POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Text Mining 5
Information Extraction

Madrid Summer School on
Advanced Statistics and Data Mining

Florian Leitner
Data Catalytics, S.L.
leitner@datacatytics.com

Retrospective

We have seen how to…

- Design generative **models** of natural **language**.
- **Segment**, **tokenize**, and **compare** text and/or sentences.
- [Multi-] **labeling** whole documents or chunks of text.

Some open questions…

- **How to assign labels to individual tokens in a stream?**
  - without using a dictionary/gazetteer: **sequence taggers**
- How to detect **semantic relationships** between tokens?
  - **dependency parsers**

Incentive and applications

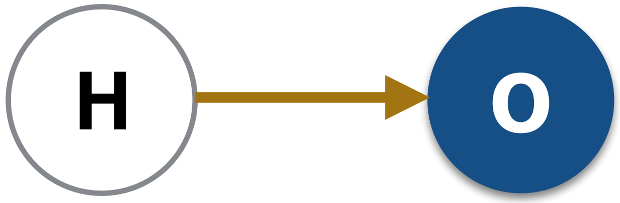
➡ Statistical analyses of [token] **sequences**

- Part-of-Speech (PoS) tagging & chunking
  - noun, verb, adjective, … & noun/verb/preposition/… phrases
- Named Entity Recognition (NER)
  - organizations, persons, places, genes, chemicals, …
- Information extraction
  - locations/times, phys. constants & chem. formulas, entity linking, event extraction, entity-relationship extraction, …

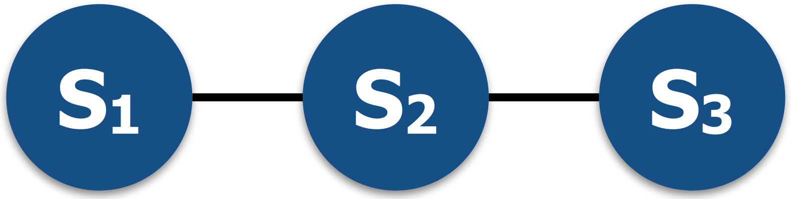
Probabilistic graphical models

- Graphs of hidden (blank) and **observed** (shaded) **variables** (vertices/nodes).
- The edges depict **dependencies**, and if directed, show [ideally causal] **relationships** between nodes.

directed ➡ Bayesian Network (BN)



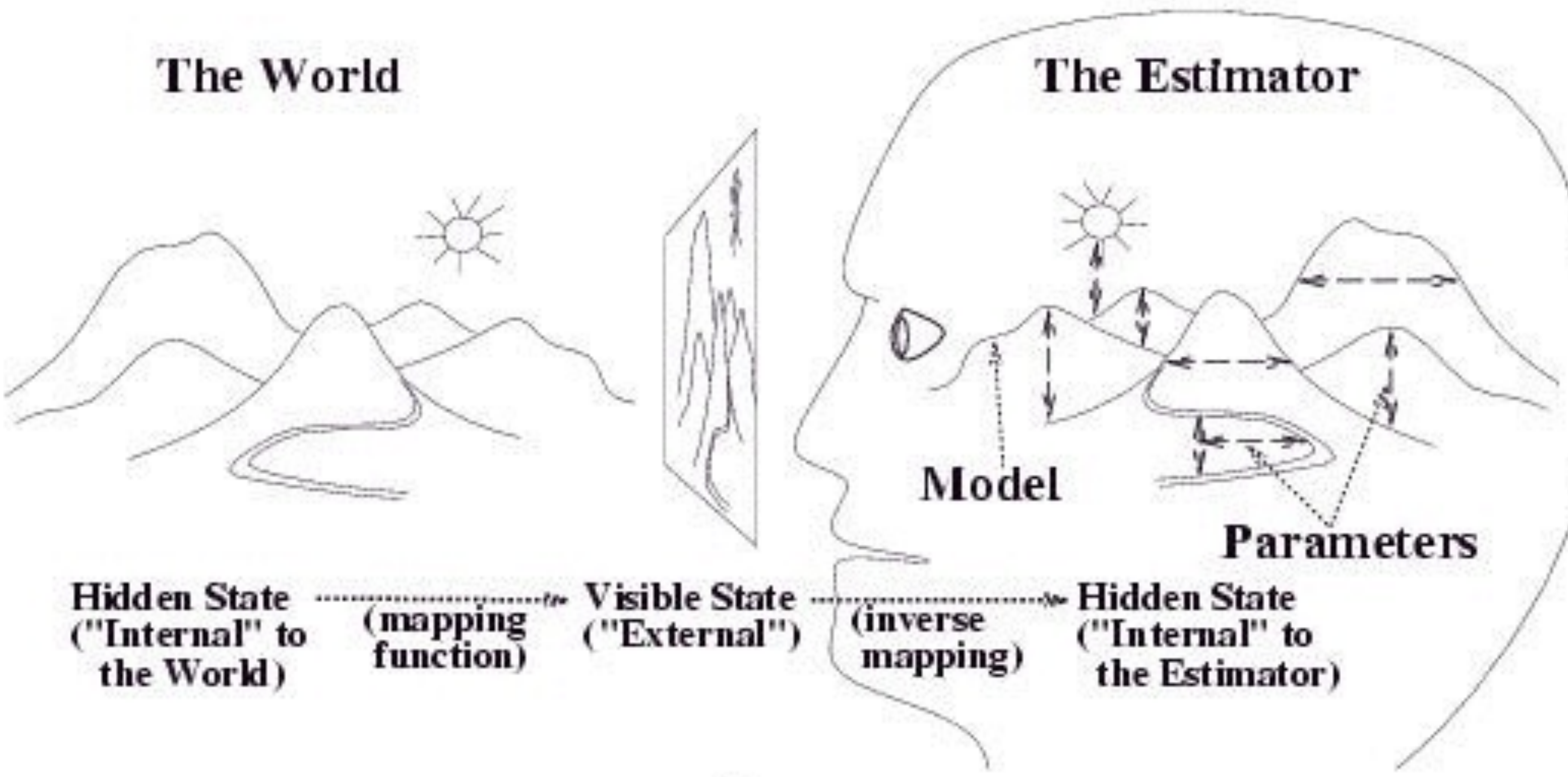
undirected ➡ Markov Random Field (MRF)



