Intelligent Systems (AI-2)

Computer Science cpsc422, Lecture 19

Oct, 23, 2017

Slide Sources
Raymond J. Mooney University of Texas at Austin

D. Koller, Stanford CS - Probabilistic Graphical Models
D. Page, Whitehead Institute, MIT

Several Figures from
"Probabilistic Graphical Models: Principles and Techniques" *D. Koller, N. Friedman* 2009

Lecture Overview

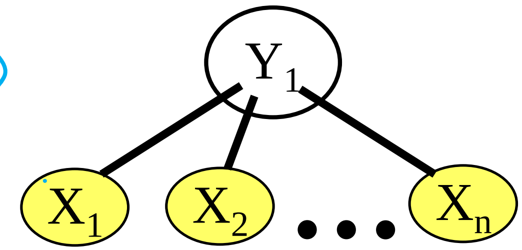
- Recap: Naïve Markov – Logistic regression (simple CRF)
- CRFs: high-level definition
- CRFs Applied to sequence labeling
- NLP Examples: Name Entity Recognition, joint POS tagging and NP segmentation

Let's derive the probabilities we need

$$\phi_i(X_i, Y_1) = \exp\{w_i \mathbb{1}\{X_i = 1, Y_1 = 1\}\}$$

how strongly $Y_1 = 1$ given that $X_i = 1$

$$\phi_0(Y_1) = \exp\{w_0 \mathbb{1}\{Y_1 = 1\}\}$$



$$P(\overset{A}{Y_1} \mid \overset{B}{x_1, \dots, x_n}) = \frac{\tilde{P}(Y, x_1 \dots x_n)}{\tilde{P}(x_1 \dots x_n)}$$

$$\tilde{\tilde{P}}(Y_1, x_1, \dots, x_n) = \phi_0(Y_1) * \prod_{i=1}^{n} \phi_i(x_i, Y_1)$$

$$\tilde{\tilde{P}}(Y_1 = 0, x_1, \dots, x_n) = 1$$

$$\tilde{\tilde{P}}(Y_1 = 1, x_1, \dots, x_n) = e^{w_0 + \sum w_i x_i}$$

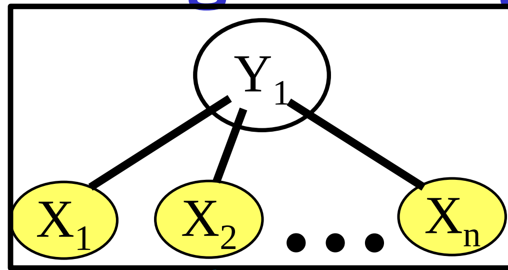
Continue…..

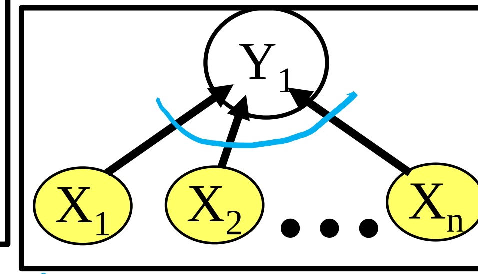
$$P(Y_1 = 1 \mid X_1 \ldots X_n) = \frac{e^{w_0 + \sum w_i X_i}}{1 + e^{w_0 + \sum w_i X_i}}$$

$$= \frac{e^{z}}{1 + e^{z}} \, \frac{e^{-z}}{e^{-z}} = \frac{1}{e^{-z} + 1}$$

$$P(Y_1 \mid X_1 \ldots X_n) = \left\{ \frac{1}{e^{-z} + 1}, \ \frac{e^{-z}}{e^{-z} + 1} \right\}$$

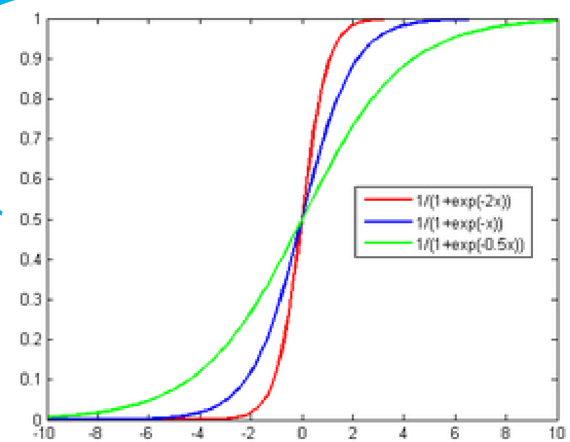
Sigmoid Function used in Logistic Regression

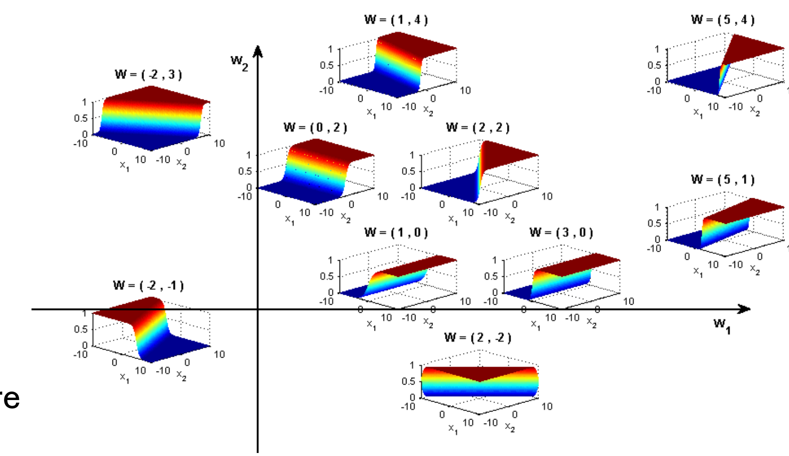




- Great practical interest
- Number of param w_i is linear instead of exponential in the number of parents
- Natural model for many real-world applications
- Naturally aggregates the influence of different parents

