Structured Probabilistic Models for Deep Learning

费行健

Why we talk about this?

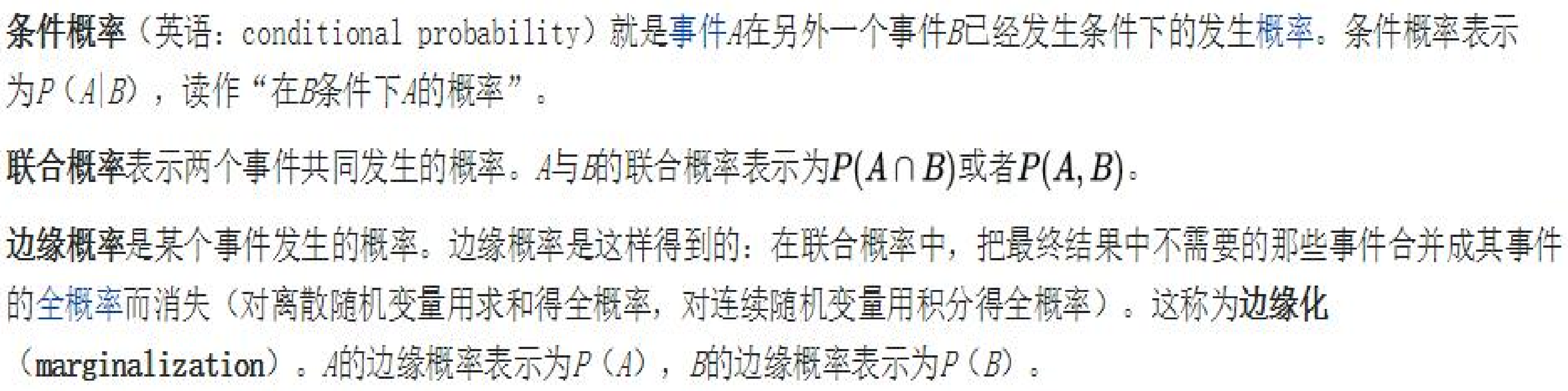
- Deep learning draws upon many modeling formalisms that researchers can use to guide their design efforts and describe their algorithms.
- One of these formalisms is the idea of **structured probabilistic models**.

What is structured probabilistic models?

- A structured probabilistic model is a way of describing a **probability distribution**, using a **graph** to describe which random variables in the probability distribution interact with each other **directly**.
- Because the structure of the model is defined by a graph, these models are often also referred to as **graphical models**.
- HMM, CRF, topic model(LDA)

First, let's solve a problem

- 现在有一个4*100米的接力比赛，其中一个队，有四名队员，分别记为A, B, C, D, 他们每个人跑完各自一圈，对应的总时间（算上前面队友花掉的时间）分别为t0, t1, t2, t3, 现在要求
- P(t0, t1, t2, t3) 这个联合概率分布。



条件概率（英语：conditional probability）就是事件*A*在另外一个事件*B*已经发生条件下的发生概率。条件概率表示为$P(A|B)$，读作“在*B*条件下*A*的概率”。

联合概率表示两个事件共同发生的概率。*A*与*B*的联合概率表示为$P(A\cap B)$或者$P(A,B)$。

边缘概率是某个事件发生的概率。边缘概率是这样得到的：在联合概率中，把最终结果中不需要的那些事件合并成其事件的全概率而消失（对离散随机变量用求和得全概率，对连续随机变量用积分得全概率）。这称为**边缘化**（**marginalization**）。*A*的边缘概率表示为$P(A)$，*B*的边缘概率表示为$P(B)$。

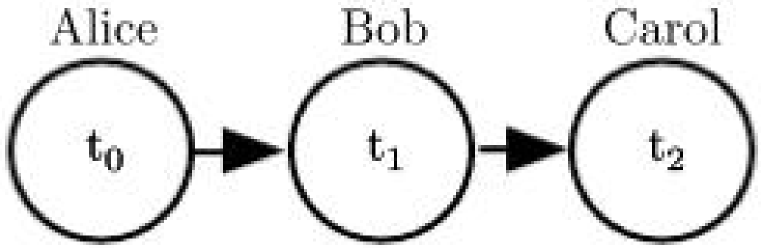
The basic of graphical models

- Each **node** represents a **random variable**.
- Each **edge** represents a **direct interaction**.

- There is more than one way to describe the **interactions** in a probability distribution using a graph.
- Graphical models can be largely divided into two categories:
- **directed <u>acyclic</u> graphs**
- **undirected graphs**

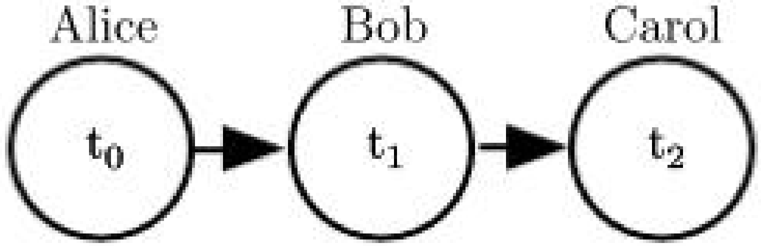
Directed graphical model

- Directed graphical model, belief network, Bayesian network.
- Drawing an arrow from a to b means that we define the probability distribution over b via a **conditional distribution**, with a as one of the variables on the right side of the conditioning bar.



