第四讲 自然语言理解

—文本与音频内容分析与理解

刘静 研究员
jliu@nlpr.ia.ac.cn

中科院自动化研究所
模式识别国家重点实验室

《多媒体分析与理解》课程设置

- 第 1 讲：多媒体内容概述
- 第 2 讲：多媒体特征提取与表示
- 第 3 讲：基于深度神经网络的特征学习
- 第4-5讲：多媒体内容分析与理解（文本/音频/图像）
- 第 6 讲：视频内容分析与理解
- 第 7 讲： 研讨课
- 第 8 讲： 跨媒体分析与理解
- 第9-10讲：多媒体信息检索
- 第 11 讲：多媒体内容推荐
- 第 12 讲：课程作业
- 第 13 讲：多媒体应用实例

多媒体内容分析与理解

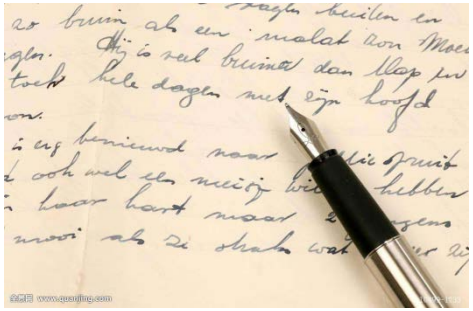
- 文本与音频（一维序列）– 自然语言理解
- 图像（二维空间）– 视觉内容语义理解
- 视频（三维时空）– 时序关联的视觉内容语义理解
- 跨媒体（多维复杂关联）– 各种模态、媒体综合性语义理解

本讲主要内容

- 自然语言理解概述
- 自然语言理解技术
- 两个重要模型及其应用
  - 语言模型
  - 隐马尔可夫模型

何为自然语言

- 自然语言是人类特有的用来表达情感和交流思想的工具
- 文字和语音是构成自然语言的两个基本属性
  - 文字是记录语言的书写符号
  - 语音是用来表示语言的声音符号



自然语言理解的定义

- 自然语言处理(Natural Language Processing, NLP)也称自然语言理解(Natural language understanding, NLU)是人工智能的分支学科，研究如何处理与运用人类语言，让计算机读懂人类语言的一门学科（Wiki）

自然语言理解：针对以文字和语音为表达形式的数据，让计算机能够分析、处理、理解人类语言的学科。

自然语言理解的研究层次

语言的基本单位：词素、词、词组、句子、句子群、段落、文章

- **形态学 Morphology**
- 是语言学的一个分支，主要研究词的内部结构
- 研究词如何由有意义的词素构成？

词素（morphemes）→ 词（word）

↓

词根(词干)、前缀、后缀

ab-norm-al　　老+虎 = 老虎

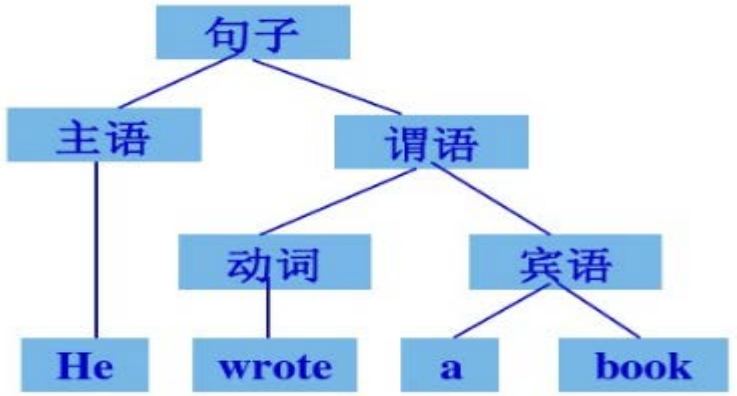
uni-versi-ty　　图+书+馆 = 图书馆

自然语言理解的研究层次

语言的基本单位：词素、词、词组、句子、句子群、段落、文章

- **语法学 Syntax**
- 研究句子成分之间的相互关系和组成句子序列的规则，包括词和短语在语句中的作用等
- 关注语言表达中各句子成分的组织形式

句子是由各种不同句子成分组成，这些成分可以是单词、词组或从句。句子成分还可以按其作用分为主、谓、宾、补、定、状、表等。



自然语言理解的研究层次

语言的基本单位：词素、词、词组、句子、句子群、段落、文章

- **语义学 Semantics**
- 是研究语言各级单位的意义的学科，探明各级单位所指对象之间的关系，从而指导人们的言语活动
  - 关注某个语言单位到底说了什么
  - 同义词、反义词、多义词

自然语言理解的研究层次

语言的基本单位：词素、词、词组、句子、句子群、段落、文章

- **语用学 Pragmatics**
- 从语言使用者角度，研究在不同上下文中的语句的应用，以及上下文对语句理解所产生的影响
  - 关注在特定上下文中为什么要说这句话

Q：看看鱼怎么样了？
A：我刚才翻了一下，两面金黄

Q：看看鱼怎么样了？
A：在鱼缸里活蹦乱跳的

自然语言理解技术（应用角度）

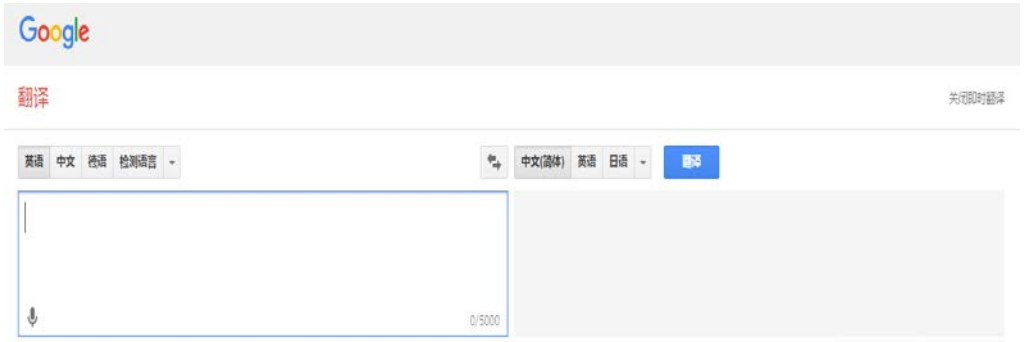
- 机器翻译
- 自动文摘
- 信息检索
- 文档分类
- 问答系统
- 语音翻译
- 语音识别
- 语音合成
- 说话人识别
- ……

文本

音频

机器翻译

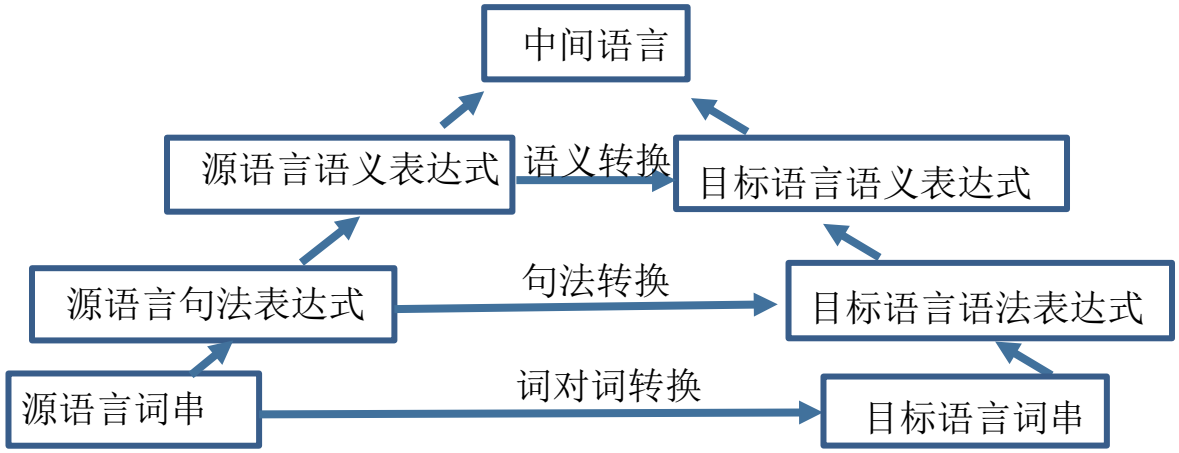
- 基本定义：由计算机程序将文字或演说从一种自然语言翻译成另一种自然语言。



- 解决方案
  - 基于规则的方法
  - 基于统计的方法
  - 基于深度神经网络的方法

机器翻译-基于规则的方法

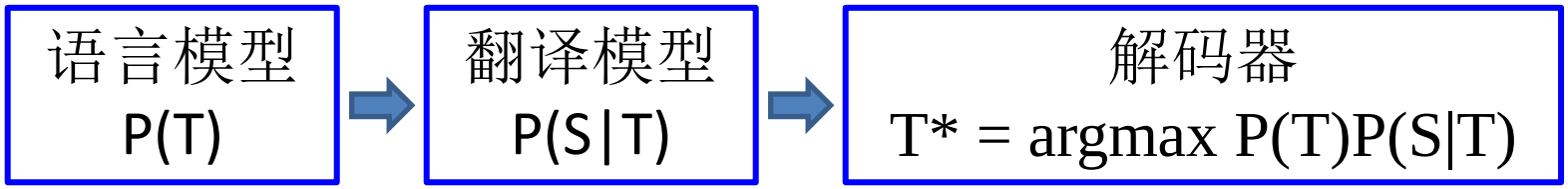
- 建立大量的规则，在翻译过程中计算机基于这些规则，进行是与否的二则判断，完成分析、转换和生成的过程，由此获得翻译的答案。
  - 直接翻译法
  - 转换翻译法
  - 中间语言法