mixed ➡ Mixture Models

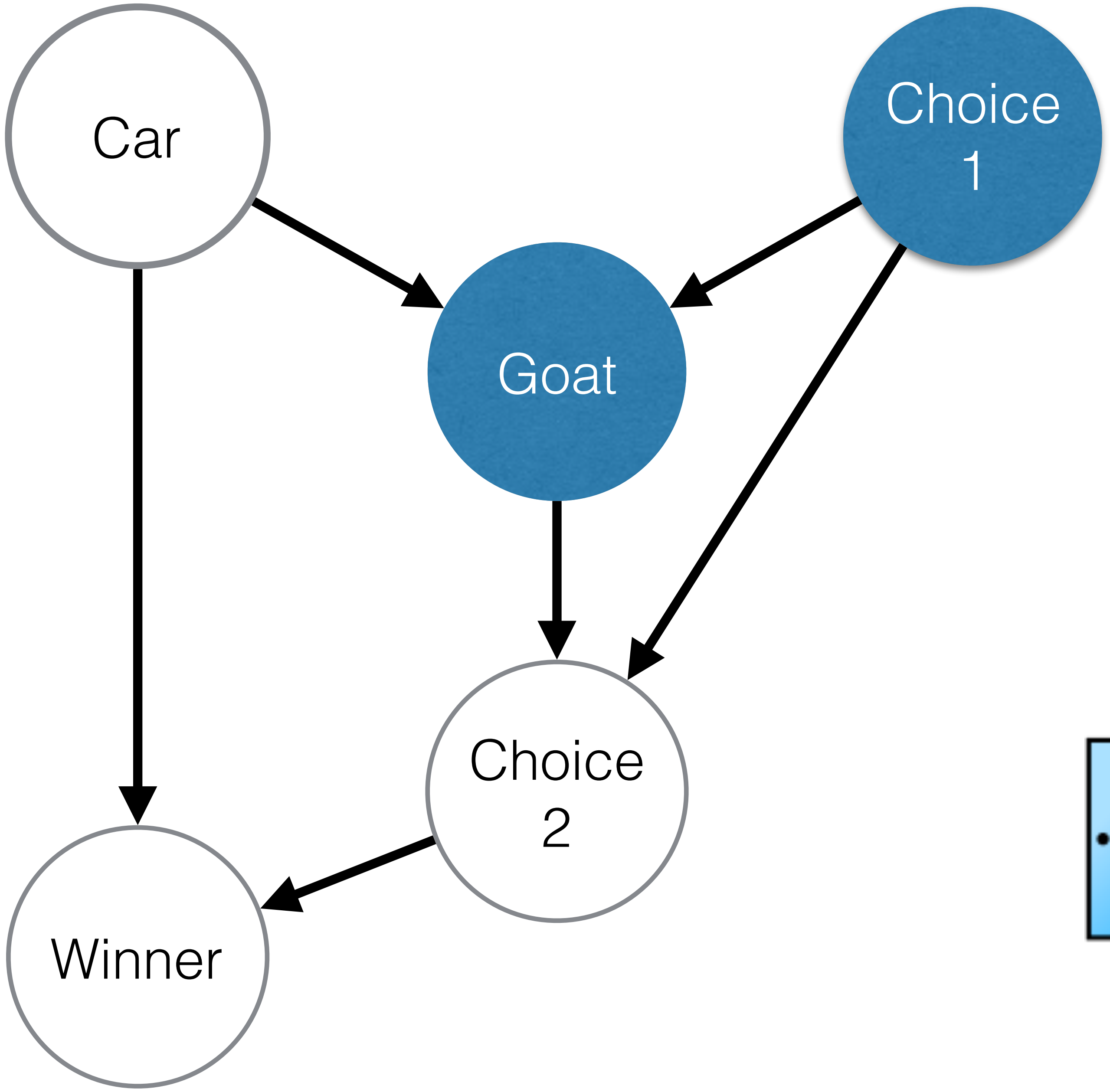
Koller & Friedman. Probabilistic Graphical Models. 2009

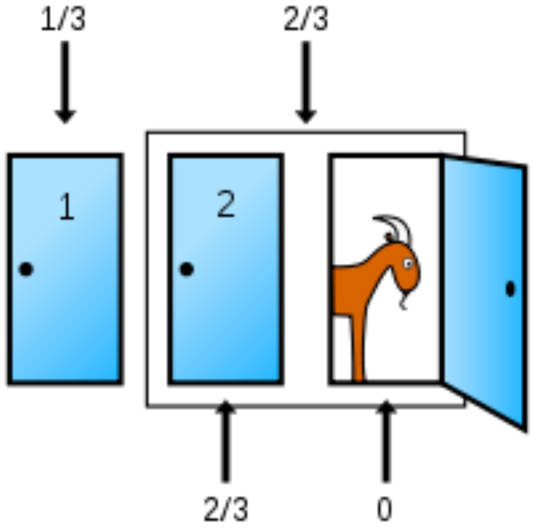
Hidden vs. observed state and statistical parameters


The World
The Estimator
Model
Parameters
Hidden State ("Internal" to the World)
(mapping function)
Visible State ("External")
(inverse mapping)
Hidden State ("Internal" to the Estimator)

Rao. A Kalman Filter Model of the Visual Cortex. Neural Computation 1997

A Bayesian network for the Monty Hall problem





Markov random field

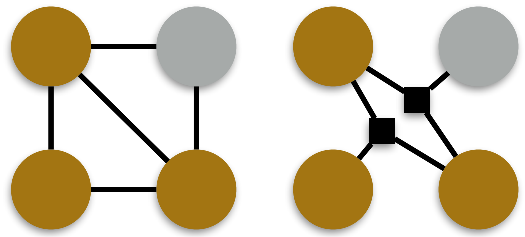
History class: Ising developed a linear field to model (binary) atomic spin states (Ising, 1924); the 2-dim. model problem then was solved by Onsager in 1944.

factor (clique potential)

$$P(X=\vec{x}) = \frac{\prod_{cl\in\vec{x}} \phi_{cl}(cl)}{\sum_{\vec{x}\in X}\prod_{cl\in\vec{x}} \phi_{cl}(cl)}$$

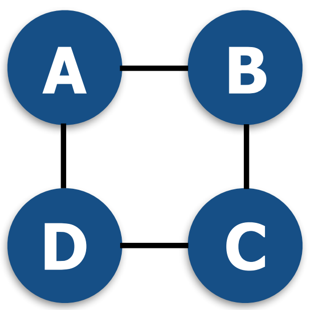
normalizing constant (partition function **Z**)

cl … [**maximal**] **clique**; a subset of factors in the graph where every pair is connected



ϕ(A, B)			ϕ(B, C)			ϕ(C, D)			ϕ(D, A)		
a_0	b_0	30	b_0	c_0	100	c_0	d_0	1	d_0	a_0	100
a_0	b_1	5	b_1	c_0	1	c_0	d_1	100	d_1	a_0	1
a_1	b_0	1	b_0	c_1	1	c_1	d_0	100	d_0	a_1	1
a_1	b_1	10	b_1	c_1	100	c_1	d_1	1	d_1	a_1	100

factor table



factor graph

$P(a_1, b_1, c_0, d_1) =$
$10\cdot1\cdot100\cdot100 \div$
$7'201'840 =$
0.014

Probabilistic models for sequential data

- Origin: Kálmán filters (L[in.]Q[uad.]E[estim.]; Kálmán 1960)
- a.k.a. **Temporal** or **dynamic** **Bayesian networks**
  - **Static** process w/ **constant** model ➡ **temporal** process w/ **dynamic** model
  - Model structure and parameters are [still] **constant**
  - The model topology within a [constant] “**time slice**” is depicted

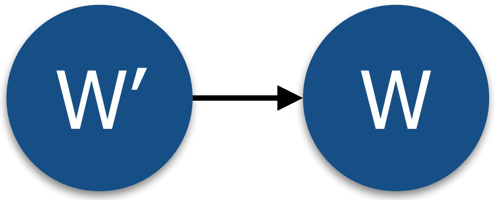
generative

- **Markov** Chain (MC; Markov. 1906)
- Hidden **Markov** Model (HMM; Baum et al. 1970)