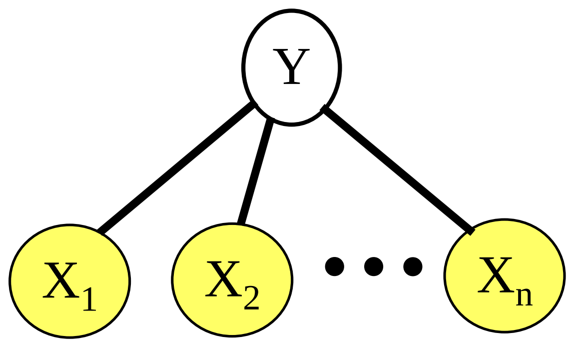
$\frac{1}{1+e^{-x}}$





Logistic Regression as a Markov Net (CRF)

Logistic regression is a simple Markov Net (a CRF)
aka **naïve markov model**



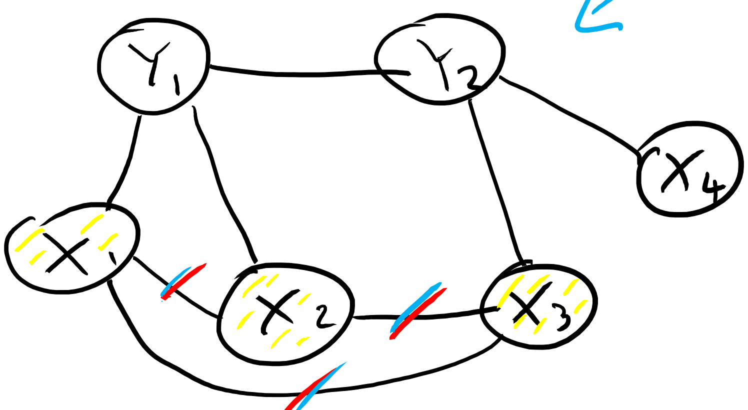
- But only models the **conditional distribution**, P(Y | X) and not the full joint P(X, Y)

Let's generalize ….

Assume that you always observe a set of variables **X** = $\{X_1 \cdots X_n\}$ and you want to predict one or more variables **Y** = $\{Y_1 \cdots Y_k\}$

A CRF is an undirected graphical model whose nodes corresponds to **X** ∪ **Y.**

$\varphi_1(D_1) \cdots \varphi_m(D_m)$ represent the factors which annotate the network (but we disallow factors involving only vars in **X** – why?)



They would be

A. too large

B. constant

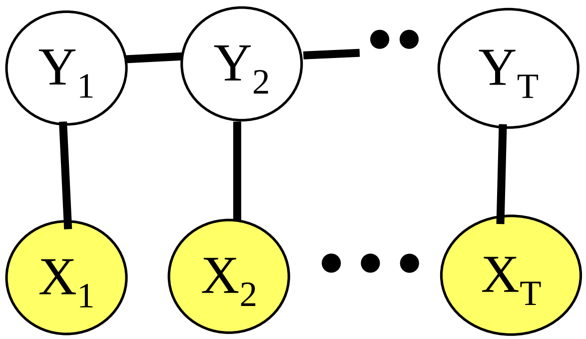
C. difficult to aquire

$$P(Y \mid X) = \frac{1}{Z(X)} \left(\prod_{i=1}^{m} \phi_i(D_i)\right) \qquad Z(X) = \sum_{Y} \left(\prod_{i=1}^{m} \phi_i(D_i)\right)$$

Lecture Overview

- Recap: Naïve Markov – Logistic regression (simple CRF)
- CRFs: high-level definition
- **CRFs Applied to sequence labeling**
- NLP Examples: Name Entity Recognition, joint POS tagging and NP segmentation