- 缺点：建立和维护规则库依赖人力，规则的覆盖性差，主观因素大

机器翻译-基于统计的方法

- 通过对大量的平行语料进行统计分析，构建统计翻译模型，进而使用此模型进行翻译
- 噪声信道模型：翻译系统被看作噪声信道，对于一个观察到的信道输出语句S，寻找一个最大可能的信道输入语句T，即求解T是P(T|S)最大

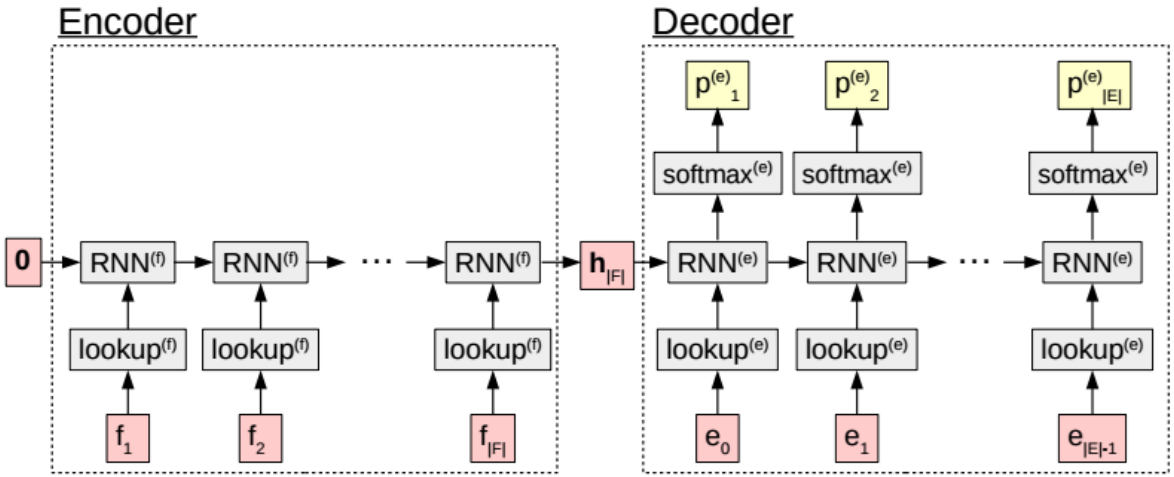


- **优点**：无须人工建立规则，开发周期短
- **缺点**：受限于语料库规模和质量，数据稀疏问题严重，时空开销大，鲁棒性差

机器翻译–基于深度学习的方法

- 一种语言的句子被向量化之后，转化为计算机可以“理解”的表示形式，再经过多层复杂的传导运算，生成另一种语言的译文。
  - 循环网络语言模型
  - 编码解码机制
  - 注意力机制

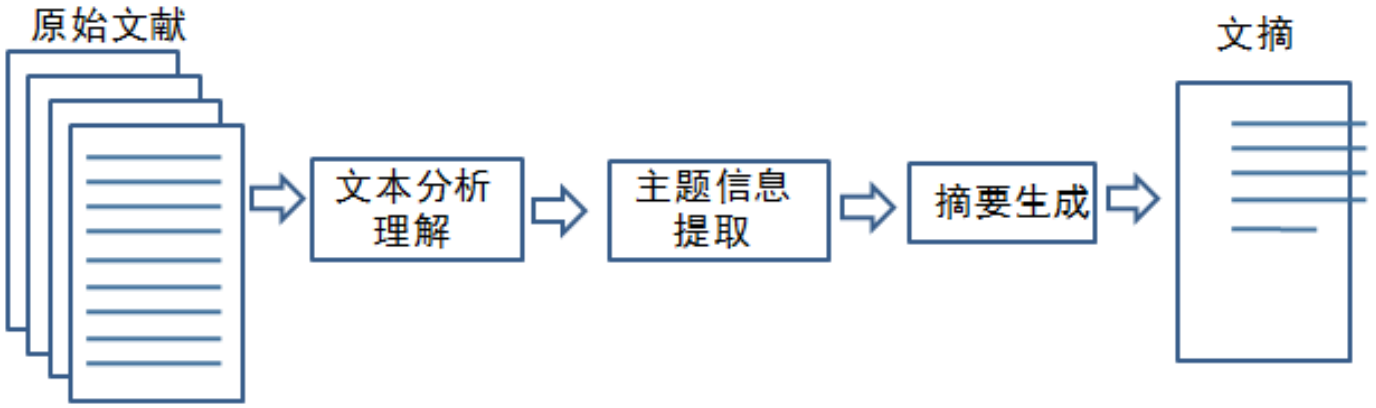
优点：端到端训练的序列学习模型，训练直接、性能鲁棒。



A computation graph of the encoder-decoder model.

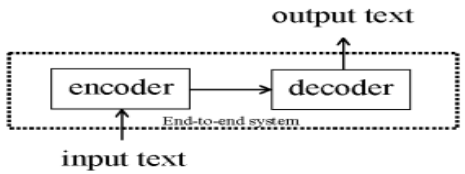
自动文摘

- 自动实现文本分析、内容归纳和摘要生成的技术
  - 文本冗余内容的识别与处理、重要信息辨认、生成文摘的连贯性
  - 功能：指示型、报道性、评论型
  - 文档数量：单文档与多文档
  - 语言种类：单语言与跨语言
  - 与原文关系：摘录型与理解型

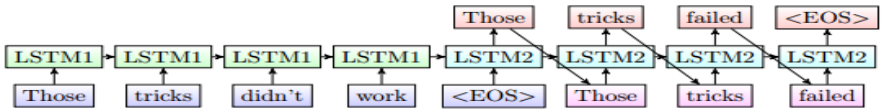


自动文摘

- **摘录型**：从原文中选择一些重要信息的句子，不加修饰构成文摘
  - 基于统计的方法：词频，标题，位置，句法结构，线索词，指示性短语
  - 基于图模型的方法：PageRank
- **理解型**：通过自然语言处理技术理解原文，生成文摘
  - 基于语义分析：语法-语义分析-信息提取-生成文摘
  - 基于RNN序列生成的方式