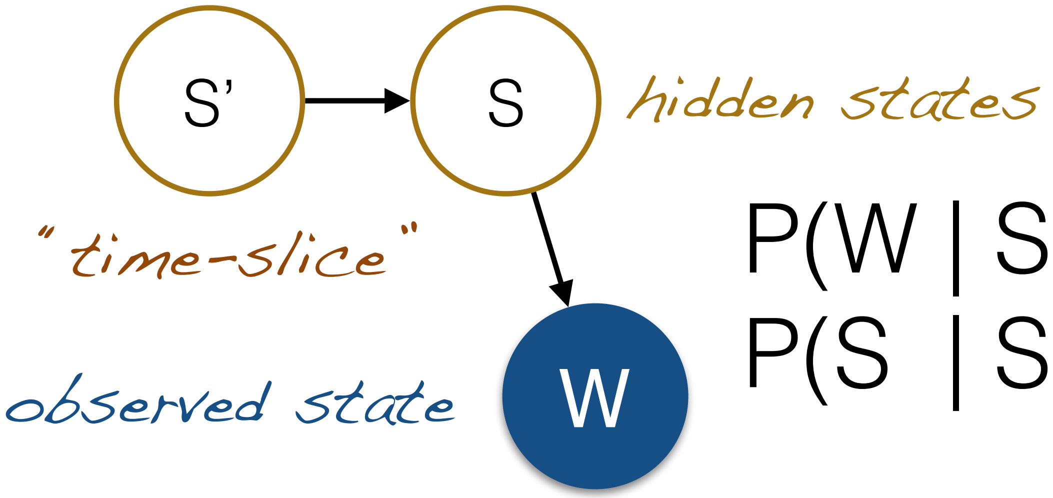
discriminative

- MaxEnt **Markov** Model (MEMM; McCallum et al. 2000)
- [**Markov**] Conditional Random Field (CRF; Lafferty et al. 2001)
  - Naming: all four models make the **Markov assumption** (see part 2)

From a Markov chain to a Hidden Markov Model (HMM)

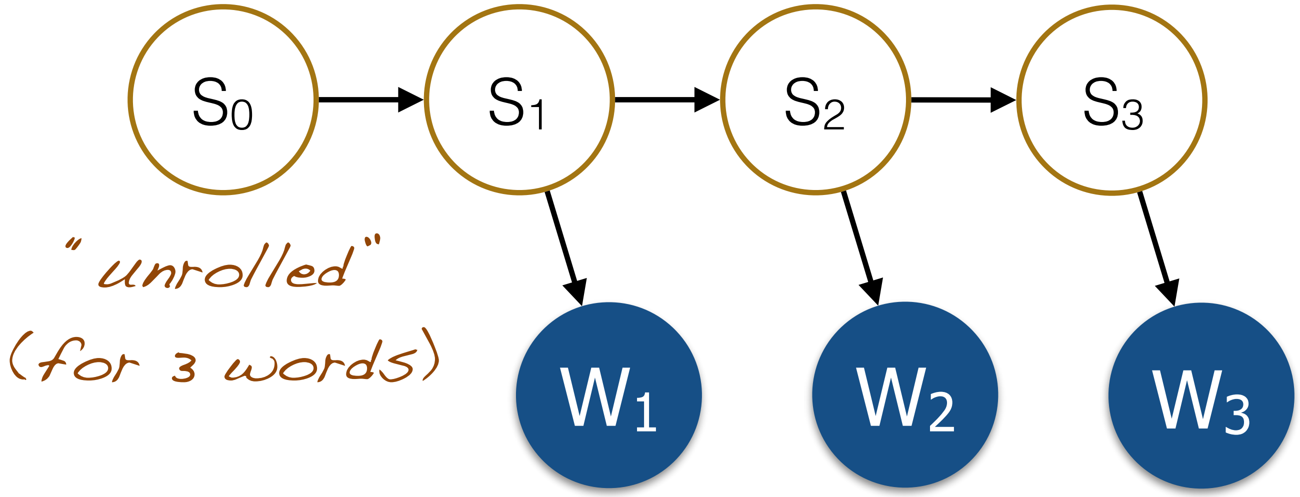


$P(W \mid W')$



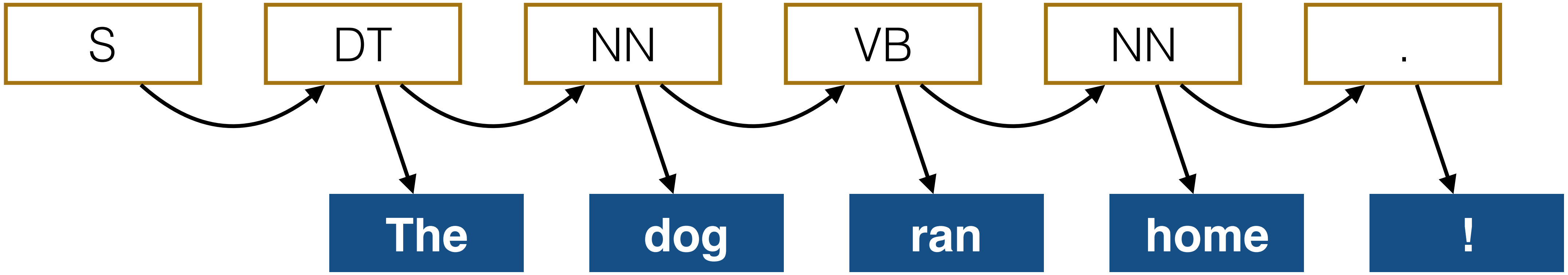
$P(W \mid S)$
$P(S \mid S')$

W depends on S
and S in turn
depends on S'



A language-based intuition for HMMs

- A Markov Chain: $P(W) = \prod P(w \mid w')$
  - assumes the observed words are in and of themselves the cause of the observed sequence.
- A HMM: $P(S, W) = \prod P(s \mid s') P(w \mid s)$
  - assumes the observed words are emitted by a hidden (not observable) sequence, for example the chain of part-of-speech-states.



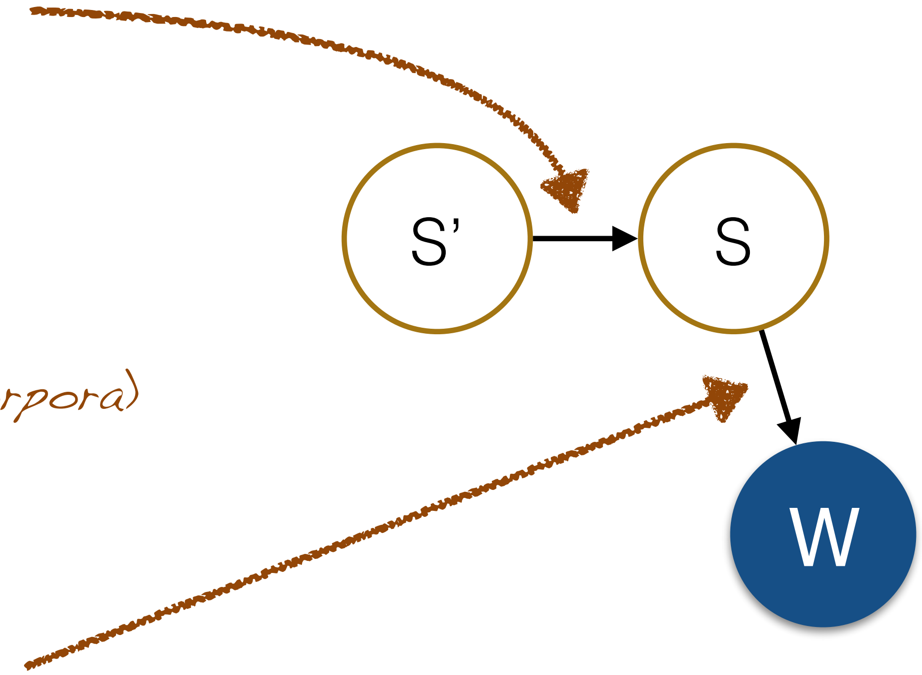
again, this is the "unrolled" model that does not depict the conditional dependencies

The two matrices of a HMM

a.k.a. "CPDs": Conditional Probability Distributions

P(s I s')	DT	NN	VB	…
DT	0,03	0,7	0	
NN	0	0	0,5	
VB	0	0,5	0,2	
…				

Transition Matrix



(measured as discrete factor tables from annotated PoS corpora)

P(w I s)	word$_1$	word$_2$	word$_3$	…
DT	0,3	0	0	
NN	0,0001	0,002	0	
VB	0	0	0,001	
…				

underflow danger ➡ use "log Ps"!

