning HMM from Data



- Possible to use **unsupervised learning** to find probabilities concerning *hidden* variables from data with just *visible* evidence
- Used in Speech Recognizers, Activity Recognizers