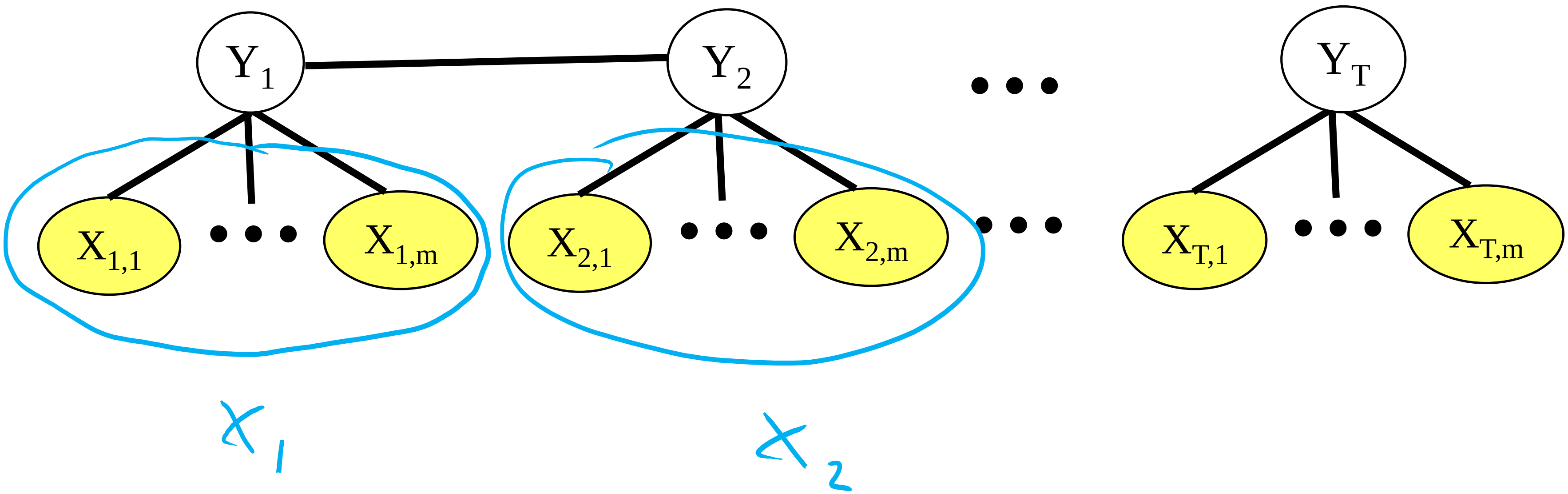
Sequence Labeling



Linear-chain CRF

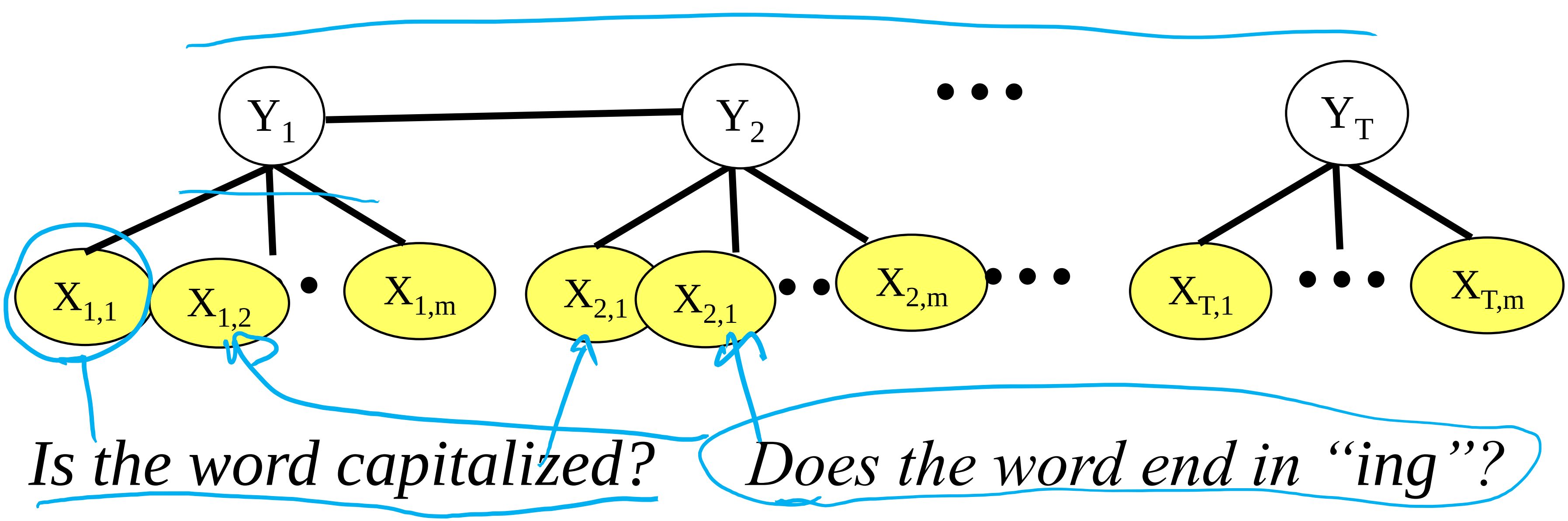
Increase representational Complexity: Adding Features to a CRF

- Instead of a single observed variable X_i we can model multiple features X_{ij} of that observation.



CRFs in Natural Language Processing

- One target variable Y for each word X, encoding the possible labels for X
- Each target variable is connected to a set of feature variables that capture properties relevant to the target distinction



Named Entity Recognition Task

- Entity often span multiple words *"British Columbia"*
- Type of an entity may not be apparent for individual words *"University of British Columbia"*
- Let's assume three categories*: **Person, Location, Organization***
- BIO notation (for sequence labeling)

possible labels B-PER I-PER B-LOC I-LOC
B-ORG I-ORG OTHER

O B-ORG I-ORG I-ORG I-ORG
The University of British Columbia

O O B-LOC I-LOC
is in Vancouver B.C.