Math

$$p(\mathbf{x}) = \Pi_i p(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)).$$



$$p(\mathrm{t}_0, \mathrm{t}_1, \mathrm{t}_2) = p(\mathrm{t}_0)p(\mathrm{t}_1 \mid \mathrm{t}_0)p(\mathrm{t}_2 \mid \mathrm{t}_1).$$

Computational complexity analysis

- 假设范围是K，有m个变量，打表计算联合概率复杂度是
- O(k^m)

$$p(\mathrm{t}_0, \mathrm{t}_1, \mathrm{t}_2) = p(\mathrm{t}_0)p(\mathrm{t}_1 \mid \mathrm{t}_0)p(\mathrm{t}_2 \mid \mathrm{t}_1).$$

- 而若只需计算这些条件概率，假设每个变量的父节点个数是n，那么复杂度是O(m*k^(n+1)) = O(k^n)

- 因为通常n << m, 所以用图模型计算的复杂度，远小于用无结构模型计算的复杂度。

- 与DL的对比。

Training, sampling, inference

- 回想朴素贝叶斯，再试着用图模型表示朴素贝叶斯。

- 采样，拓扑排序。

- 学习与推断，学习和推断都是对参数的估计，因此可以都归为推断。推断，有变量消去和信念传播(Belief Propagation)两种算法，它们本质上是动态规划，但问题是这些方法还是复杂度很高！NP hard
- 所以一般做近似推断，19章讲。

Undirected Models

- Undirected models, Markov random fields, Markov networks.

- 边的含义：两个变量间互相影响。
- Unlike directed models, the edge in an undirected model has noarrow, and is **not associated with a conditional probability** distribution.

Example

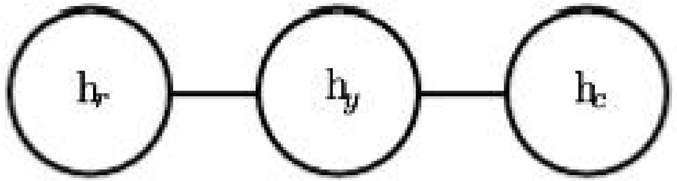
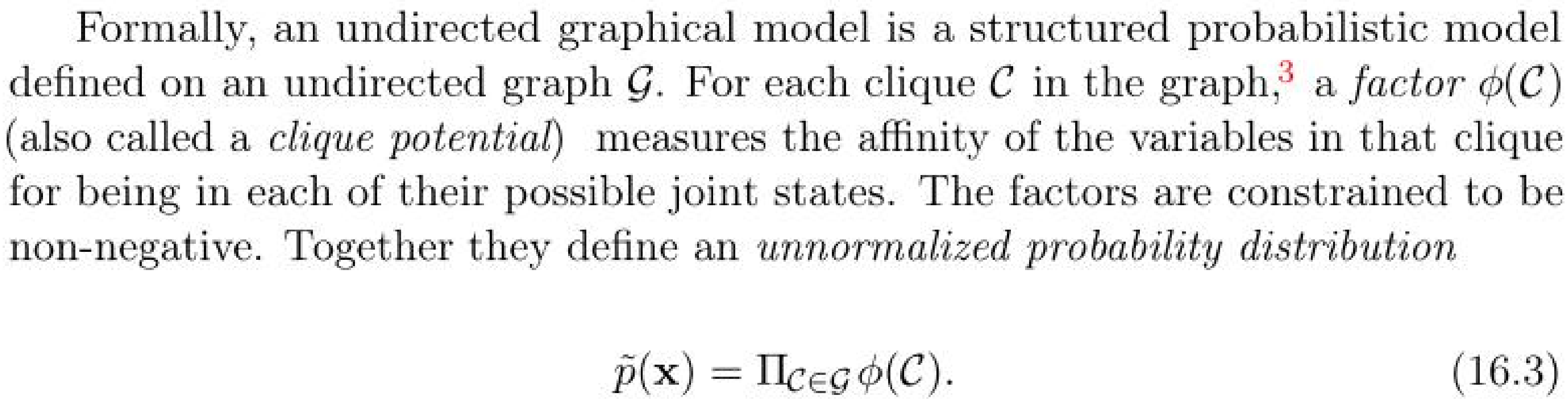


Figure 16.3: An undirected graph representing how your roommate's health h_r, your health h_y, and your work colleague's health h_c affect each other. You and your roommate might infect each other with a cold, and you and your work colleague might do the same, but assuming that your roommate and your colleague do not know each other, they can only infect each other indirectly via you.

Math



Formally, an undirected graphical model is a structured probabilistic model defined on an undirected graph $\mathcal{G}$. For each clique $\mathcal{C}$ in the graph,[3] a *factor* $\phi(\mathcal{C})$ (also called a *clique potential*) measures the affinity of the variables in that clique for being in each of their possible joint states. The factors are constrained to be non-negative. Together they define an *unnormalized probability distribution*

$$\tilde{p}(\mathbf{x}) = \Pi_{\mathcal{C}\in\mathcal{G}}\,\phi(\mathcal{C}). \tag{16.3}$$