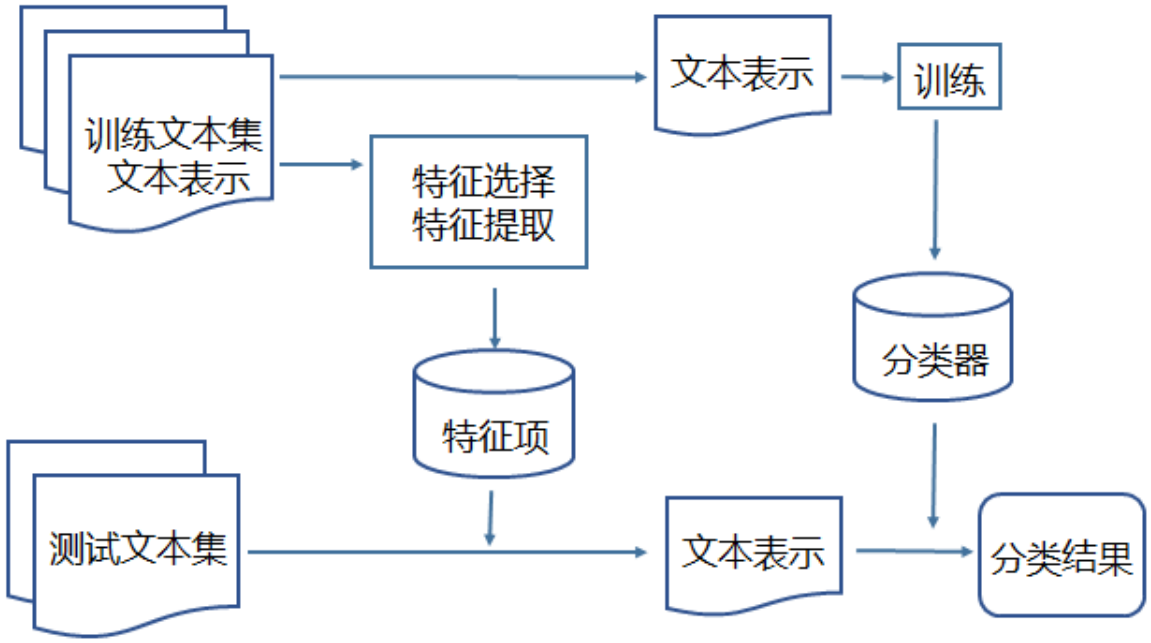
(a) General architecture



(b) An RNN-based instance for sequence-to-sequence transduction

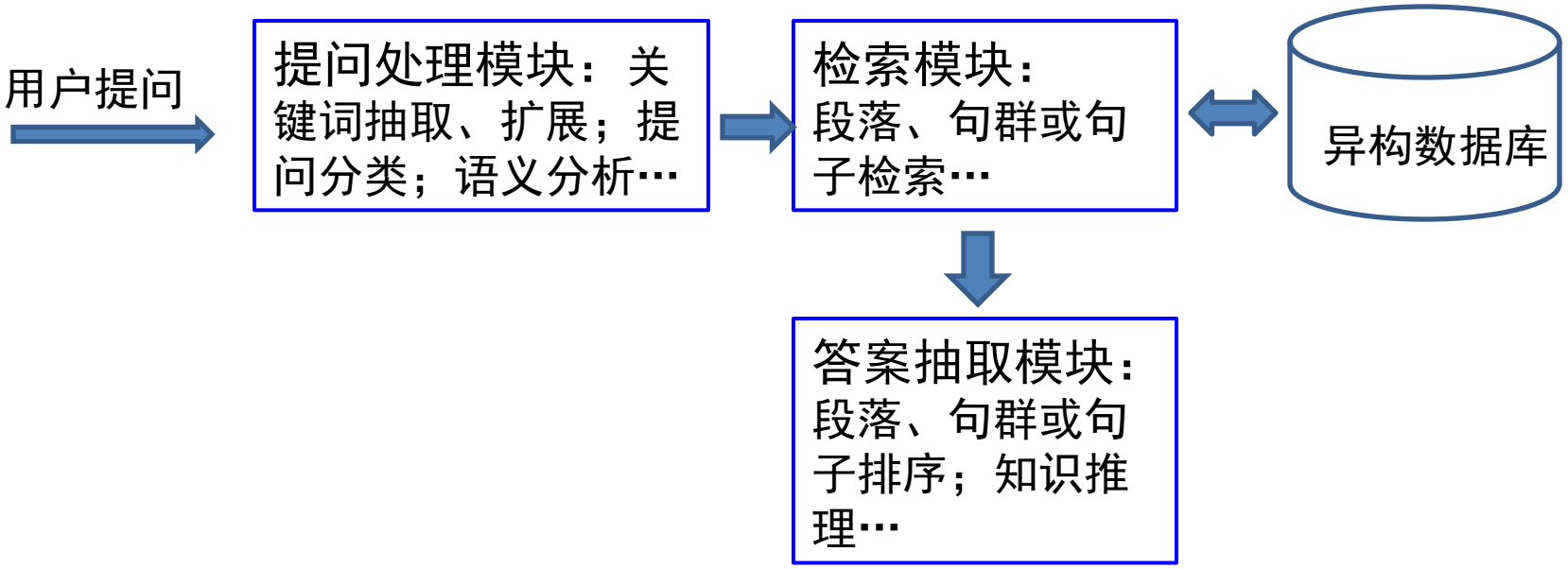
文档分类

- 文档分类是指根据预先定义的主题类别，按照一定的规则将文档集合中未知类别的文档自动确定一个或几个类别的过程
  - 主题分类、情感分类…



问答系统(question and answering, QA)

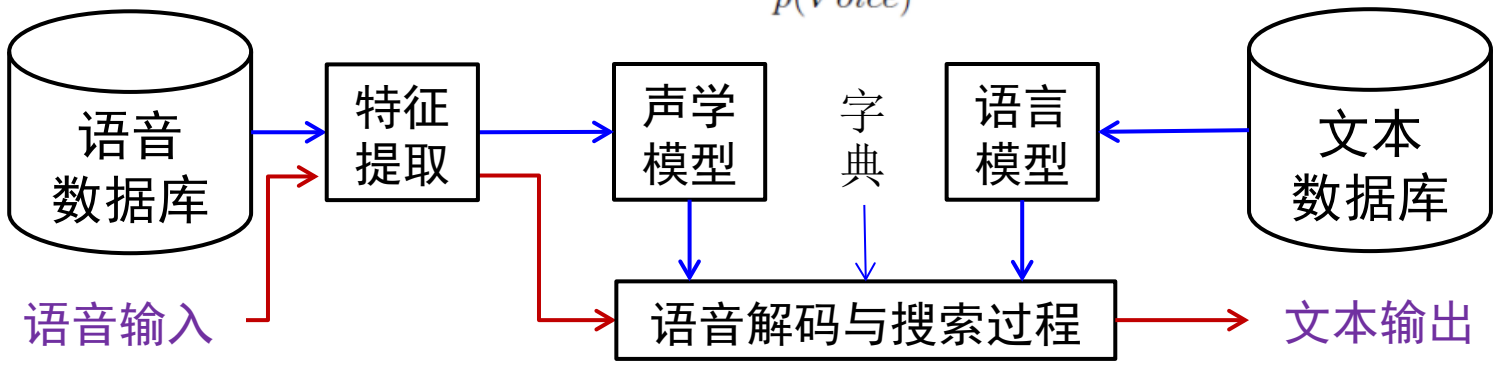
- 信息检索系统的一种高级形式，它能用准确、简洁的自然语言回答用户用自然语言提出的问题
- 是集知识表示、信息检索、语义理解、智能推理于一身的新一代搜索引擎



语音识别

- 定义：通过识别和理解过程把语音信号转变为相应的文本或命令。
- 目的：与机器进行语音交流，让机器明白你说什么。
- 基本实现：作为一个机器学习与模式识别的具体应用，语音识别同样遵守与一般机器学习任务相似的大框架，主要包括特征提取，模型训练，模式匹配与分类

$$p(Text|Voice) = \frac{p(Voice|Text) \cdot p(Text)}{p(Voice)},$$



语音翻译

- 从一种语言的语音到另一种语言的语音自动翻译的过程
- 三个技术模块：语音识别、机器翻译、语音合成
- 主要特点：
  - 同声即时翻译
  - 口语翻译：非规范、口语化（重复、省略、颠倒、修正、口头语） 例：不 不 现在先不订 过两天 噢 还是明天再订吧
  - 鲁棒的语音识别：同音字、说话人声调、语速、发音特点、环境噪音
  - 个性化的语音合成：说话人的情感、个人特征（性别、年龄、音色）

自然语言理解的难点

- 自然语言的多样性
  - 自然语言是可以组合的，字到词，词到短语，短语到从句、句子，句子到篇章，这种组合性使得同一个意思可有不同的表达方式

我要听大王叫我来巡山
给我播大王叫我来巡山
我想听歌，大王叫我来巡山
放首歌大王叫我来巡山

- 语言的上下文
  - 缺少上下文的语境约束，语言会有很大的歧义性

A：来首歌听
B：请问你想听什么歌？

A：我要去拉萨 – 订票？听歌？景点？

自然语言理解的难点

- 自然语言的鲁棒性
  - 语言在输入的过程中，尤其是通过语音识别转录过来的文本，会存在多字、少字、错字、噪音等

错字：大王叫我来新山
多字：大王叫让我来巡山
少字：大王叫我巡山
别称：熊大熊二（指熊出没）
不连贯：我要看那个恩花千骨
噪音：全家只有大王叫我去巡山咯

- 语言的知识依赖
  - 语言是对世界的符号化描述，语言天然连接着知识

大鸭梨：除了表示水果，还可以表示餐厅名
七天：除了表示时间，还可以表示酒店名

自然语言理解技术（应用角度）

- 机器翻译
- 自动文摘
- 信息检索
- 文档分类
- 问答系统
- 语音翻译
- 语音识别
- 语音合成
- 说话人识别
- ……

文本

音频

文本表示/自然语言输出

↓

语言模型

数据对象：时间或者空间上一维连续的序列信号

↓

隐马尔可夫模型
深度循环网络

*深度循环网络（RNN、LSTM）：可进行Seq2Seq的端到端学习

本讲主要内容

- 自然语言理解概述
- 自然语言理解技术
- 两个重要模型及其应用
  - 语言模型
  - 隐马尔可夫模型

语言模型

- 基本定义：
  - 自然语言中标记序列的概率分布，标记可为词、字符或字节等，记为：$\mathrm{P}(w_1, w_2, \ldots, w_n)$
  - 衡量词序列（句子或文章）符合自然语言表达的程度
- 应用领域：
  - 语音识别
    - ✓ I went to a party vs. Eye went two a party
  - 机器翻译
    - ✓ 王刚出现在电视上: Wang Gang appeared on TV vs. In Wang Gang appeared TV
  - 上下文敏感的拼写检查
    - ✓ appear on TV vs. appear of TV

语言模型

- 理想化建模：
  - 给定词序列W = $\{w_1, w_2, \ldots, w_T\}$, 计算p(W)
  - 根据链式法则：

$$p(W) = p(w_1)p(w_2|w_1)\cdots p(w_T|w_1, w_2, \ldots, w_{T-1})$$

  - 只是理想公式，参数空间($T|V|^T$)
- n元语言模型 n-gram，简化为 $O(T|V|^n)$
  - 每个词只与其前n-1个词有关，可近似认为n-1阶马尔可夫链（n>2）

$$p(W) = p(w_1, w_2, \ldots, w_n)\prod_{t=1}^{T+1} p(w_t|w_{t-n+1}, w_{t-n+2}, \ldots, w_{t-1})$$

  - 通常, unigram(n=1),bigram(n=2)和trigram(n=3)

语言模型 – n-gram

$$\mathrm{p}(W) = p(w_1, w_2, \dots, w_n) \prod_{t=1}^{T+1} \mathrm{p}(w_t | w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1})$$

- 给定训练集，基于最大似然估计容易求得

$$\mathrm{p}(w_t | w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1}) = \frac{P_n(w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1}, w_{t-1})}{P_{n-1}(w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1})}$$

- 参数空间太大$|V|^n$，$|V|$为词典大小
  - 引入词类别，基于类的语言模型
- $P_n = 0$或$P_{n-1} = 0$,将造成无法计算
  - 平滑技术：非零化

$$\hat{P}(w_t | w_{t-1}, w_{t-2}) = \alpha_0(q_t)p_0 + \alpha_1(q_t)p_1(w_t) + \alpha_2(q_t)p_2(w_t | w_{t-1}) + \alpha_3(q_t)p_3(w_t | w_{t-1}, w_{t-2})$$