Observation Matrix

very sparse (W is large)

➡ Smoothing!

Three tasks solved by HMMs

Evaluation: Given a HMM, **infer** the P of the observed sequence (because a HMM is a generative model).

in Bioinformatics: Likelihood of a particular DNA element

Solution: **Forward Algorithm**

Decoding: Given a HMM and an observed sequence, **predict** the hidden states that lead to this observation.

in Statistical NLP: PoS annotation

Solution: **Viterbi Algorithm**

Training: Given only the graphical model and an observation sequence, **learn** the best [smoothed] parameters.

Solution: **Baum-Welch Algorithm**

all three algorithms are are implemented using dynamic programming

Three limitations of HMMs

“unsolved”

(CRF)

Markov assumption: The next state only depends on the current state.

Example issue: trigrams *(long-range dependencies!)*

MEMM

Output assumption: The output (observed value) is independent of all previous outputs (given the current state).

Example issue: word morphology *(inflection, declension!)*

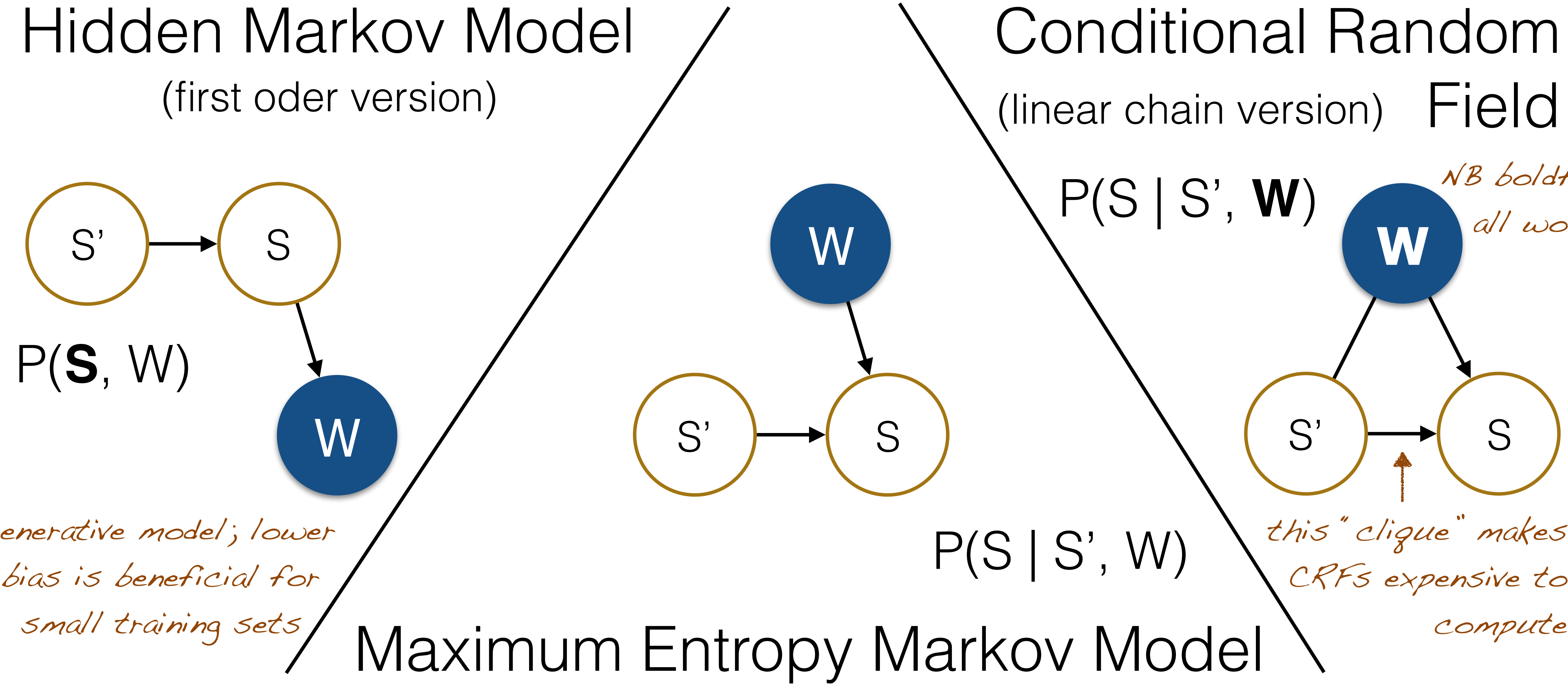
CRF

Stationary assumption: Transition probabilities are independent of the actual time when they take place.

Example issue: position in sentence

(label bias problem, see next!)

From generative to discriminative Markov models



Hidden Markov Model
(first oder version)

P(**S**, W)

generative model; lower bias is beneficial for small training sets

Maximum Entropy Markov Model
(first oder version)

P(S | S', W)

Conditional Random Field
(linear chain version)

P(S | S', **W**)

NB boldface W: all words!

this "clique" makes CRFs expensive to compute

Maximum entropy (MaxEnt 2/2) [again]

REMINDER

- In summary, MaxEnt is about selecting the “maximal” model p*:

$$p^* = \underset{p \in P}{argmax} \; - \sum_{x \in X} p(x) \; \sum_{y \in Y} p(y|x) \; log_2 \; p(y|x)$$

select some model that maximizes the conditional entropy...

- That obeys the following conditional equality constraint:

$$\sum_{x \in X} P(x) \; \sum_{y \in Y} P(y|x) \; f(x,y) = \sum_{x \in X, y \in Y} P(x,y) \; f(x,y)$$

...using a conditional model that matches the (observed) joint probabilities

- Next: Using, e.g., **Langrange multipliers**, one can establish the optimal λ parameters of the model that maximize the entropy of this probability:

$$p^*(y|X) = \frac{exp(\sum \lambda_i f_i(X,y))}{\sum_{y \in Y} exp(\sum \lambda_i f_i(X,y))}$$

“ Exponential Model ”