Linear chain CRF parameters

With two factors "types" for each word

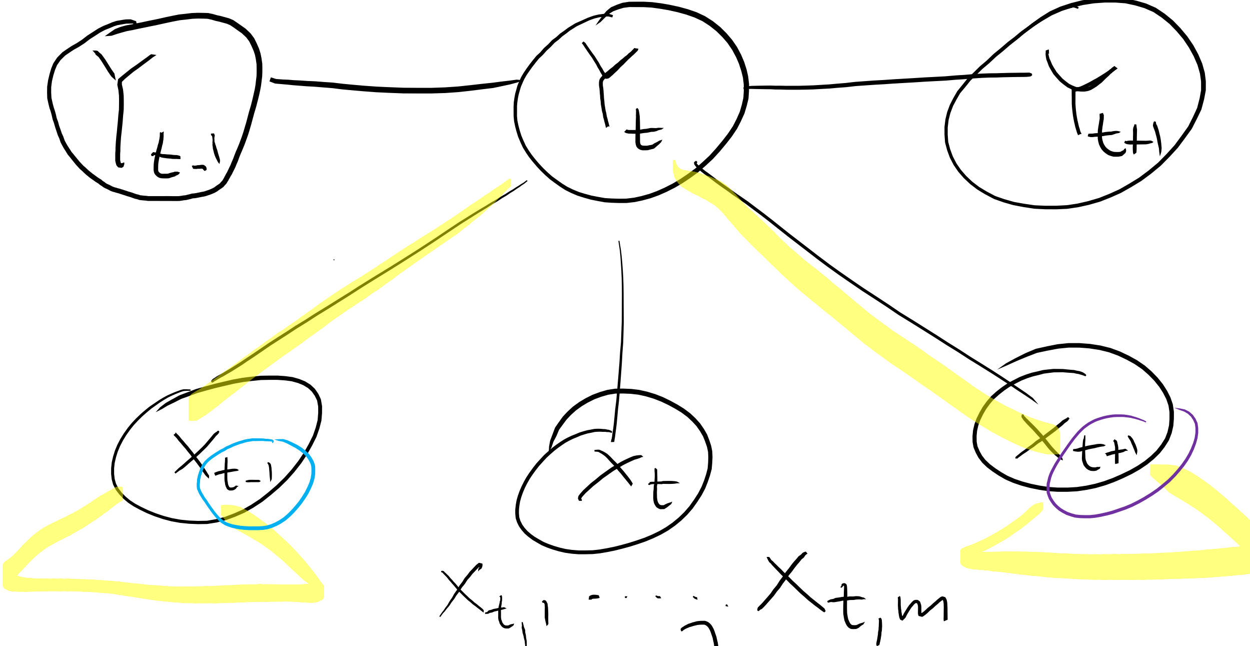
$\phi_t^1(Y_t, Y_{t-1})$ $\phi_t^1(Y_t, Y_{t+1})$ Dependency between neighboring target vars

$\phi_t^2(Y_t, X_1, .., X_T)$ Dependency between target variable and its context in the word sequence, which can include also **features of the words** (capitalized, appear in an atlas of location names, etc.)

Factors are similar to the ones for the Naïve Markov (logistic regression)

$$\phi_t(Y_t, X_{tk}) = \exp\{w_{tk} \times \mathbb{1}\{Y_t = \text{I-LOC}, X_{tk} = 1\}\}$$

appears in atlas of location names



$$\mathbb{1}\{Y_t = \text{I-ORG}, X_{t,k} = \text{"Times"}\}$$

$$\mathbb{1}\{Y_t = \text{I-PER}, X_{t+1,k} = \text{"spoke"}\}$$

$$\mathbb{1}\{Y_t = \text{I-PER}, X_{t-1,k} = \text{"Mrs."}\}$$

Features can also be

- The word
- Following word
- Previous word

More on features

Including features that are **conjunctions of simple features** increases accuracy

$\mathbb{1}\{Y_t = \text{I-PER}, X_{t+1,k} = \text{"spoke"}\}$

$\mathbb{1}\{Y_t = \text{I-PER}, X_{t-1,k} = \text{"Mrs."}\}$

Y_t

Total number of features can be 10^5–10^6

However features are sparse i.e. most features are 0 for most words

Linear-Chain Performance

Per-token/word accuracy in the high 90% range for many natural datasets label is wrong for 2 words out of 9

Per-field precision and recall are more often around 80–95% , depending on the dataset. Entire Named Entity Phrase must be correct. only one is correct out of 2

O B-ORG I-ORG B-LOC I-LOC
The University of British Columbia

O O B-LOC I-LOC
is in Vancouver B.C

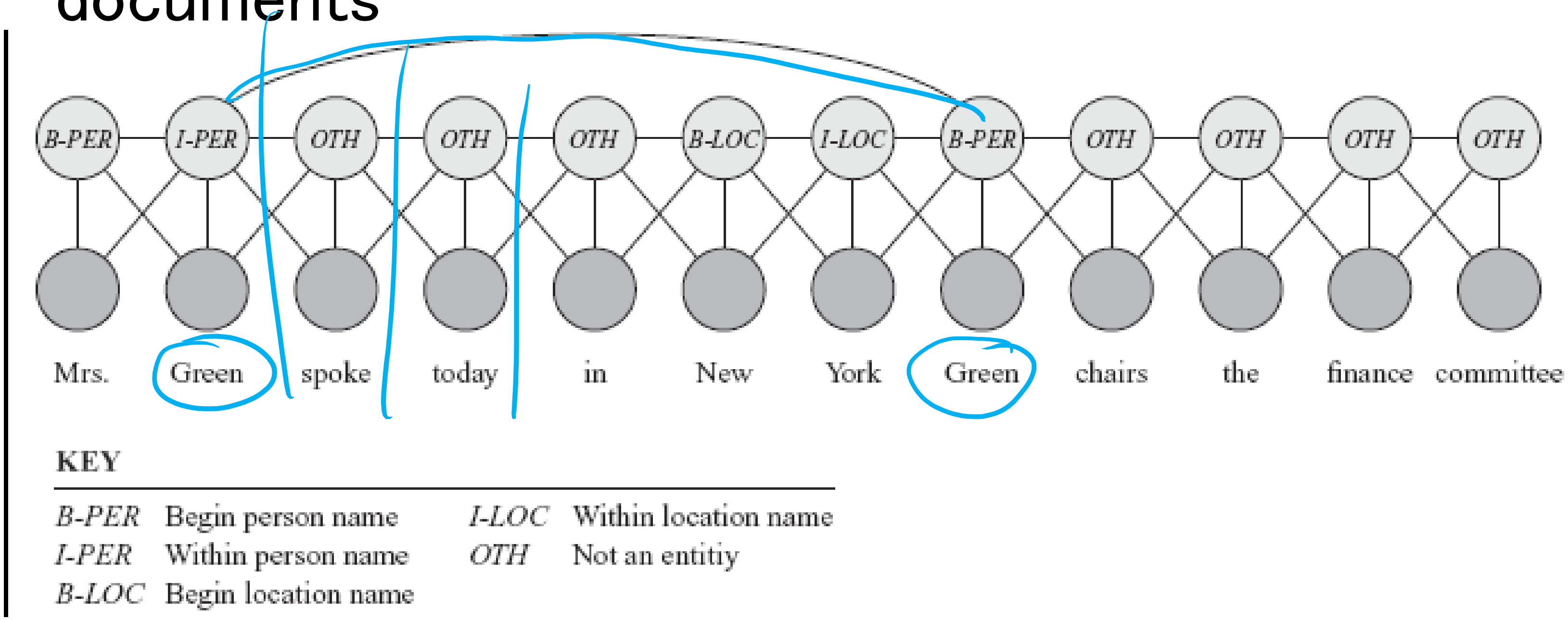
	A.	B	C.
Per-word accuracy?	1/2	7/9	7/9
Per-field precision?	1/2	3/9	1/2