  - 回退方法：查找低阶n-gram

语言模型

- 采用机器学习的方法来进行构造语言模型
- 最大似然法

$$\mathcal{L} = \sum_{w \in \mathcal{C}} \log p(w|Context(w)),$$

$$p(w|Context(w)) = F(w, Context(w), \theta),$$

- 给定语料库C，针对参数集合 θ 进行优化问题求解
- 与n-gram相比，不需要事先计算所有条件概率
- 如何选择合适的函数F？
  - 神经网络模型

语言模型–基于神经网络

- 2001年由Bengio等人[*]提出基于四层神经网络的n-gram语言模型

学习一个函数

$$f(w_t, \ldots, w_{t-n+1}) = p(w_t | w_{t-n+1}^{t-1})$$

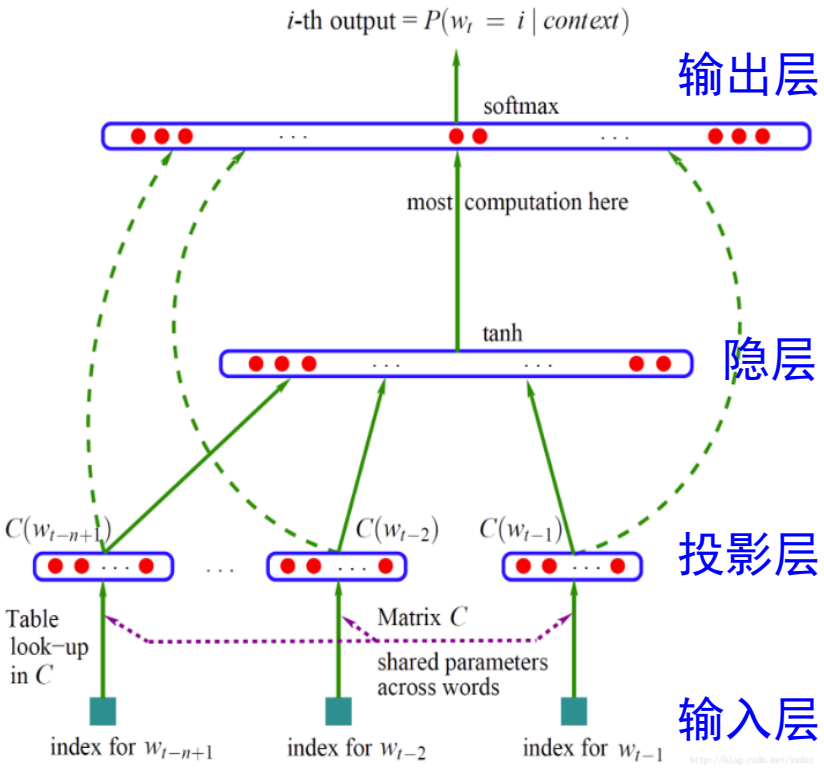
(1)从词表中任意元素到实向量
C(i)∈ R^m,C为|V|×C的转换矩阵

(2)基于词向量表示C的概率函数：
输入为x={C(w_{t-n+1}),…, C(w_{t-1})},
输出为给定{$w_{t-n+1}, \ldots, w_{t-1}$}输出$w_t$的条件概率

$$\hat{P}(w_t | w_{t-1}, \cdots w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

$$\boldsymbol{y = b + Wx + U\,tanh(d + Hx)}$$

$\boldsymbol{y \in R^{|V|}}$



[*]Y. Bengio et. al., A Neural Probabilistic Language Model, NIPS 2001, 932-938

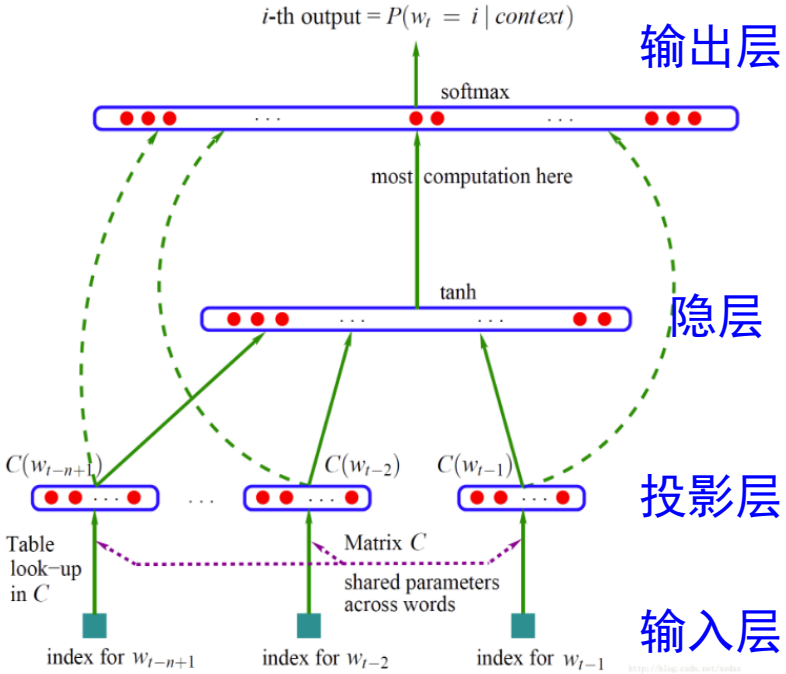
语言模型–基于神经网络

Softmax 函数输出 $$\hat{P}(w_t|w_{t-1},\cdots w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

$$\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\,\boldsymbol{tanh}(\boldsymbol{d} + \boldsymbol{H}\boldsymbol{x})$$

参数包括：$\theta = (b, d, W, U, H, C)$　C为|V|×C的转换矩阵

b	$\|V\|$	输出偏置
d	h	隐藏层偏置
U	$\|V\|*h$	隐藏层到输出层的权重
W	$\|V\|*(n-1)m$	词特征层到输出层的权重
H	$h*(n-1)m$	隐藏层的权重



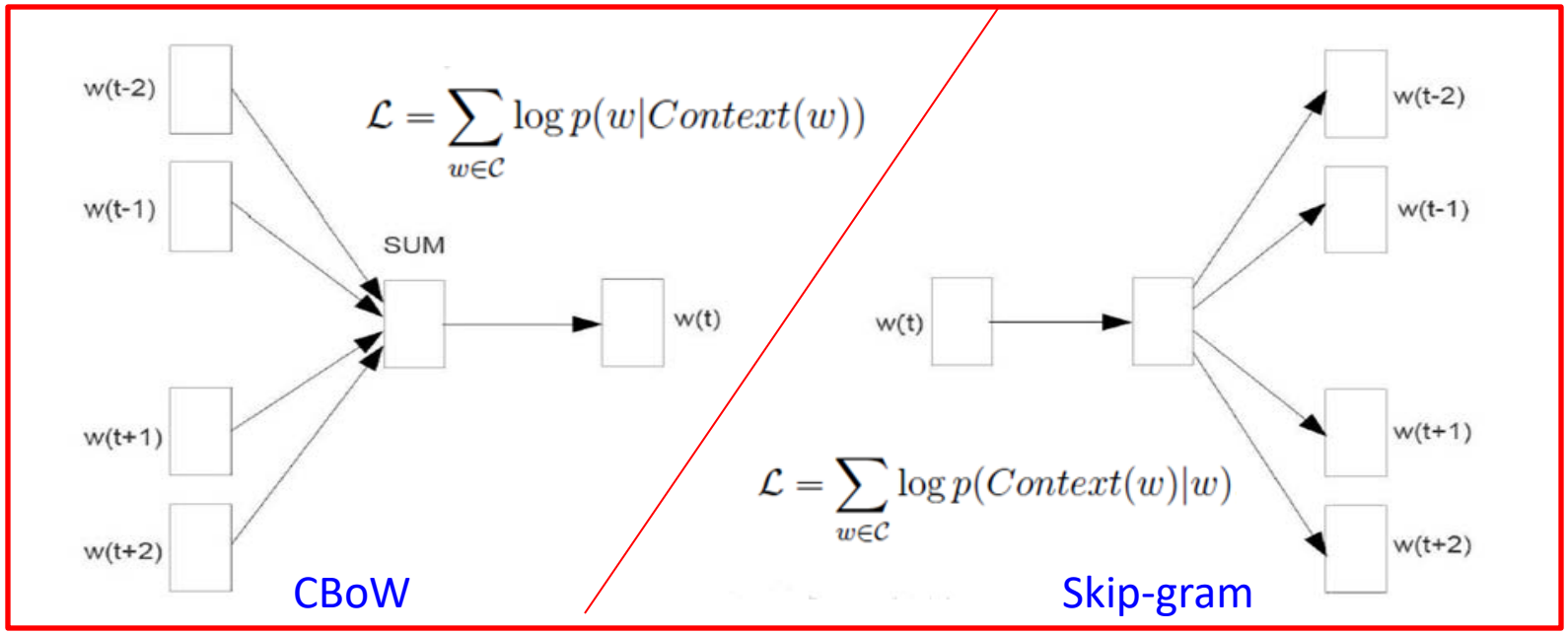
- ✓ 词向量作为网络学习的副产品，可用于度量词语之间的相似性
- ✓ 输出概率非零，自带平滑
- x 输出层计算须在所有|V|输出上归一化，计算成本过高
- x 只能处理定长序列