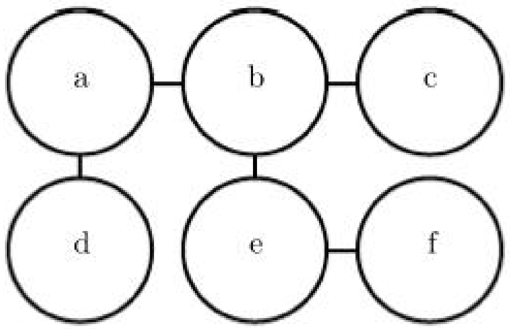
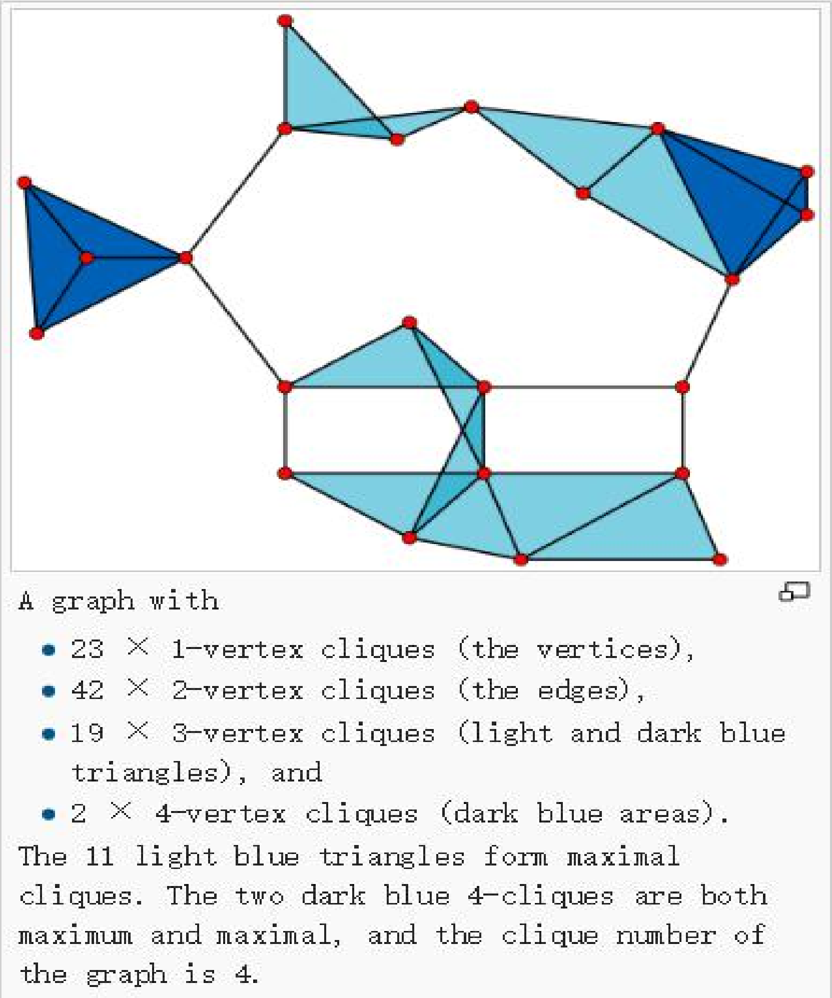


Figure 16.4: This graph implies that $p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e},\mathrm{f})$ can be written as $\frac{1}{Z}\phi_{\mathrm{a},\mathrm{b}}(\mathrm{a},\mathrm{b})\phi_{\mathrm{b},\mathrm{c}}(\mathrm{b},\mathrm{c})\phi_{\mathrm{a},\mathrm{d}}(\mathrm{a},\mathrm{d})\phi_{\mathrm{b},\mathrm{e}}(\mathrm{b},\mathrm{e})\phi_{\mathrm{e},\mathrm{f}}(\mathrm{e},\mathrm{f})$ for an appropriate choice of the ϕ functions.

What is clique

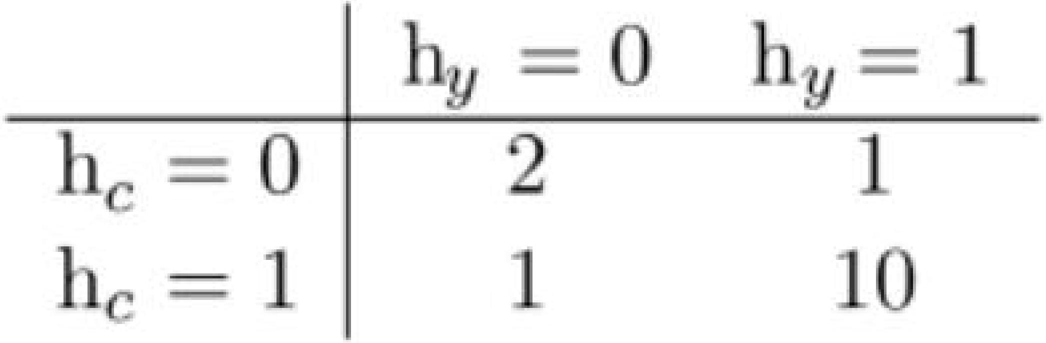
- In the mathematical area of graph theory, a clique is a subset of vertices of an undirected graph such that its induced subgraph is complete; that is, every two distinct vertices in the clique are adjacent.



A graph with

- 23 × 1-vertex cliques (the vertices),
- 42 × 2-vertex cliques (the edges),
- 19 × 3-vertex cliques (light and dark blue triangles), and
- 2 × 4-vertex cliques (dark blue areas).

The 11 light blue triangles form maximal cliques. The two dark blue 4-cliques are both maximum and maximal, and the clique number of the graph is 4.