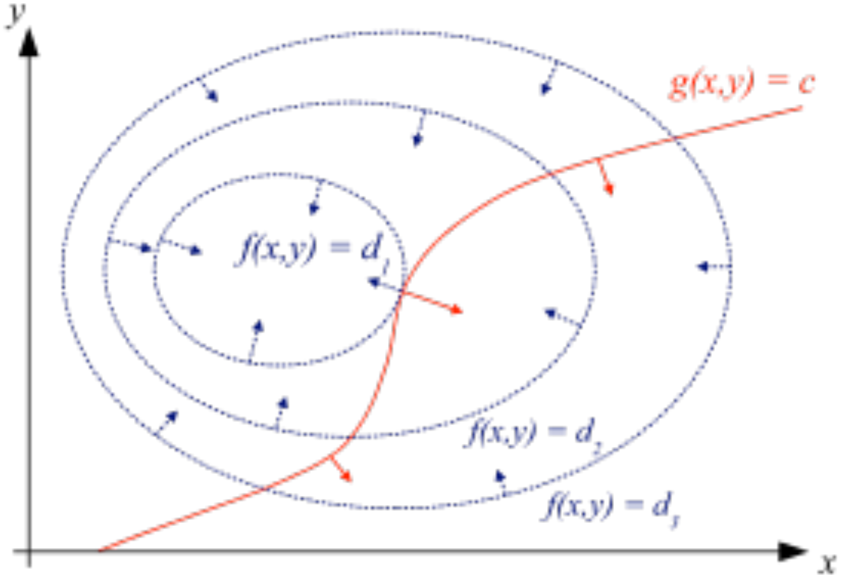
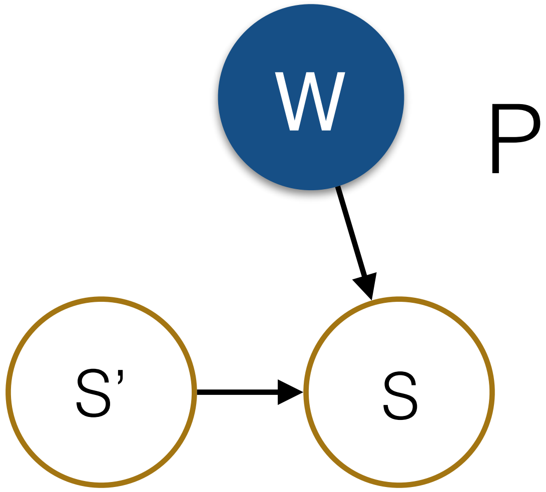


Image Source: WikiMedia Commons, Nexcis

Maximum Entropy Markov Models (MEMM)

Simplify training of P(s | s’, w) by splitting the model into |S| separate transition functions $P_{s'}(s | w)$ for each s’



P(S | S’, W)

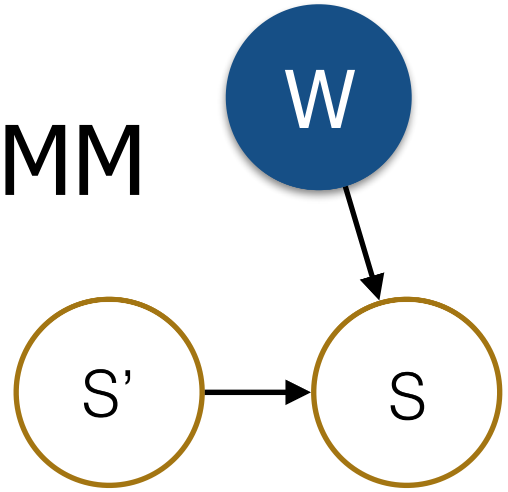
$$P(s | s', w) = P_{s'}(s | w)$$

$$P_{s'}(s|w) = \frac{exp\left(\sum \lambda_i f_i(w, s)\right)}{\sum_{s^* \in S} exp\left(\sum \lambda_i f_i(w, s^*)\right)}$$

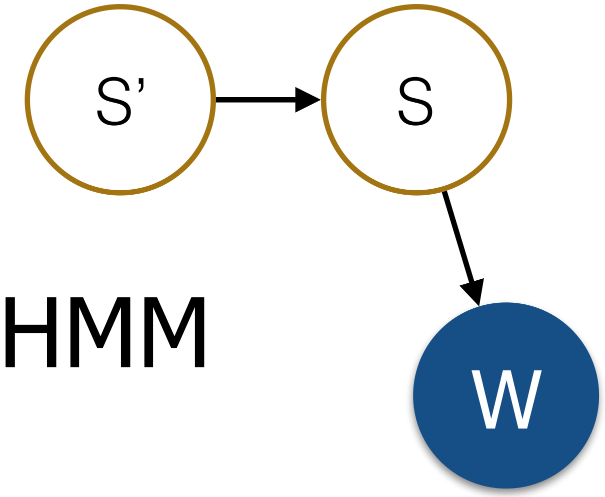
(The MaxEnt slide had {x, y} which here are {w, s}.)

The **label bias problem** of directional Markov models

MEMM



because of their directionality constraints, MEMMs & HMMs suffer from the label bias problem



HMM

The robot wheels Fred around.

DT **NN** **VB** NN RB

The robot wheels were broken.

DT **NN** **NN** VB JJ

The robot wheels are round.

DT **NN** **??** *yet unseen!*

Wallach. Efficient Training of CRFs. MSc 2002

Conditional Random Field

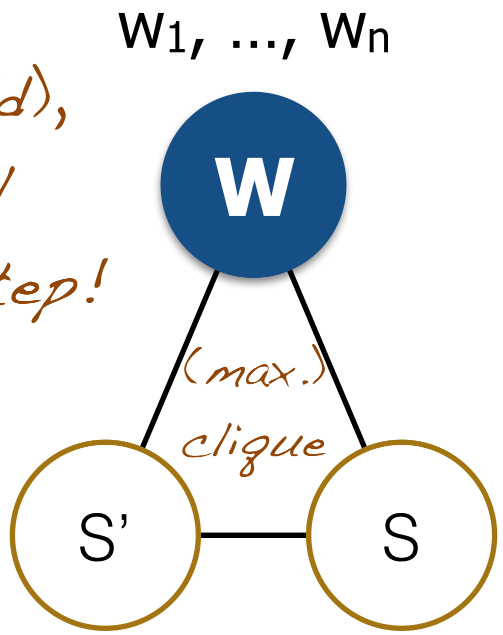
MRF:

$$P(X = \vec{x}) = \frac{\prod_{cl \in \vec{x}} \phi_{cl}(cl)}{\sum_{\vec{x} \in X} \prod_{cl \in \vec{x}} \phi_{cl}(cl)}$$

CRF:

$$P(Y = \vec{y}\, | X = \vec{x}) = \frac{\prod_{y \in \vec{y}} \phi_{cl}(y', y, \vec{x})}{\sum_{\vec{y} \in Y} \prod_{y \in \vec{y}} \phi_{cl}(y', y, \vec{x})}$$

note W (upper-case/bold), not w (lower-case): all words are used in each step!



The label bias problem is "solved" by conditioning the MRF Y-Y' on the entire observed sequence.

$$P(S|W) = \frac{exp\left(\sum \lambda_i f_i(W, s', s)\right)}{\sum_{s^* \in S} exp\left(\sum \lambda_i f_i(W, s'^*, s^*)\right)}$$

Models a per-state **exponential function** of joint probability over the **entire** observed **sequence** W.