语言模型–词向量

- One-hot表示
  - 向量长度为词典大小，每个向量只有一个维度为1，其余为0
  - 缺点：维数灾难，语料库字典非常大；任意两个不同词的距离相同，无法刻画词词相似性
- 分布式表示
  - 1986年由Hinton提出，词语的语义是通过上下文信息来确定的，即相同语境出现的词，其语义也相近
  - 将词映射到新的空间，从原始词向量的稀疏表示转变为低维空间的密集表示
  - LSA矩阵分解模型，PLSA潜在语义分析概率模型…
  - 神经网络语言模型

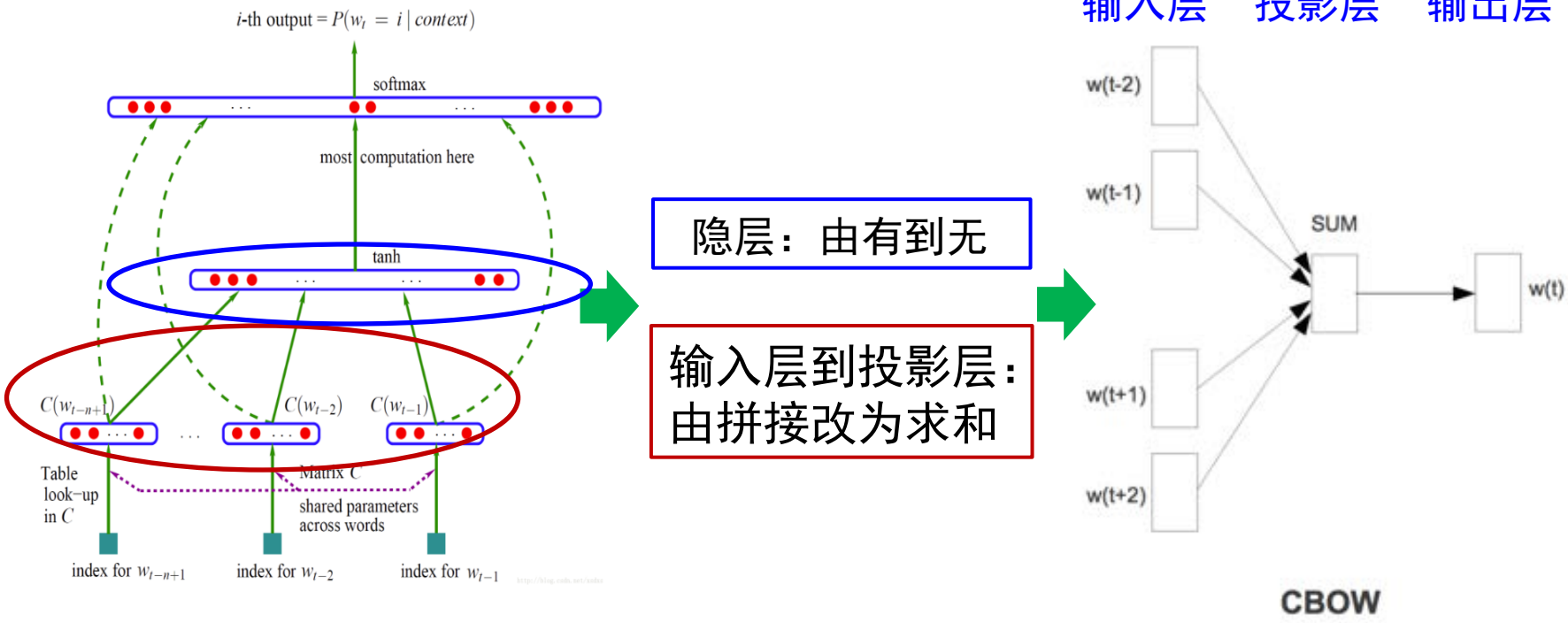
语言模型–基于神经网络

- Word2Vec
  - 2013年，Google推出的用于获取词向量表示的工具包
  - 核心算法由 Mikolov提出，是Bengio工作的加速改进版
  - 两种模型：CBoW & Skip-gram
  - 两种加速求解：层次Softmax & 负采样



语言模型– CBoW

- CBoW- Continuous Bag-of-Words



- ✓ CBoW模型等价于一个词袋模型的向量乘以一个嵌入矩阵，得到一个连续的词向量
- ✓ 从context对target word的预测中学习到词向量的表达

加速求解一：层次Softmax

- 以CBoW为例

✓ 层次Softmax本质是用基于霍夫曼编码的二叉树层级结构代替扁平化的Softmax层。

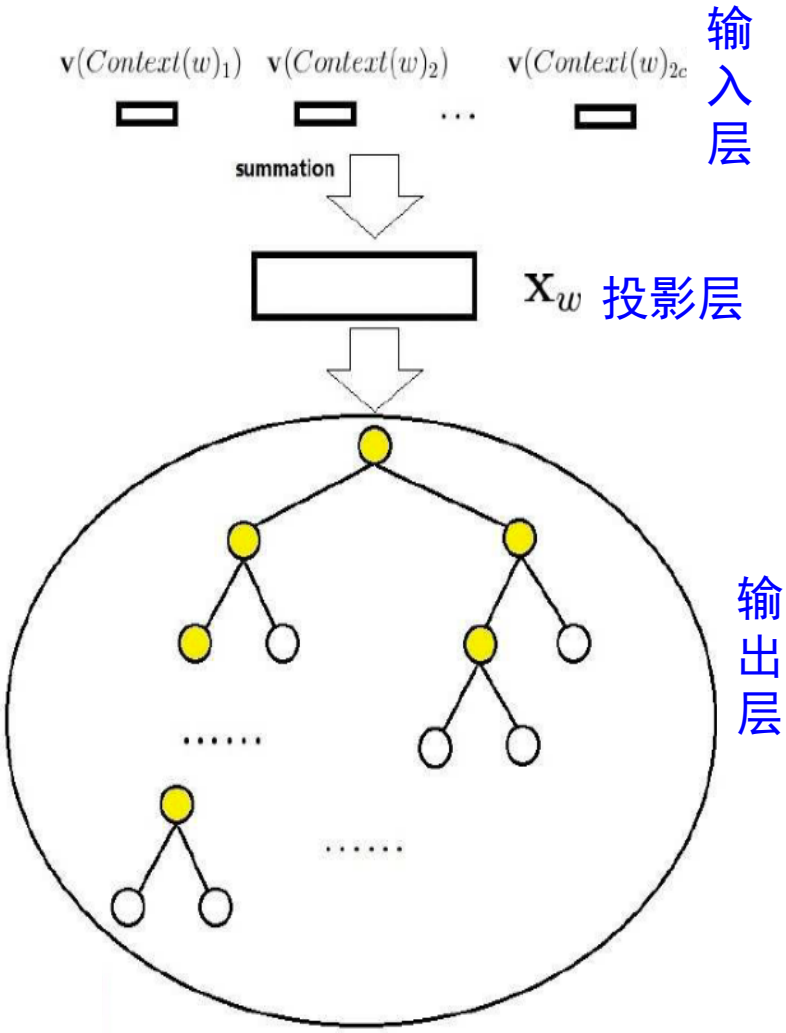
✓ 每个叶子节点表示一个词语，则给定上下文的每个词条件概率值的计算过程可被拆解为最多log2|V|个概率值的计算

$$p(v|context) = \prod_{i=1}^{m} p(b_i(v)|b_1(v), \ldots, b_{i-1}(v), context)$$

✓ 遍历整颗树的概率和常为1

✓ 每一层条件概率对应一个二分类问题，可以通过逻辑回归函数去拟合

$$\sigma(\mathbf{x}_w^\top \theta) = \frac{1}{1 + e^{-\mathbf{x}_w^\top \theta}},$$