	$\mathrm{h}_y = 0$	$\mathrm{h}_y = 1$
$\mathrm{h}_c = 0$	2	1
$\mathrm{h}_c = 1$	1	10

Sampling and inference and training

- Sampling: 无向图不存在拓扑序，因此只能用其他方法。
- 吉布斯采样，17章讲。

- inference and training
- 同有向图模型

How to determine the structure of the model

- The primary advantage of using structured probabilistic models is that they allow us to dramatically reduce the cost of representing probability distributions as well as learning and inference.
- The primary mechanism that allows all of these operations to use less runtime and memory is choosing to **not model certain interactions**.

Structure learning

- Most structure learning techniques are a form of **greedy search**.
- 1 A structure is proposed, a model with that structure is trained, then given a score. The score rewards high training set accuracy and penalizes model complexity.
- 2 Candidate structures with a small number of edges added or removed are then proposed as the next step of the search.
- 3 The search proceeds to a new structure that is expected to increase the score.

与 Deep learning 对比

- Deep learning model 是不是图模型？

- 在DL 模型中，我们是如何建模变量间的依赖关系的？
- 隐变量！

- 有什么好处？
- 不需要改变网络结构，省去了搜索和多轮训练的麻烦。

- 怎么做到的？
- A fixed structure over visible and hidden variables can use direct interactions between visible and hidden units to impose indirect interactions between visible units.

与 Deep learning 对比

- 非线性相关性：
- DL model:
- Complicated nonlinear interactions between variables are accomplished via indirect connections that flow through multiple latent variables.

- Traditional model:
- Traditional models mostly use higher-order terms and structure learning to capture complicated nonlinear interactions between variables.

与 Deep learning 对比

- 隐变量的数目：
- 图模型通常大部分变量都是观察到的变量，隐变量较少。
- 而深度学习模型，通常隐变量数目，比观察到的变量数目要多得多。

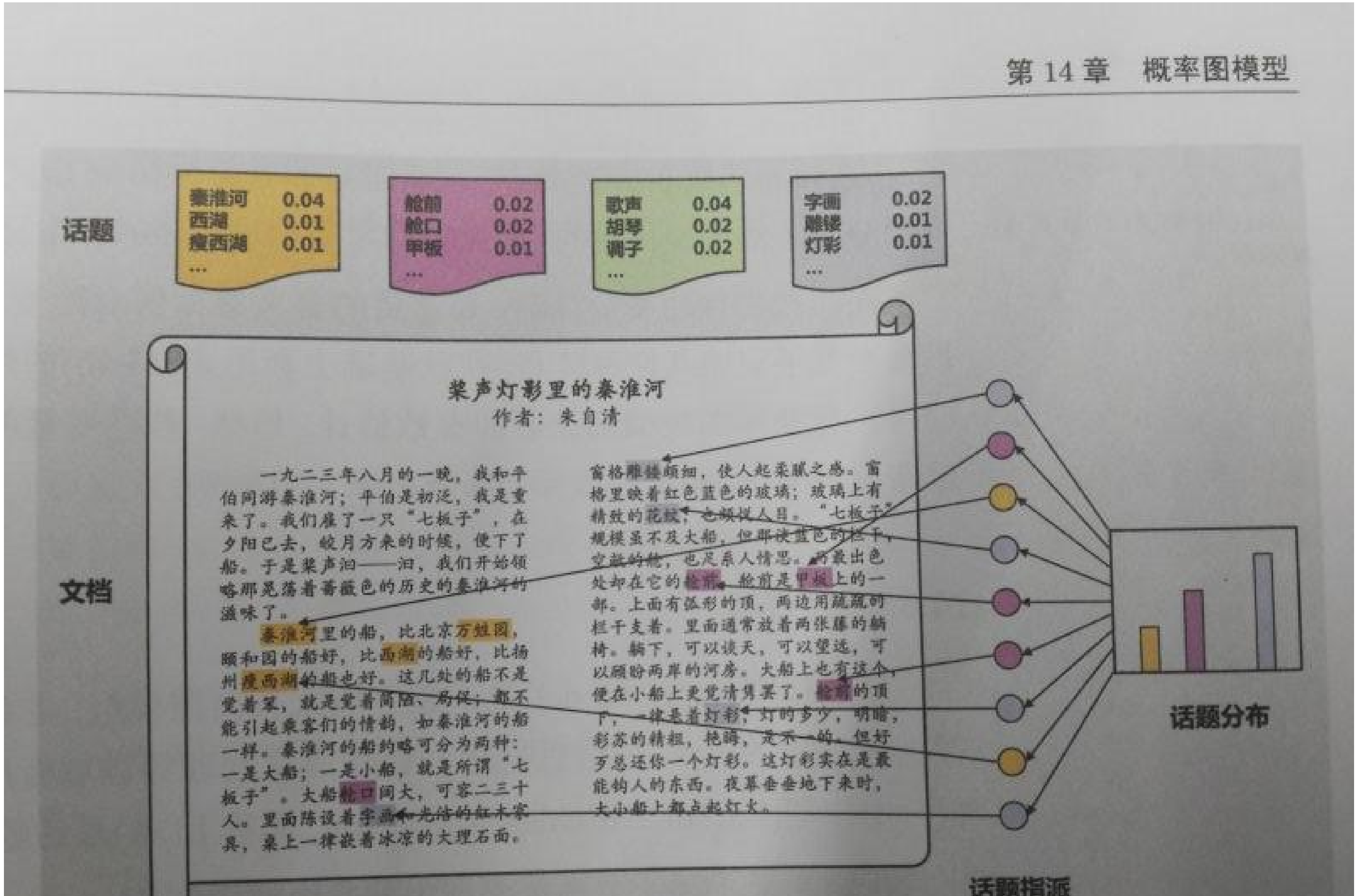
- 隐变量的含义：
- 深度学习中隐变量，通常事先没有定义特别的语义，而图模型有，例如文档主题
- 所以深度学习模型，隐变量的含义是通过学习算法创造出来的，但具体是什么语义，人类并不能做很好解释，而图模型可以