Wallach. Conditional Random Fields: An introduction. TR 2004

Parameter estimation and L2 regularization of CRFs

(regularization reduces the effects of overfitting)

- For training $\{Y^{(n)}, X^{(n)}\}^{N}_{n=1}$ sequence pairs with K features
- Parameter estimation using **conditional log-likelihood**

$$\lambda = \underset{\lambda \in \Lambda}{argmax} \sum_{n}^{N} log\ P(Y^{(n)} | X^{(n)}; \lambda)$$

- Substitute log $P(Y^{(n)} \mid X^{(n)})$ with log **exponential model**

normalizing constant (partial function Z)

$$\ell(\lambda) = \sum_{n=1}^{N} \sum_{y \in Y^{(n)}} \sum_{i=1}^{K} \lambda_i f_i(y', y, X^{(n)}) - \sum_{n=1}^{N} log\ Z(X^{(n)})$$

- Add a penalty for parameters with a to high **L2-norm**

L2: Euclidian norm

(∑ of squared errors)

$$\ell(\lambda) = \sum_{n=1}^{N} \sum_{y \in Y^{(n)}} \sum_{i=1}^{K} \lambda_i f_i(y', y, X^{(n)}) - \sum_{n=1}^{N} log\ Z(X^{(n)}) - \sum_{i=1}^{K} \frac{\lambda_i^2}{2\sigma^2}$$

$$\frac{\|\Lambda\|_2^2}{2\sigma^2}$$

free regularization parameter

Model summaries: HMM, MEMM, CRF

- A **HMM**
  - **generative** model
  - **efficient** to learn and deploy
  - trains with **little data**
  - generalizes well (**low bias**)
- A **MEMM**
  - better labeling **performance**
  - modeling of **features**
  - **label bias** problem
- **CRF**
  - conditioned on **entire observation**
  - **complex features** over full input
  - training time **scales exponentially**
    - $O(NTM^2G)$
    - N: # of sequence pairs;
      T: E[sequence length];
      M: # of (hidden) states;
      G: # of gradient computations for parameter estimation

a PoS model w/45 states and 1 M words can take a week to train...

Sutton & McCallum. An Introduction to CRFs for Relational Learning. 2006

Information extraction: Application of dynamic graphical models

The Parts of Speech

I	ate	the	pizza	with	green	peppers	.
PRP	VB	DT	NN	IN	JJ	NN	.

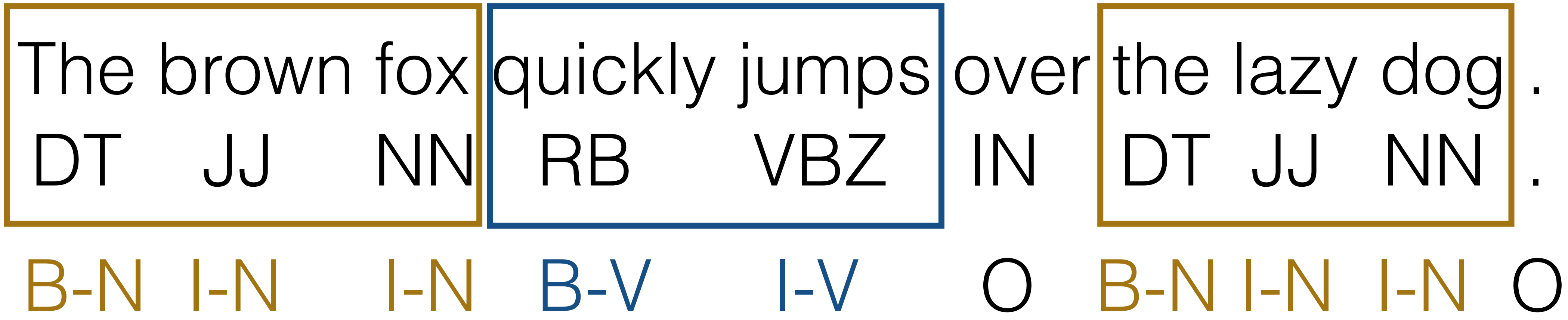
- **Corpora** for the [supervised] **training** of PoS taggers
  - **Brown** Corpus (AE from ~1961)
  - British National Corpus: **BNC** (20th century British)
  - Wall Street Journal Corpus: **WSJ Corpus** (AE from the 80s)
  - American National Corpus: **ANC** (AE from the 90s)
  - Lancaster Corpus of Mandarin Chinese: **LCMC** (Books in Mandarin)
  - The **GENIA** corpus (Biomedical abstracts from PubMed)
  - **NEGR@** (German Newswire from the 90s)
  - Spanish and Arabian corpora should be (commercially…) available… ???

Best tip: Ask on the "corpora list" mailing list!

Noun and verb phrase chunking with BIO-encoded labels

“shallow parsing”

a pangram (hint: check the letters)



The	brown	fox	quickly	jumps	over	the	lazy	dog	.
DT	JJ	NN	RB	VBZ	IN	DT	JJ	NN	.
B-N	I-N	I-N	B-V	I-V	O	B-N	I-N	I-N	O

Performance (2nd order CRF) ~ 94%

Main problem: embedded & chained NPs (N of N and N)

Chunking is “more robust to the highly diverse corpus of text on the Web” and [exponentially] faster than [deep] parsing.

Banko et al. Open Information Extraction from the Web. IJCAI 2007 a paper with over 700 citations

Wermter et al. Recognizing noun phrases in biomedical text. SMBM 2005 error sources

Word Sense Disambiguation (WSD)

Note the PoS-tag "dependency": otherwise, the two examples would have even more senses!

- Basic Example: **hard** [JJ]
  - physically hard (a hard stone)
  - difficult [task] (a hard task)
  - strong/severe (a hard wind)
  - dispassionate [personality] (a hard bargainer)
- Entity Disambig.: **bank** [NN]
  - finance ("bank account")
  - terrain ("river bank")
  - aeronautics ("went into bank")
  - grouping ("a bank of …")

- **SensEval**
  - http://www.senseval.org/
  - SensEval/SemEval Challenges
  - Provides corpora where every word is tagged with its sense
- **WordNet**
  - http://wordnet.princeton.edu/
  - A labeled graph of word senses
- **Applications**
  - Named entity recognition
  - Machine translation
  - Language understanding

Word vector representations: Unsupervised WSD

- Idea 1: Words with similar meaning have similar **environments**.
- Use a word vector to count a word's **surrounding words**.
- Similar words now will have similar word vectors.
  - See lecture 2, neural network models of language and lecture 4, cosine similarity
  - Visualization: Principal Component Analysis