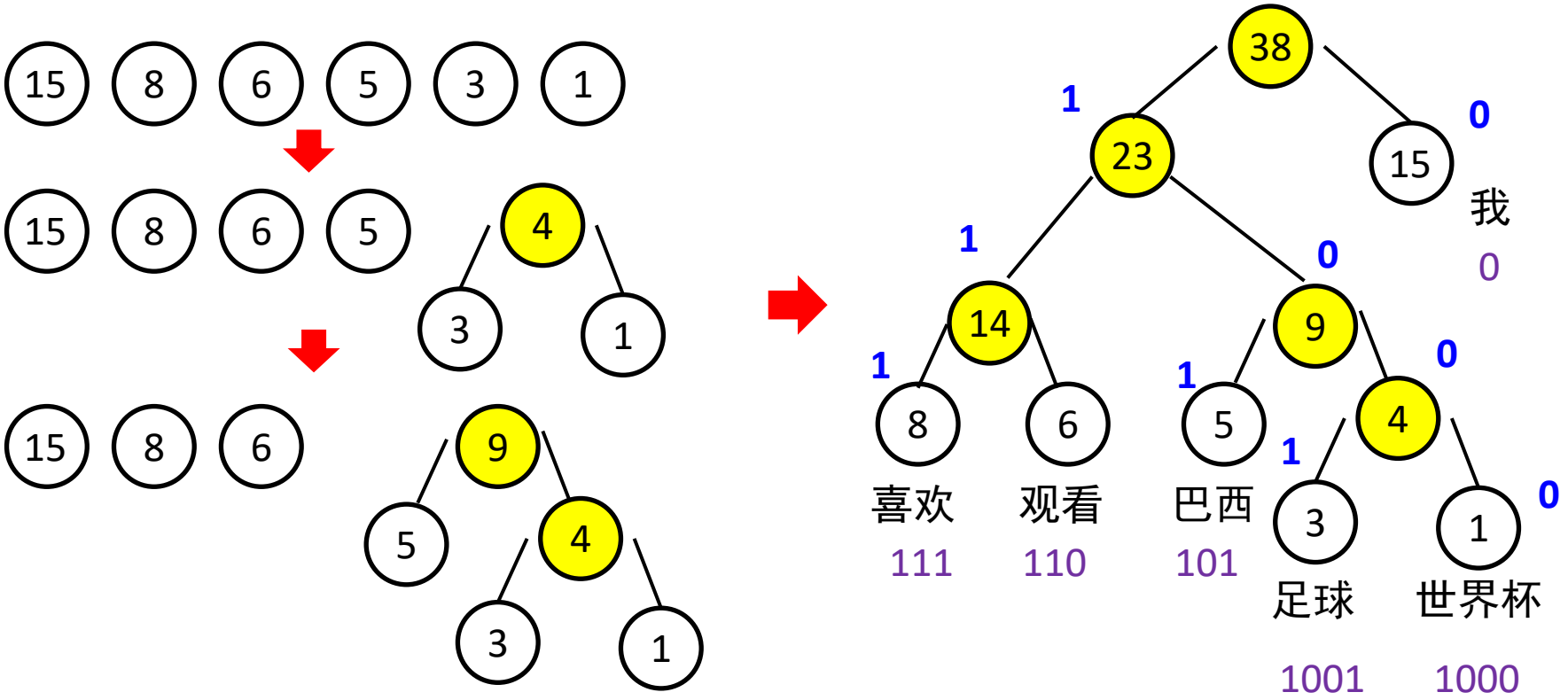


霍夫曼编码-Huffman

- 基于霍夫曼树设计的二进制编码，即霍夫曼编码
- 霍夫曼树
  - 给定n个权值作为n个叶子节点，由此构造的带权路径长度最小的二叉树
  - 从根节点到该节点之间路径长度与该节点权值的乘积
  - 权值越大离根节点越大
  - 通常，权值大的为左孩子节点，小的为右孩子节点
- 语言模型将词典中的词作为叶子节点，基于词频构造霍夫曼树
  - 计算softmax输出时，只需计算log2|V|

霍夫曼编码-Huffman

- 举例：经统计某新闻语料库中，“我”“喜欢”“观看”“巴西”“足球”“世界杯”六个词出现的次数分别为15, 8, 6, 5, 3, 1。以这6个词为叶子节点，构造Huffman树



层次Softmax计算过程举例

编码为1定义为负类，0为正类

$$Label(p_i^w) = 1 - d_i^w, i = 2, 3, \cdots, l^w.$$

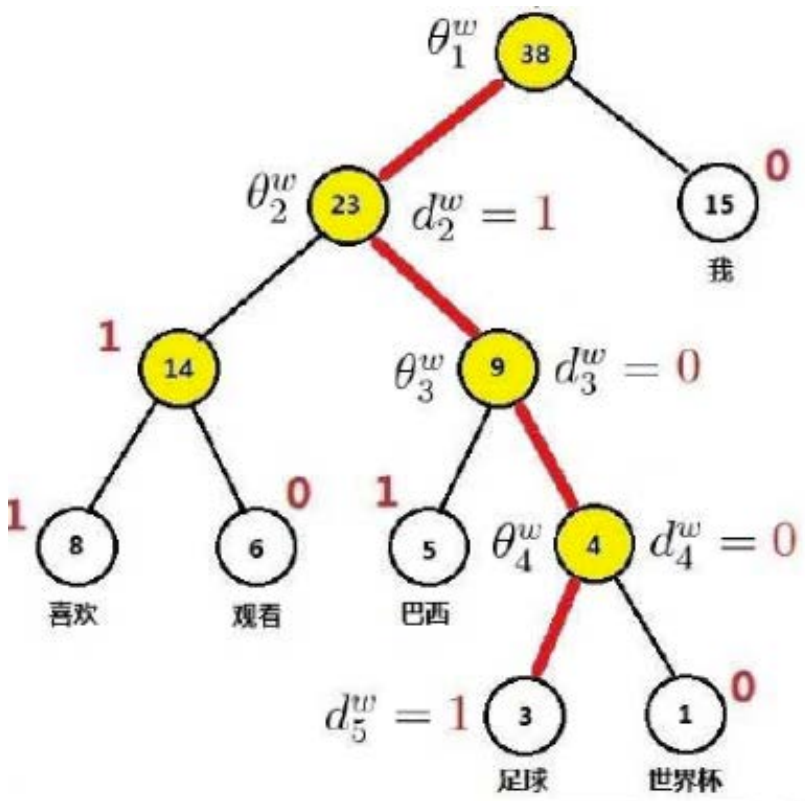
正类的概率为：$\sigma(\mathbf{x}_w^\top \theta) = \dfrac{1}{1 + e^{-\mathbf{x}_w^\top \theta}}$,

从根节点到“足球”，经历编码1001

1. 第 1 次: $p(d_2^w|\mathbf{x}_w, \theta_1^w) = 1 - \sigma(\mathbf{x}_w^\top \theta_1^w)$;
2. 第 2 次: $p(d_3^w|\mathbf{x}_w, \theta_2^w) = \sigma(\mathbf{x}_w^\top \theta_2^w)$;
3. 第 3 次: $p(d_4^w|\mathbf{x}_w, \theta_3^w) = \sigma(\mathbf{x}_w^\top \theta_3^w))$;
4. 第 4 次: $p(d_5^w|\mathbf{x}_w, \theta_4^w) = 1 - \sigma(\mathbf{x}_w^\top \theta_4^w)$,

$$p(\text{足球}|Contex(\text{足球})) = \prod_{j=2}^{5} p(d_j^w|\mathbf{x}_w, \theta_{j-1}^w).$$

计算概率次数最多为log2|V|，而非传统的|V|



$\theta_1^w, \theta_2^w, \cdots, \theta_{l^w-1}^w \in \mathbb{R}^m$

非叶子节点对应的向量，为模型参数

加速求解二：负采样

- 用来提高训练速度，并改善词向量质量
- 相比层次Softmax，不再使用Huffman树，而是通过随机负采样方式计算概率
- 给定*context(w)*,*w*为正样本，其他词为负样本，负样本太多，因此采用采样方式

$$g(w) = \prod_{u \in \{w\} \cup NEG(w)} p(u|Context(w)),$$

进一步 $$g(w) = \sigma(\mathbf{x}_w^\top \theta^w) \prod_{u \in NEG(w)} \left[1 - \sigma(\mathbf{x}_w^\top \theta^u)\right],$$

增大正样本的概率，同时降低负样本的概率

Word2Vec的应用

- 基于特定语料库训练得到的Word2Vec表示
  - 作为初始文本内容表示，用于文本内容分析任务（文本分类、机器翻译、命名实体识别等）
  - 还可扩展为Sentence2Vec表示
- 词的语义表达空间 -Word Pair Relationships
  - V(King)-V(Man)+V(Woman)=V(Queen)
  - V(Paris)-V(France)+V(Italy)=V(Rome)
  - 知识表达与知识库构建

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack

Skip-gram model trained on 783M words with 300 dimensionality

Word2Vec的应用

- 机器翻译
  - 首先从大量的单语种语料中学习到每种语言的word2vec表达
  - 利用较小的双语语料库学习两种语言word2vec表达的线性映射W，其损失函数构造为：

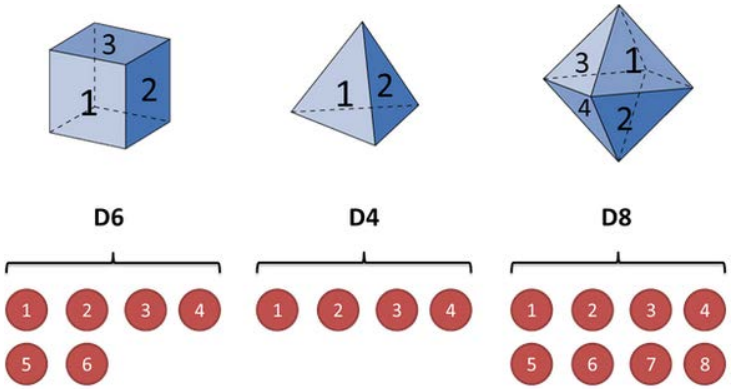
$$J(W) = \sum_{i=1}^{n} ||Wx_i - z_i||^2$$

X源语言，Z目标语言

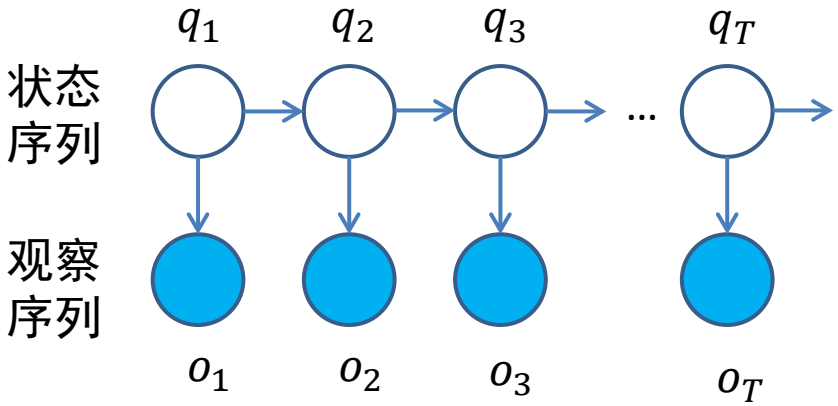
  - 翻译过程：将源语言word2vec向量通过矩阵W映射到目标语言向量空间上，再找出与投影向量距离最近的目标语言词作为翻译结果