i>clicker.

Skip-Chain CRFs

Include additional factors that connect non-adjacent target variables

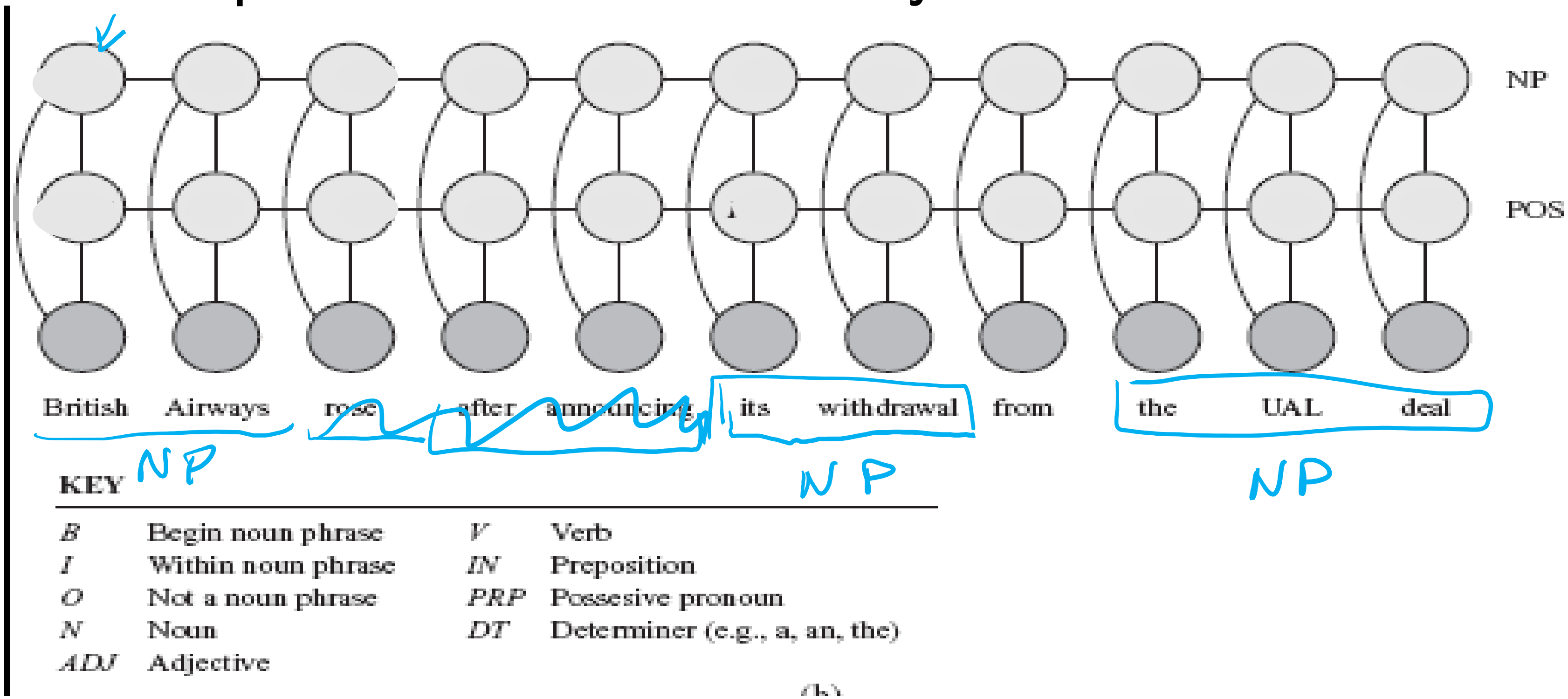
E.g., When a word occur multiple times in the same documents



Graphical structure over Y can depend on the values of the Xs !