隐变量的含义

- CNN中间结果可视化分析
- http://blog.csdn.net/wiking__acm/article/details/51273512

隐变量的含义（感谢叶子同学的拍照!）

- LDA的例子



第 14 章　概率图模型

话题

秦淮河	0.04
西湖	0.01
瘦西湖	0.01
…	

舱前	0.02
舱口	0.02
甲板	0.01
…	

歌声	0.04
胡琴	0.02
调子	0.02
…	

字画	0.02
雕镂	0.01
灯彩	0.01
…	

文档

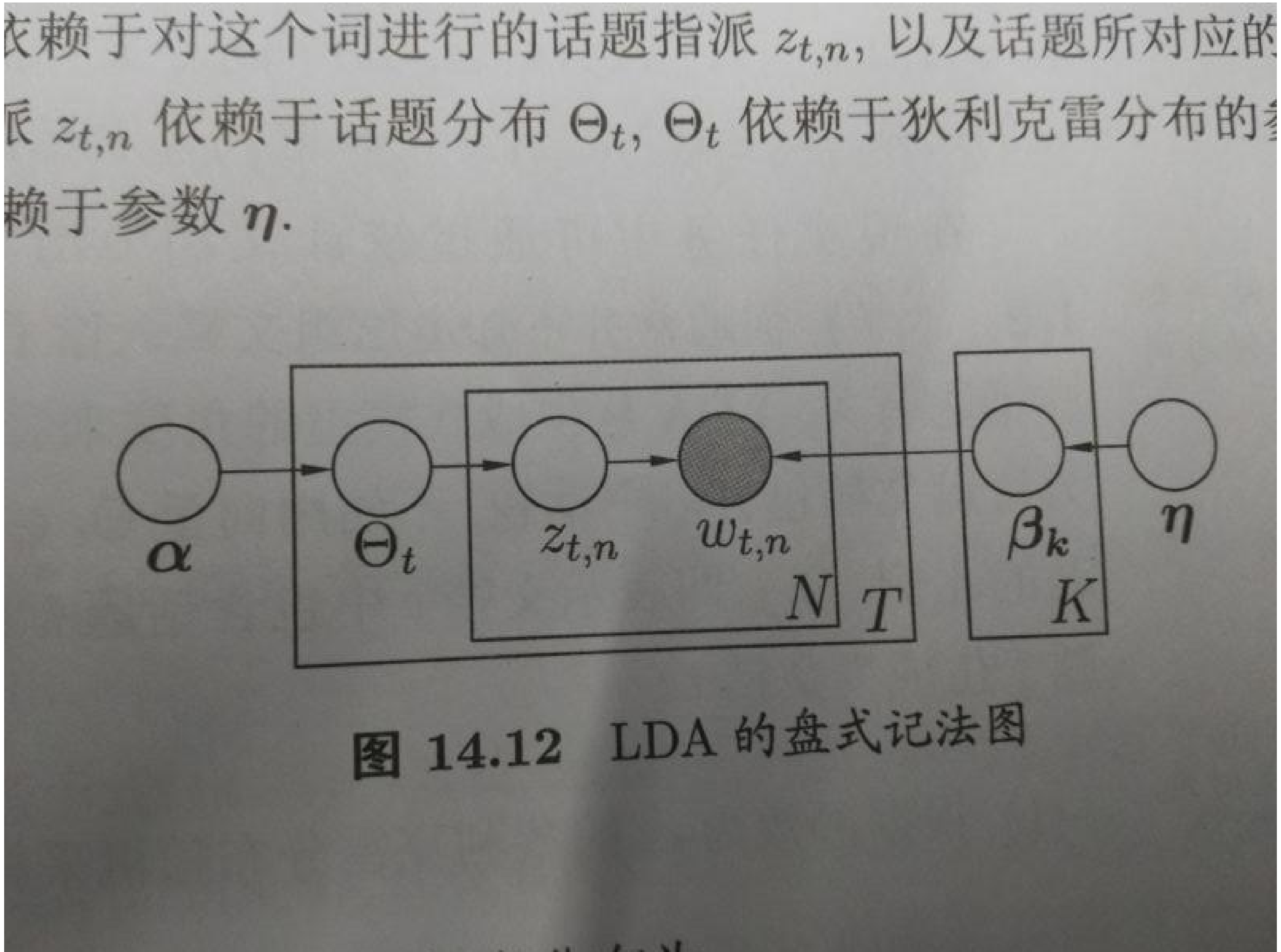
桨声灯影里的秦淮河

作者：朱自清

一九二三年八月的一晚，我和平伯同游秦淮河；平伯是初泛，我是重来了。我们雇了一只“七板子”，在夕阳已去，皎月方来的时候，便下了船。于是桨声汩——汩，我们开始领略那晃荡着蔷薇色的历史的秦淮河的滋味了。