Turian et al. Word representations. ACL 2010

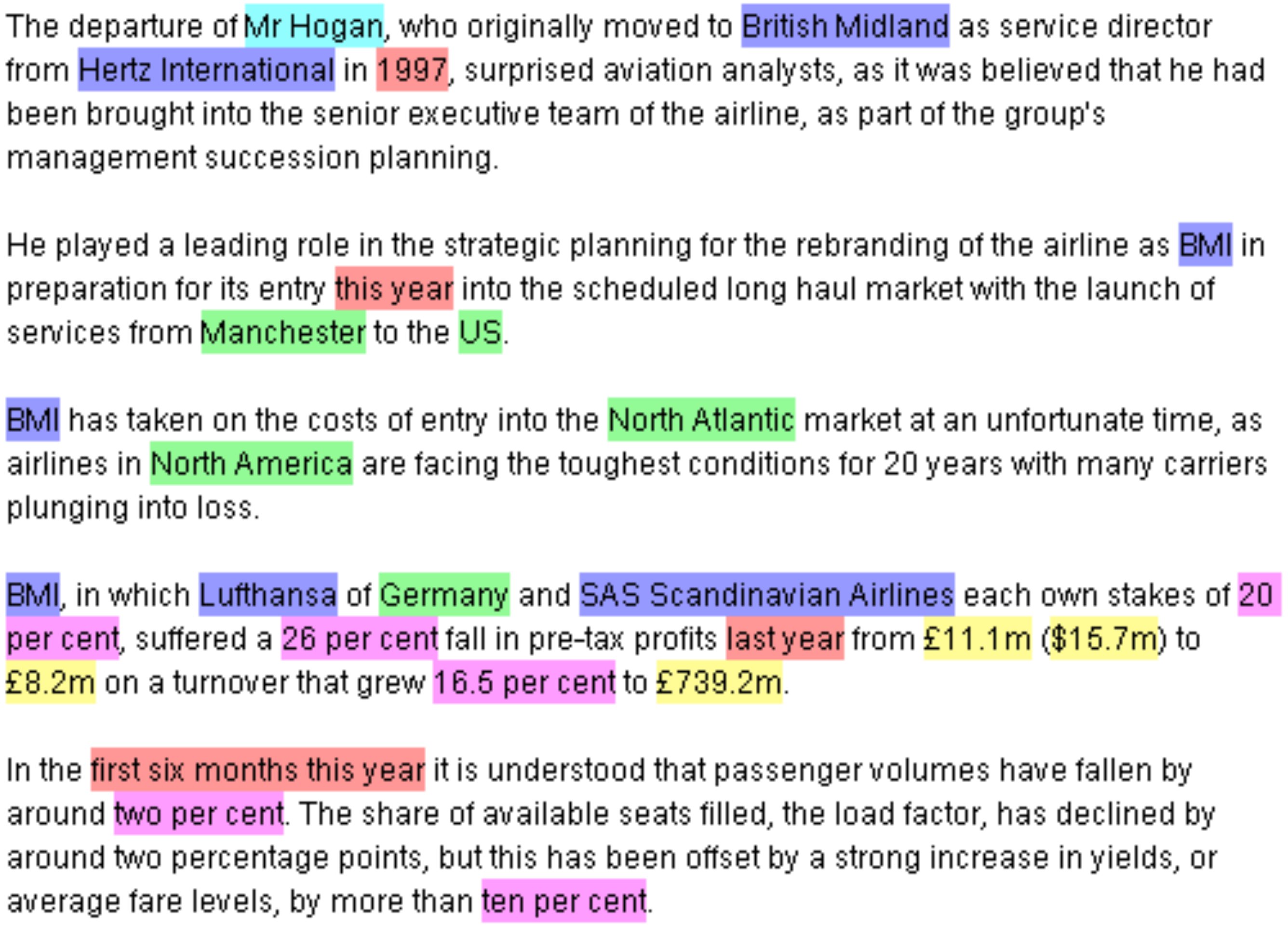
- Idea 2: Words with similar meaning have similar **environments**.
- Use the surrounding of **unseen words** to "smoothen" language models (i.e., the correlation between word w_i and its context c_j).
  - see Text Mining 4: TF-IDF weighting, Cosine similarity and point-wise MI
  - Levy & Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. CoNLL **2014**

Levy & Goldberg took two words on each side to "beat" a neural network model with a four word window!

Named Entity Recognition (NER)

How much training data do I need?
"corpora list"

Image Source: v@s3k [a GATE session; http://vas3k.ru/blog/354/]



The departure of Mr Hogan, who originally moved to British Midland as service director from Hertz International in 1997, surprised aviation analysts, as it was believed that he had been brought into the senior executive team of the airline, as part of the group's management succession planning.

He played a leading role in the strategic planning for the rebranding of the airline as BMI in preparation for its entry this year into the scheduled long haul market with the launch of services from Manchester to the US.

BMI has taken on the costs of entry into the North Atlantic market at an unfortunate time, as airlines in North America are facing the toughest conditions for 20 years with many carriers plunging into loss.

BMI, in which Lufthansa of Germany and SAS Scandinavian Airlines each own stakes of 20 per cent, suffered a 26 per cent fall in pre-tax profits last year from £11.1m ($15.7m) to £8.2m on a turnover that grew 16.5 per cent to £739.2m.

In the first six months this year it is understood that passenger volumes have fallen by around two per cent. The share of available seats filled, the load factor, has declined by around two percentage points, but this has been offset by a strong increase in yields, or average fare levels, by more than ten per cent.

Date
Location
Money
Organization
Percentage
Person

- ➡ Conditional Random Field
- ➡ Ensemble Methods; +SVM, HMM, MEMM, … ➡ pyensemble

NB these are corpus-based approaches (supervised)

CoNLL03: http://www.cnts.ua.ac.be/conll2003/ner/

PoS tagging and lemmatization for Named Entity Recognition (NER)

N.B.: This is all supervised (i.e., manually annotated corpora)!

Token	PoS	Lemma	NER
Constitutive	JJ	constitutive	O
binding	NN	binding	O
to	TO	to	O
the	DT	the	O
peri-κ	NN	peri-kappa	B-DNA
B	NN	B	I-DNA
site	NN	site	I-DNA
is	VBZ	be	O
seen	VBN	see	O
in	IN	in	O
monocytes	NNS	monocyte	B-cell
.	.	.	O

chunk

de facto standard
PoS tagset
{NN, JJ, DT, VBZ, …}
Penn Treebank

noun-phrase (chunk)