Coupled linear-chain CRFs

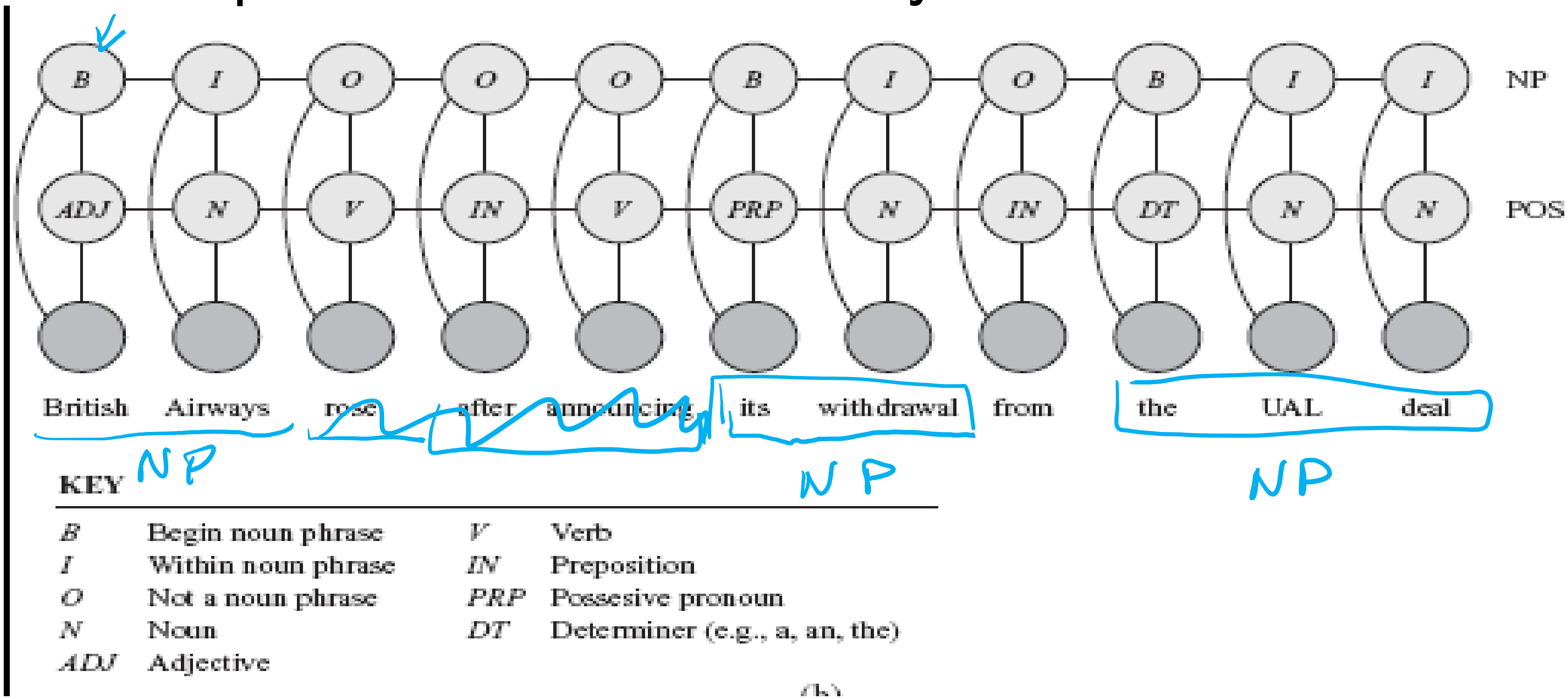
- Linear-chain CRFs can be combined to perform multiple tasks simultaneously



- Performs part-of-speech labeling and noun-phrase segmentation

Coupled linear-chain CRFs

- Linear-chain CRFs can be combined to perform multiple tasks simultaneously



- Performs part-of-speech labeling and noun-phrase segmentation

Inference in CRFs (just intuition)

An HMM can be viewed as a factor graph $p(\mathrm{y},\mathrm{x}) = \boxed{\prod_t \Psi_t(y_t, y_{t-1}, x_t)}$ where $Z = 1$, and the factors are defined as:

$$\Psi_t(j, i, x) \stackrel{\text{def}}{=} p(y_t = j | y_{t-1} = i) p(x_t = x | y_t = j). \tag{4.1}$$

Forward / Backward / Smoothing and Viterbi can be rewritten (not trivial!) using these factors