HMM的由来

- **隐马尔可夫模型** (Hidden Markov Model) 是一个双重随机过程
  - 马尔科夫链：描述隐含状态的转移，用转移概率描述
  - 一般随机过程：描述状态与观察序列之间的关系，用观察值概率描述



隐含状态集合：{D6, D4, D8}
观察集合：{1,2,3,4,5,6,7,8}

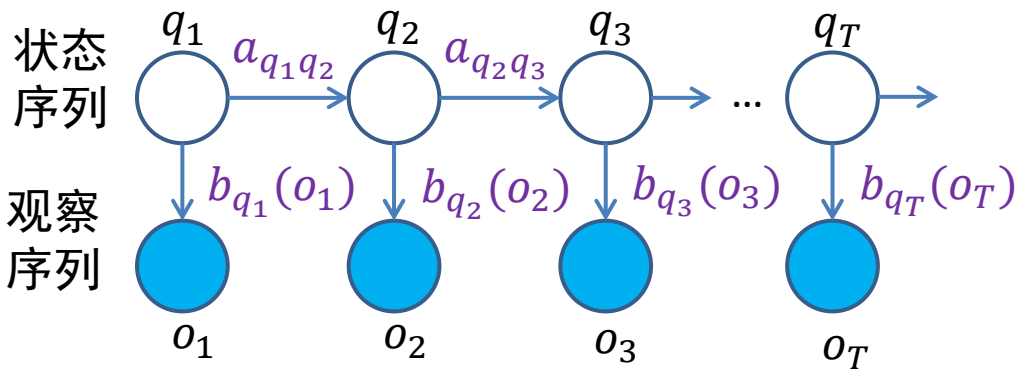


HMM图解

HMM的基本要素

- 用模型五元组$\lambda = (N, M, A, B, \pi)$来描述HMM

参数	含义	举例
N	隐含状态数目	{D4, D6, D8}
M	每个状态对应的观察值数目	{1,2,3,4,5,6,7,8}
A	与时间无关的状态转移概率	选不同骰子之间的转移概率
B	给定状态下，观察值概率分布	每个骰子可掷出的数字概率
π	初始状态空间的概率分布	初始选择某个骰子的概率



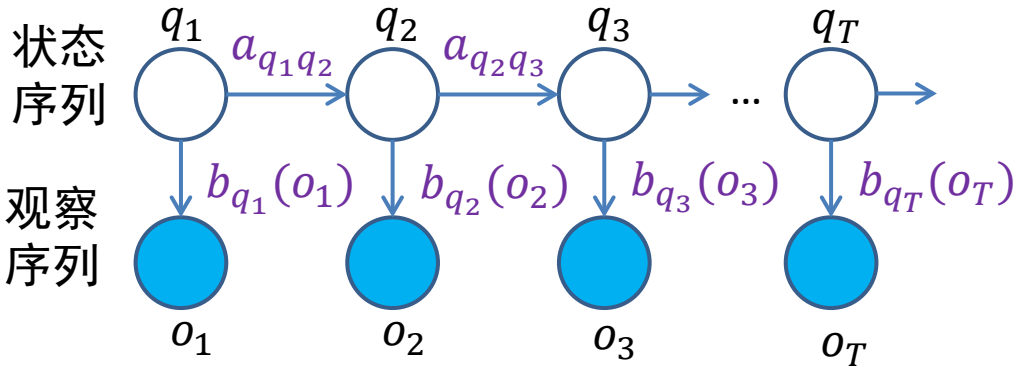
$$A = \{a_{ij} \geq 0\}, 1 \leq i, j \leq N$$
$$a_{ij} = p(q_{t+1} = S_j | q_t = S_i)$$
$$\sum_{j=1}^{N} a_{ij} = 1$$

$$B = \{b_j(k) \geq 0\}, 1 \leq j \leq \mathrm{N}$$
$$b_j(k) = p(o_t = v_k | q_t = S_j)$$
$$\sum_{k=1}^{M} b_j(k) = 1$$

$$\pi_i = p(q_1 = S_i) \geq 0, \sum_{i=1}^{N} \pi_i = 1, 1 \leq i \leq N$$

HMM的基本假设

- 齐次马尔可夫性
  - 当前状态只与前一时刻状态有关，与观测、时间和其他时刻状态无关
  - $\mathrm{P}(q_{t+1}|q_t,\dots,q_1)=\ \mathrm{P}(q_{t+1}|q_t)$
  - $\mathrm{P}\left(q_{t+1}=S_i\middle|q_t=S_j\right)=\mathrm{P}(q_{k+1}=S_i|q_k=S_j)$
- 观测输出独立性
  - 当前时刻的观察值只与该时刻的状态有关
  - $\mathrm{P}(o_1,o_2,\dots,o_T|q_1,q_2,\dots,q_T)=\prod P(o_t\,|q_t)$



HMM可解决的三类问题

- 问题1：给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$和模型$\lambda = (A, B, \pi)$，如何计算概率$p(O|\lambda)$?
  - 概率计算问题，也可理解为模型评估问题
  - 知道骰子有几种(隐含状态数量)，每种骰子是什么（转移概率），计算掷出特定结果（观察序列）的概率
- 问题2：给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$和模型$\lambda = (A, B, \pi)$，如何选择一定意义下最优的状态序列$Q = \{q_1, q_2, \ldots, q_T\}$，试其能够最为合理的解释观察序列O?
  - 解码问题
  - 知道骰子有几种，每种骰子是什么，根据掷出结果（观察序列），计算选择的骰子序列（隐状态链）
- 问题3：给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$，如何调节模型参数使得$p(O|\lambda)$最大?
  - 模型参数估计问题
  - 知道骰子有几种，观察到多次投掷的结果（观察序列），计算选择各种骰子之间的转移概率

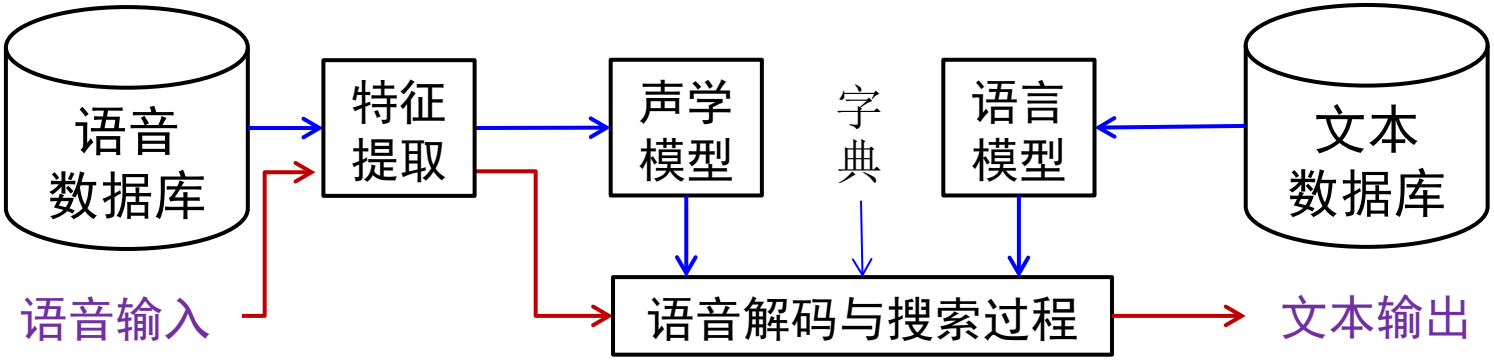
HMM的具体应用

- 中文分词与词性标注统一考虑
  - 输入中文语句 S
    - ✓ “他们两个是兄弟”
  - 观察序列 $O = \{o_1, o_2, \ldots, o_T\}$
    - ✓ “他们 两个 是 兄弟” or “他 们 两个 是 兄 弟”…
  - 状态集合{名词，动词，形容词，副词，介词…}

- **问题3：**基于大型语料估计HMM模型参数
- **问题1：**对已给定的输入S与可能的输出序列O和模型 $\lambda = (A, B, \pi)$，计算$p(O|\lambda)$，所有可能的输出序列中使概率$p(O|\lambda)$最大的解就是要找的分词结果
- **问题2：**选择最优的状态序列（词性标注），使其最好的解释观察序列（分词结果）