秦淮河里的船，比北京万甡园，颐和园的船好，比西湖的船好，比扬州瘦西湖的船也好。这几处的船不是觉着笨，就是觉着简陋、局促；都不能引起乘客们的情韵，如秦淮河的船一样。秦淮河的船约略可分为两种：一是大船；一是小船，就是所谓“七板子”。大船舱口阔大，可容二三十人。里面陈设着字画和光洁的红木家具，桌上一律嵌着冰凉的大理石面。窗格雕镂颇细，使人起柔腻之感。窗格里映着红色蓝色的玻璃；玻璃上有精致的花纹，也颇悦人目。“七板子”规模虽不及大船，但那淡蓝色的栏干，空敞的舱，也足系人情思。而最出色处却在它的舱前。舱前是甲板上的一部。上面有弧形的顶，两边用疏疏的栏干支着。里面通常放着两张藤的躺椅。躺下，可以谈天，可以望远，可以顾盼两岸的河房。大船上也有这个，便在小船上更觉清隽罢了。舱前的顶下，一律悬着灯彩；灯的多少，明暗，彩苏的精粗，艳晦，是不一的。但好歹总还你一个灯彩。这灯彩实在是最能钩人的东西。夜幕垂垂地下来时，大小船上都点起灯火。

LDA



依赖于对这个词进行的话题指派 $z_{t,n}$，以及话题所对应的

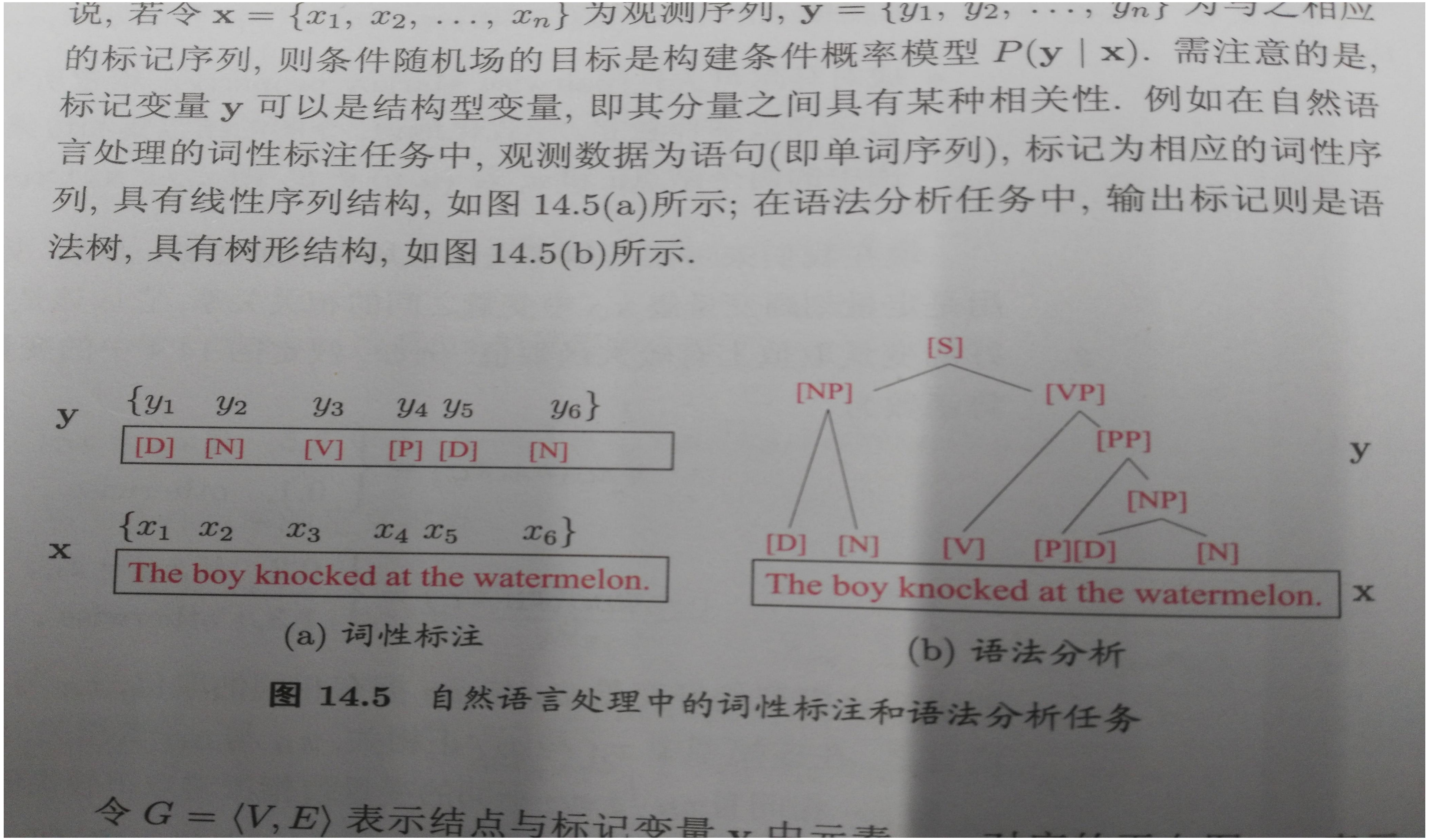
派 $z_{t,n}$ 依赖于话题分布 Θ_t，Θ_t 依赖于狄利克雷分布的参

赖于参数 $\boldsymbol{\eta}$.