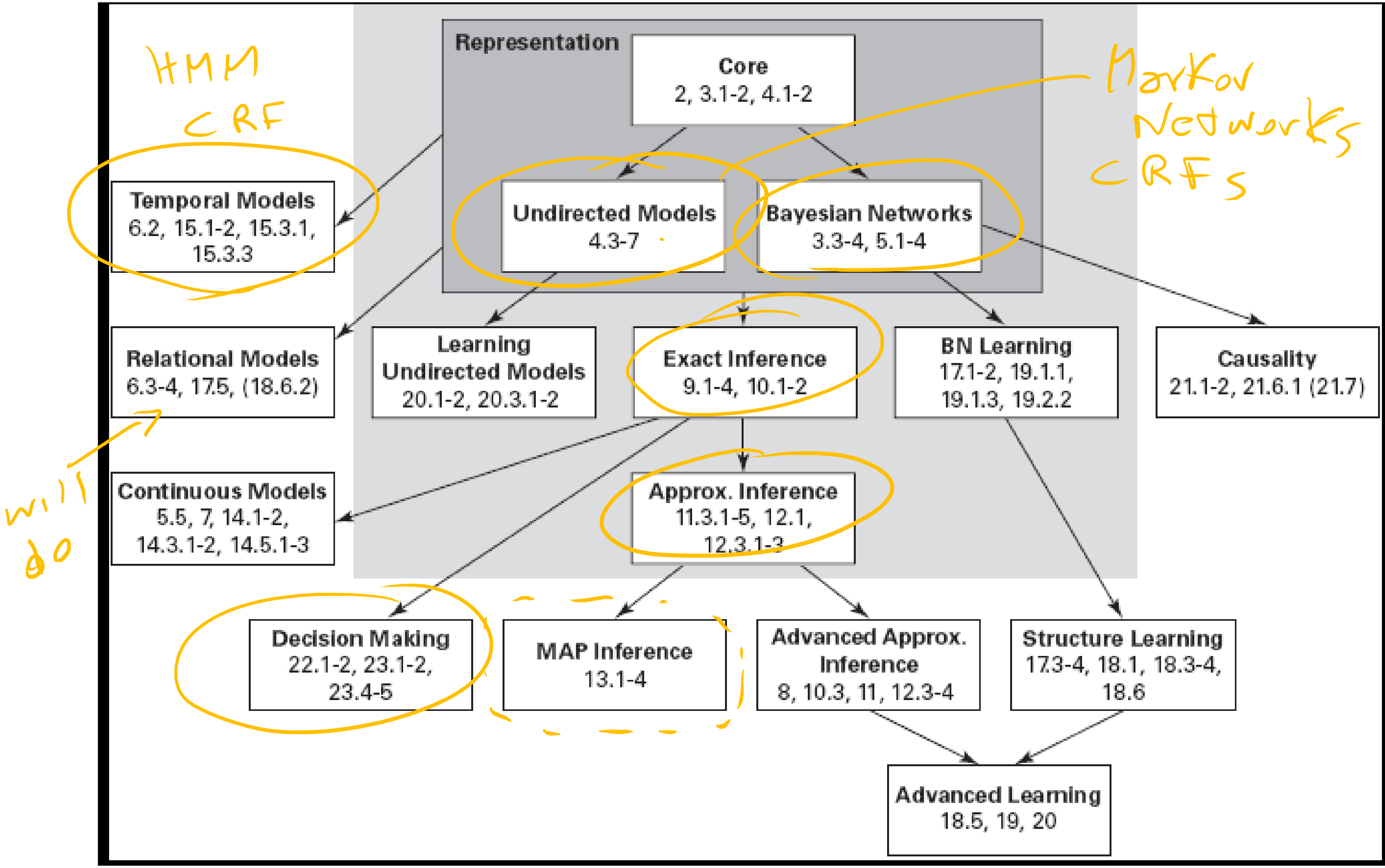
Then you plug in the factors of the CRFs and all the algorithms work fine with CRFs! ☺

CRFs Summary

- Ability to incorporate arbitrary overlapping local and global features
- Graphical structure over Y can depend on the values of the Xs (see slide 21)
- Can perform multiple tasks simultaneously (see slide 23)
- *Standard Inference algorithm* for HMM can be applied
- *Practical Learning algorithms exist*
- State-of-the-art on many labeling tasks *(deep learning recently shown to be often better when large training data are available… current research on ensembling them!)*

See MALLET package

Probabilistic Graphical Models



From "Probabilistic Graphical Models: Principles and Techniques" *D. Koller, N. Friedman* 2009