Begin-Inside-Outside
(relevant) token

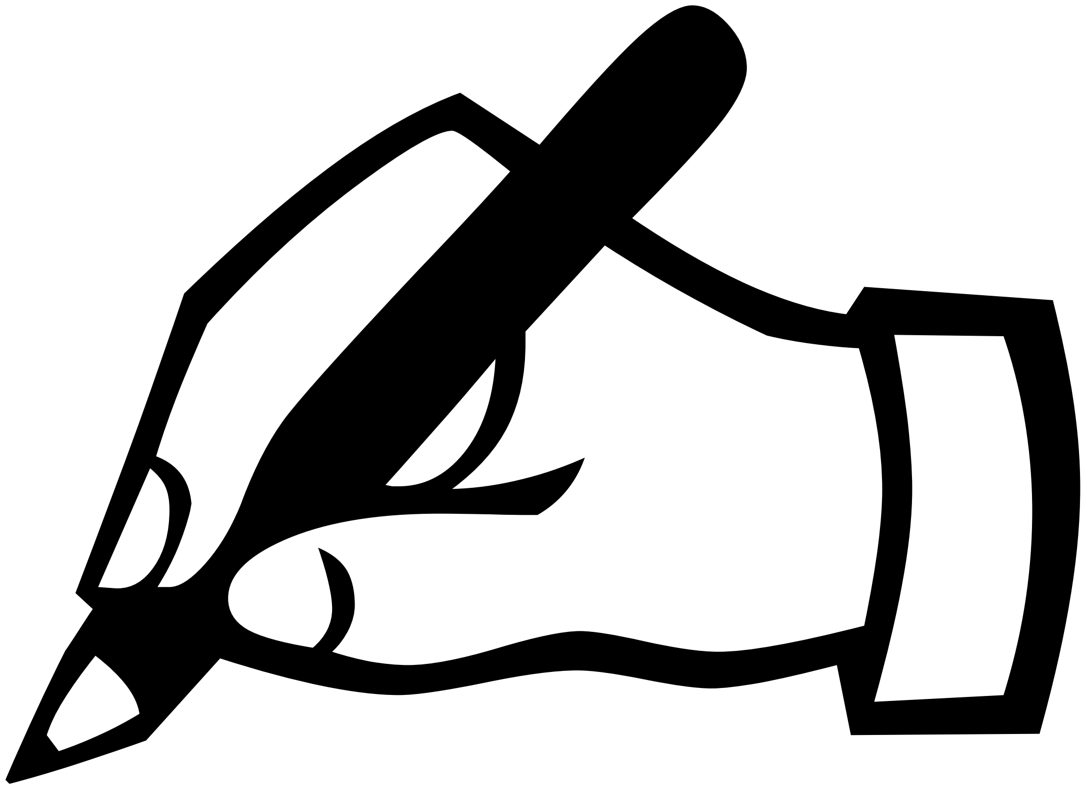
B-I-O
chunk encoding

common
alternatives:
I-O
I-E-O
B-I-E-W-O

End token

(unigram) Word

Practical:
Chunking, tagging, and NER



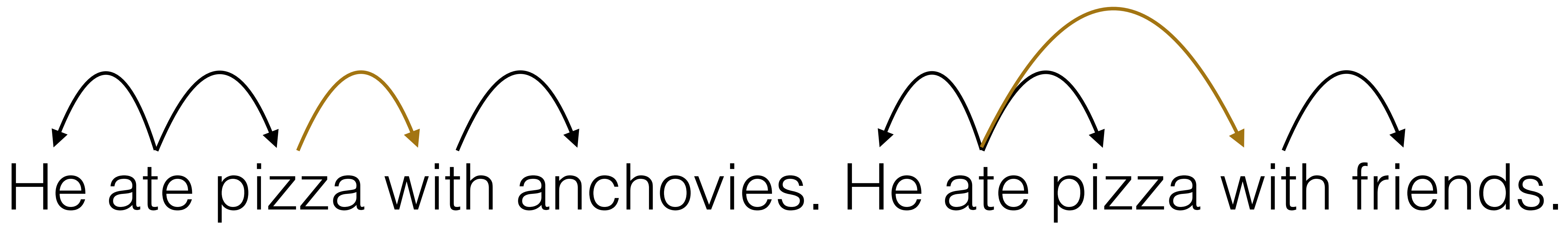
The next step: Relationship extraction

Given this piece of text:

The fourth Wells account moving to another agency is the packaged paper-products division of Georgia-Pacific Corp., which arrived at Wells only last fall. Like Hertz and the History Channel, it is also leaving for an Omnicom-owned agency, the BBDO South unit of BBDO Worldwide. BBDO South **in** Atlanta, which handles corporate advertising for Georgia-Pacific, will assume additional duties for brands like Angel Soft toilet tissue and Sparkle paper towels, said Ken Haldin, a spokesman for Georgia-Pacific **in** Atlanta.

Which organizations operate in Atlanta? (BBDO S., G-P)

Tesnière's dependency relations (1959)



He ate pizza with anchovies. He ate pizza with friends.

ate(he, pizza with anchovies)
~~ate(he, with anchovies)~~

Relationships

ate(he, pizza)
ate(he, with friends)

Dependency parsing 1/2

- Transition-based, arc-standard, shift-reduce, greedy parsing.
- The default approach to dependency parsing today in O(n).
  - **Transition-based**: Move from one token to the next.
  - **Arc-standard**: assign arcs when the dependent token (arrowhead) is fully resolved (alternative: arc-eager → assign the arcs immediately).
  - **Shift-reduce**: A stack of words and a stream buffer: either shift next word from the buffer to the stack or reduce a word from the stack by "arcing".
  - **Greedy**: Make locally optimal transitions (assume independence of arcs).

A shift-reduce parse

(left-arc, right-arc)

Buffer Economic news had little effect on financial markets .

Stack

.

markets

financial

on

effect

little

had

news

Economic

ROOT

Shift: Economic

Shift: news

Left-arc: Economic←news

Shift: had

Left-arc: news←had

Shift: little

Shift: effect

Left-arc: little←effect

Shift: on

Shift: financial

Shift: markets

Left-arc: financial←markets

Right-arc: on→markets

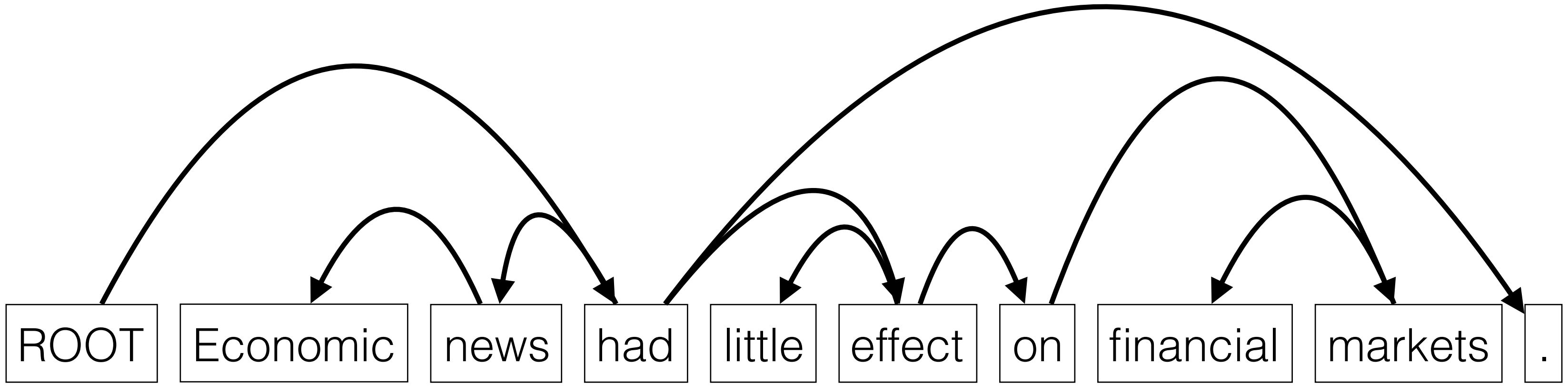
Right-arc: effect→on

Right-arc: had→effect

Shift: .

Right-arc: had→.

Right-arc: ROOT→had



Dependency Parsing. Kübler et al., 2009

Dependency parsing 2/2

- (Arc-standard) Transitions: **shift** or **reduce** (left-arc, right-arc)
- Transitions are chosen using some classifier
  - Maximum entropy classifier, support vector machine, single-layer perceptron, perceptron with one hidden layer (→ Stanford parser, 2014 edition)
- Main issues:
  - Few large, well annotated training corpora ("dependency **treebanks**"). Biomedical domain: GENIA; Newswire: WSJ, Prague, Penn, …
  - **Non-projective** trees (i.e., trees with arcs crossing each other; common in a number of other languages) with arcs that have to be drawn between nodes that are not adjacent on the stack.

Four approaches to relationship extraction

- **Co-mention window**
  - E.g.: if ORG and LOC entity within same sentence and no more than x tokens in between, treat the pair as a hit.
  - Low precision, high recall; trivial, many false positives.
- **Dependency parsing**
  - If a path covering certain nodes (e.g. prepositions like "in/IN" or predicates [~verbs]) connects two entities, extract that pair.
  - Balanced precision and recall, computationally expensive.
- **Pattern extraction**
  - e.g.: <ORG>+ <IN> <LOC>+

    preposition
  - High precision, low recall; cumbersome, but very common.
  - Pattern **learning** can help.
- **Machine Learning**
  - Features for sentences with entities and some classifier (e.g., SVM, neural net, MaxEnt, Bayesian net, …)

    token-distance, #tokens between the entities, tokens before/after them, etc.)
  - Highly variable milages.

    … but loads of fun in your speaker's opinion :)

Practical:
Stanford Tagger and SpaCy