图 14.12 LDA 的盘式记法图

图结构的对比

- 问题：



说，若令 $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ 为观测序列，$\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ 为与之相应的标记序列，则条件随机场的目标是构建条件概率模型 $P(\mathbf{y} \mid \mathbf{x})$. 需注意的是，标记变量 $\mathbf{y}$ 可以是结构型变量，即其分量之间具有某种相关性. 例如在自然语言处理的词性标注任务中，观测数据为语句(即单词序列)，标记为相应的词性序列，具有线性序列结构，如图 14.5(a)所示；在语法分析任务中，输出标记则是语法树，具有树形结构，如图 14.5(b)所示.

y
$\{y_1$ y_2 y_3 y_4 y_5 $y_6\}$
[D] [N] [V] [P] [D] [N]
x
$\{x_1$ x_2 x_3 x_4 x_5 $x_6\}$
The boy knocked at the watermelon.
[S]
[NP]
[VP]
[PP]
[NP]
[D] [N] [V] [P][D] [N]
y
The boy knocked at the watermelon.
x

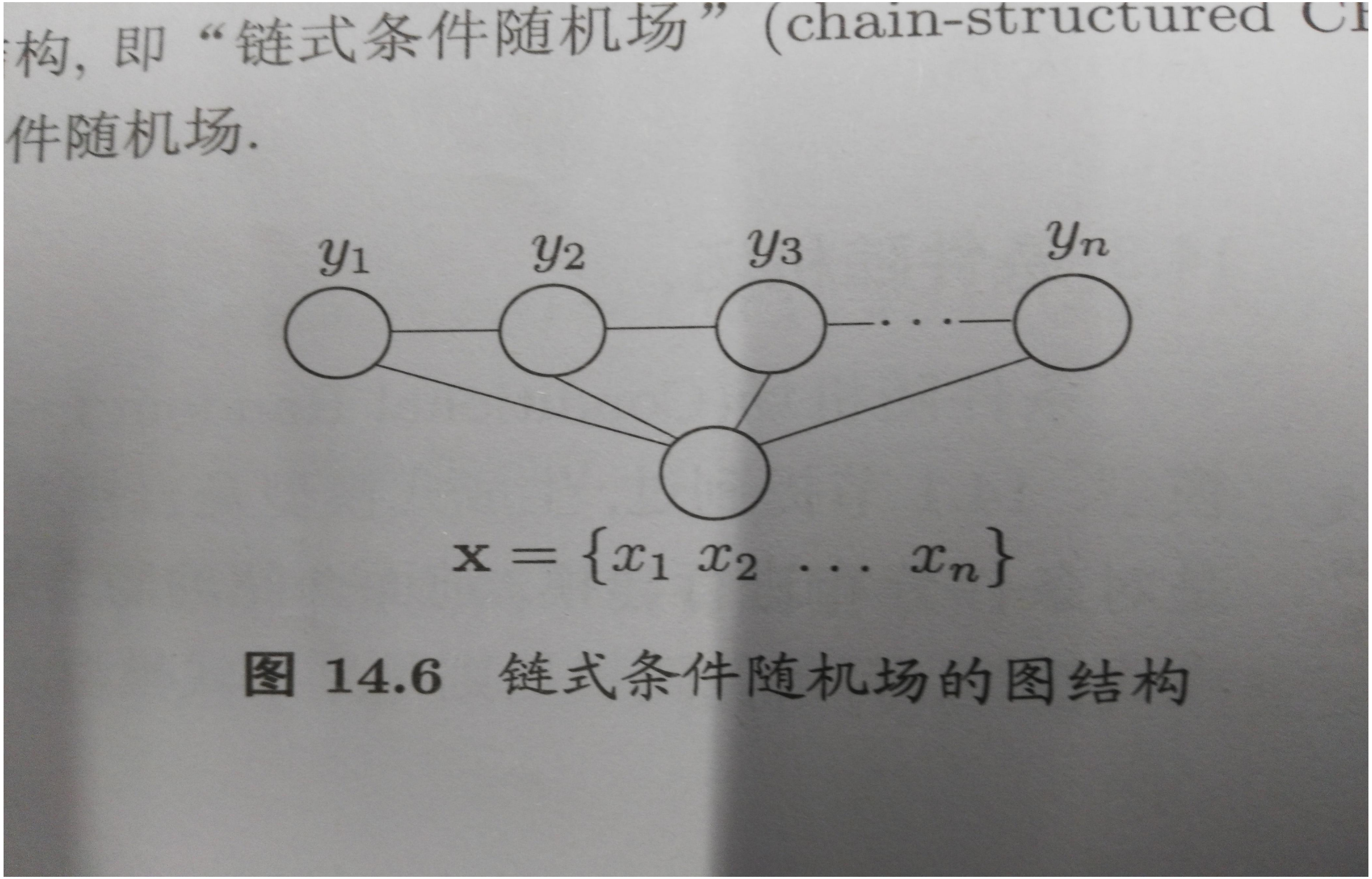
(a) 词性标注

(b) 语法分析

图 14.5 自然语言处理中的词性标注和语法分析任务

令 $G = \langle V, E\rangle$ 表示结点与标记变量 $\mathbf{y}$ 中元素

CRF



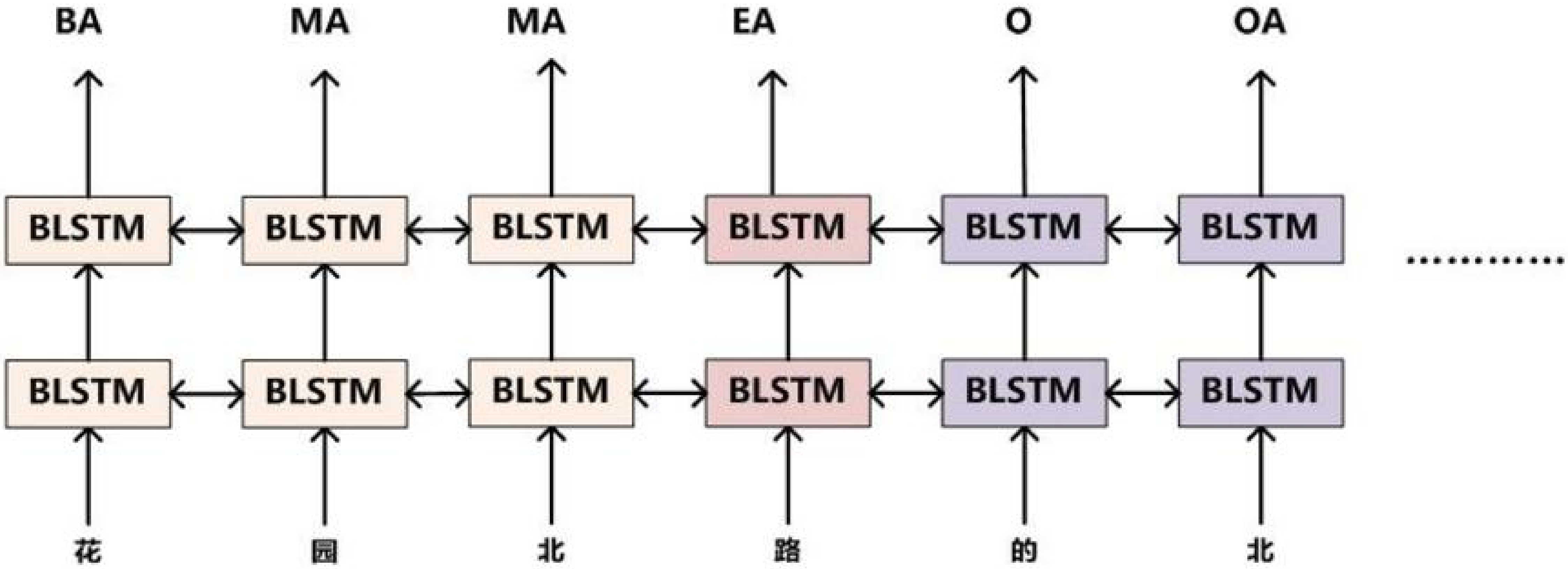
构, 即“链式条件随机场”(chain-structured C

件随机场.

y_1
y_2
y_3
y_n
$\mathbf{x} = \{x_1\ x_2\ \ldots\ x_n\}$

图 14.6 链式条件随机场的图结构

Bilstm


BA
MA
MA
EA
O
OA
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
BLSTM
…………
花
园
北
路
的
北

图模型和DL的结合

- RBM

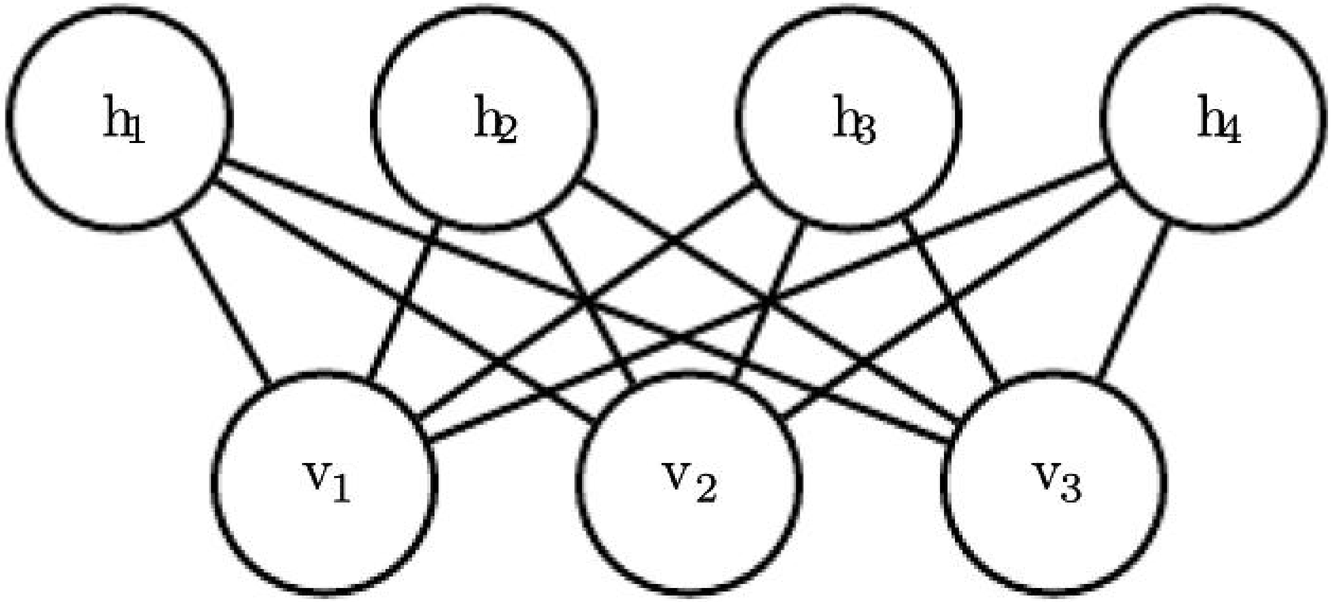


Figure 16.14: An RBM drawn as a Markov network.

- 谢谢！