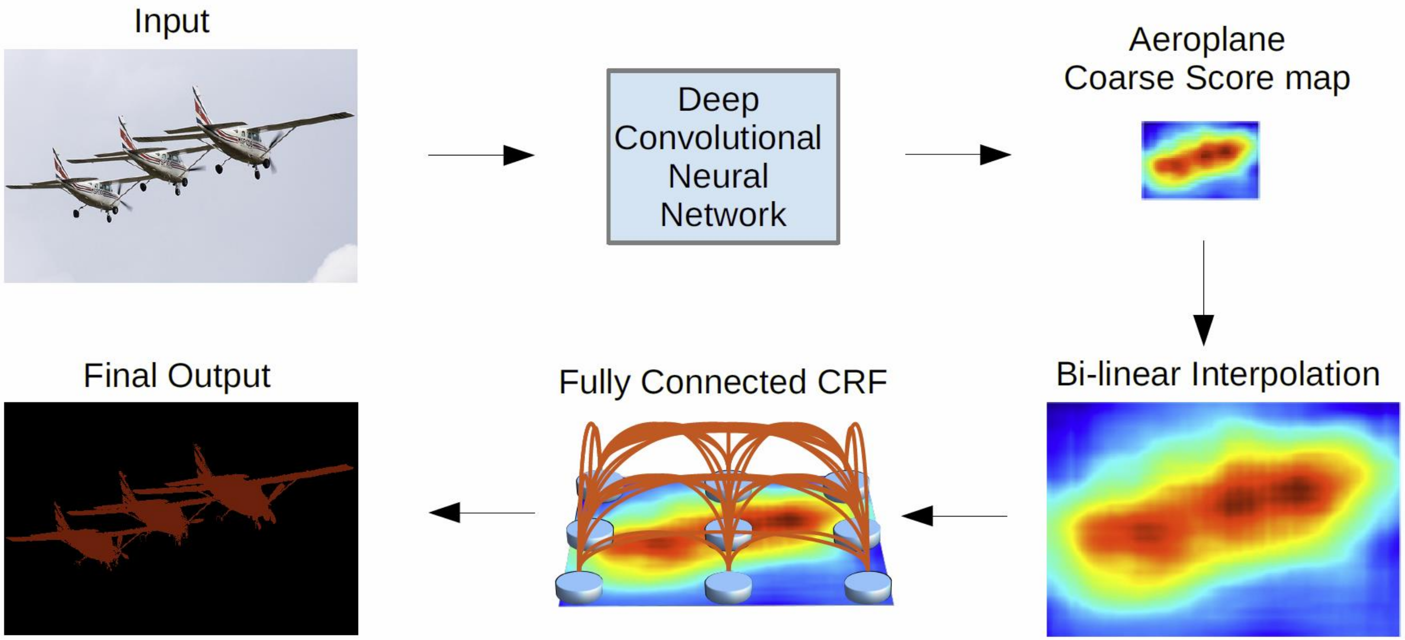
Combining CRFs and Neural Models

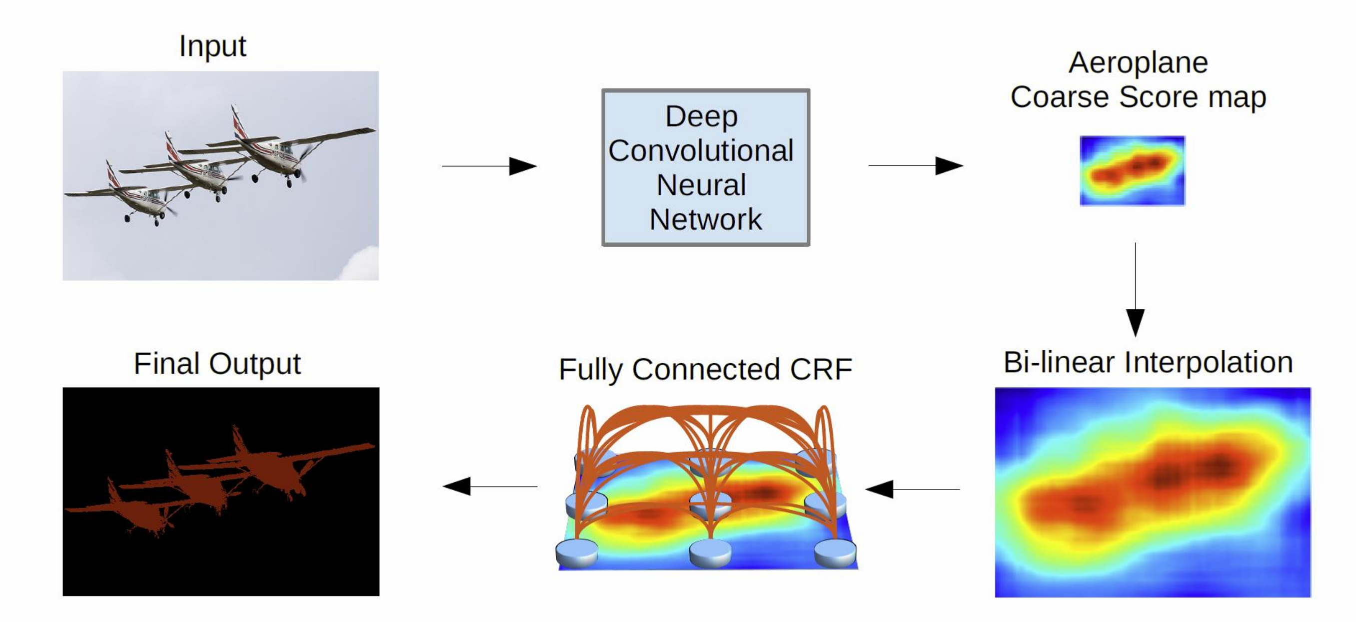
SEMANTIC IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NETS AND FULLY CONNECTED CRFS

International Conference on Learning Representations (ICLR), San Diego, California, USA, May 2015.

Liang-Chieh Chen Univ. of California, Los Angeles; George Papandreou Google Inc. ; Iasonas Kokkinos INRIA ; Kevin Murphy Google Inc. ; Alan L. Yuille Univ. of California, Los Angeles

1. Use CNN to generate a rough prediction of segmentation (smooth, blurry heat map)
2. Refine this prediction with a conditional random field (CRF)




Input
Deep Convolutional Neural Network
Aeroplane Coarse Score map
Bi-linear Interpolation
Fully Connected CRF
Final Output

422 big picture: Where are we?

Hybrid: Det +Sto

Prob CFG
Prob Relational Models
Markov Logics

Deterministic | Stochastic

Query

Logics
First Order Logics

Ontologies
Temporal rep.

- Full Resolution
- SAT

Belief Nets

Approx. : Gibbs

Markov Chains and HMMs

Forward, Viterbi….
Approx. : Particle Filtering

Undirected Graphical Models
Markov Networks
Conditional Random Fields

Planning

Markov Decision Processes and
Partially Observable MDP

- Value Iteration
- Approx. Inference

Reinforcement Learning

Applications of AI

Representation

Reasoning
Technique

Learning Goals for today's class

You can:

- Provide general definition for CRF
- Apply CRFs to sequence labeling
- Describe and justify features for CRFs applied to Natural Language processing tasks
- Explain benefits of CRFs

Midterm, Wed, Oct 25, we will start at noon sharp

How to prepare…

- Go to **Office Hours (extra hours offered)**
- **Learning Goals** (look at the end of the slides for each lecture – complete list has been posted)
- Revise all the **clicker questions** and **practice exercises**
- **More practice material** has been posted
- Check questions and answers on Piazza

Next class Fri

- Start Logics
- Revise Logics from 322!