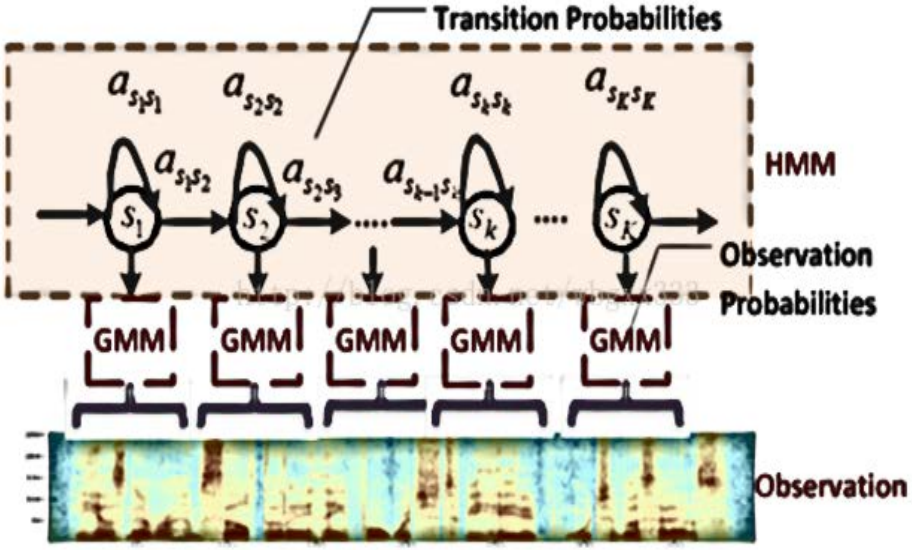
HMM的具体应用

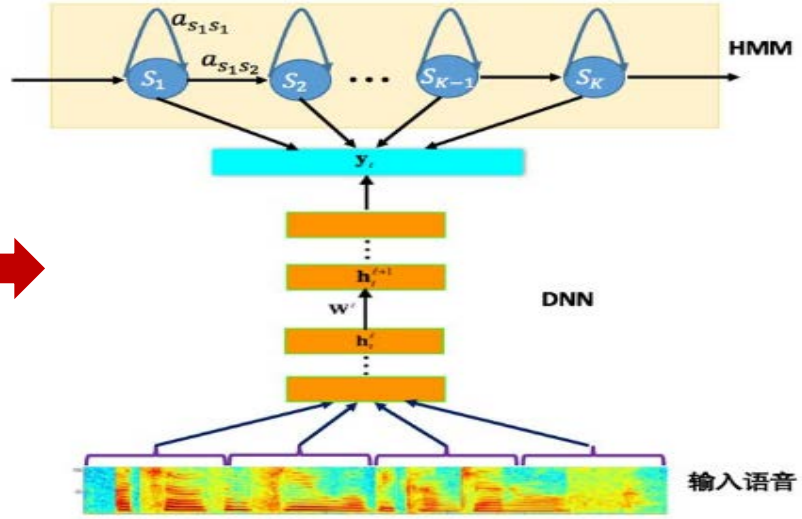
- 语音识别
  - 1988年，李开复提出第一个非特定人连续语音识别系统
  - 声学模型（GMM-HMM）+ 语言模型（N-gram）
  - 声学模型：用HMM模型对语音的时序进行建模，用GMM对语音的观察概率进行建模
  - 语言模型：语言建模能有效结合语法和语义知识，描述词之间的内在关系，从而提高识别率，减少搜索范围



HMM的具体应用

- 语音识别 - 声学模型
  - 用HMM模型对语音的时序进行建模，用GMM对语音的观察概率进行建模
  - 进一步用DNN/RNN代替GMM
  - 直到2013年才出现脱离HMM，直接端到端的深度学习语音识别系统





HMM三个基本问题的解决方案

- 前后向算法解决问题1
  - 给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$和模型$\lambda = (A, B, \pi)$，如何计算概率$p(O|\lambda)$?
- 维特比（Viterbi）算法解决问题2
  - 给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$和模型$\lambda = (A, B, \pi)$，如何选择一定意义下最优的状态序列$Q = \{q_1, q_2, \ldots, q_T\}$，试其能够最为合理的解释观察序列O？
- 最大似然法解决问题3
  - 模型参数估计：给定观察序列$O = \{o_1, o_2, \ldots, o_T\}$，如何调节模型参数使得$p(O|\lambda)$最大？

解决问题1

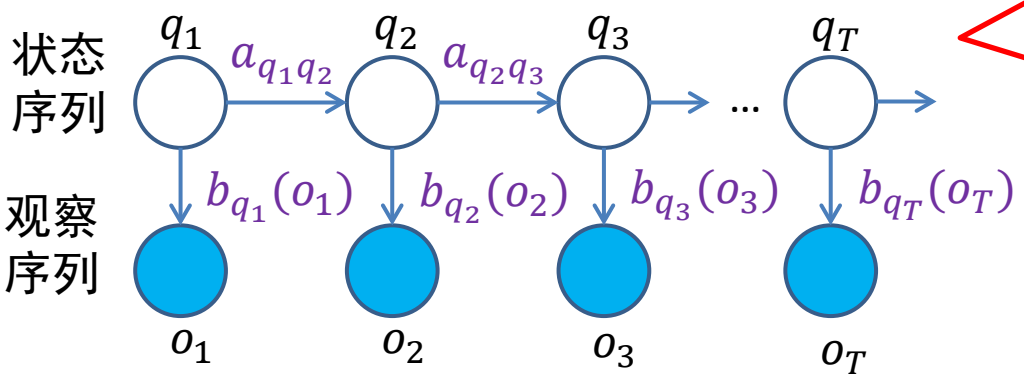
- 问题1（评估问题）：计算观察序列概率
  - 所有可能隐状态序列情况下的观察序列概率p(O|λ)

$$p(O|\lambda) = \sum_Q p(O,Q|\lambda) = \sum_Q p(Q|\lambda)\, p(O|Q,\lambda)$$

$$p(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T}$$

$$p(O|Q,\lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \ldots b_{q_T}(o_T)$$

N个状态T个时刻，所有隐状态序列组合数为N^T



解决问题1：前向算法

- 基本思想：定义前向概率

$$\alpha_t(j) = p(o_1, o_2, \dots, o_t, q_t = S_j|\lambda)$$

$$\alpha_{t+1}(j) = p(o_1, o_2, \dots, o_t, o_{t+1}, q_{t+1} = S_j)$$

$$= p(o_1, o_2, \dots, o_t, o_{t+1}|q_{t+1} = S_j)p(q_{t+1} = S_j)$$

$$= p(o_1, o_2, \dots, o_t|q_{t+1} = S_j)p(o_{t+1}|q_{t+1} = S_j)p(q_{t+1} = S_j)$$

$$= p(o_1, o_2, \dots, o_t, q_{t+1} = S_j)p(o_{t+1}|q_{t+1} = S_j)$$

$$= \sum_i p(o_1, o_2, \dots, o_t, q_t = S_i, q_{t+1} = S_j)\, p(o_{t+1}|q_{t+1} = S_j)$$

$$= \sum_i p(o_1, o_2, \dots, o_t, q_t = S_i)p(q_{t+1} = S_j|q_t = S_i)\, p(o_{t+1}|q_{t+1} = S_j)$$

$$= \sum_i \alpha_t(i)a_{ij}b_j(o_{t+1})$$

⬅ 递归计算公式

解决问题1：前向算法

- 基本思想：定义前向概率

$$\alpha_t(j) = p(o_1, o_2, \ldots, o_t, q_t = S_j|\lambda)$$

初始化：$\alpha_1(j) = \pi_j b_i(o_1)$

递归计算：$\alpha_{t+1}(j) = \sum_i \alpha_t(i) a_{ij} b_j(o_{t+1})$

结束输出：$p(O|\lambda) = \sum_j p(O, q_t = S_j|\lambda) = \sum_j \alpha_T(j)$

每个时刻需要计算N个前向变量，其中每个前向变量需考虑从前一时刻N个状态转移到当前状态的可能性，共有T个时刻，所以前向算法的复杂度为$O(N^2T)$

解决问题1：后向算法

- 基本思想：定义后向概率

$$\beta_t(j) = p(o_{t+1}, o_{t+2}, \dots, o_T | q_T = S_j, \lambda)$$

初始化：$\beta_T(j) = 1$

递归计算：$\beta_t(j) = \sum_i a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$

结束输出：$p(O|\lambda) = \sum_j \beta_1(j) \pi_j b_j(o_1)$

注：具体推导过程可参见《统计学习方法》李航著

解决问题2

- 问题2（解码问题）
  - 给定模型λ和观察序列O，计算合适的状态序列Q，使其能够最为合理的解释观察序列O
  - 最大化：$p(Q|O,\lambda) = p(O,Q|\lambda)/p(O|\lambda)$
- 穷举法：所有可能观察序列情况下，隐状态序列的最大概率
- Viterbi法：基于动态规划的快速计算方法

解决问题2：Viterbi法

- 基本思想：
  - 定义Viterbi变量$\delta_t(i)$为时刻t，模型沿某路径到达S_i，并输出观察序列$\{o_1, o_2, \dots, o_t\}$的最大概率

$$\delta_t(\mathrm{i}) = max_{q_1,q_2,\dots,q_{t-1}} p(q_1, q_2, \dots, q_{t-1}, q_t = S_i, o_1, \dots, o_T|\lambda)$$

  - T时刻最大的$\delta_T(\mathrm{i})$所对应的那个状态序列

递归计算：$\delta_t(\mathrm{j}) = \max\left[\delta_{t-1}(i)a_{ij}\right] b_j(O_t), 2 \le t \le T, 1 \le j \le N$

$$\varphi_t(j) = \mathrm{argmax}[\delta_t(i)]$$

解决问题2：Viterbi法

- 基本流程

(1) 初始化：$\delta_1(i) = \pi_i b_i(o_1), \varphi_1(i) = 0, 1 \le i \le N$

(2) 递归计算：$2 \le t \le T$

$$\delta_t(j) = \max\left[\delta_{t-1}(i)a_{ij}\right] b_j(O_t), \qquad 1 \le i \le N$$

$$\varphi_t(j) = \operatorname{argmax}\left[\delta_t(i)\right], \qquad 1 \le i \le N$$

(3) 结束：$p(q_T) = \max\left[\delta_T(i)\right], \ 1 \le i \le N$

$$q_T = \operatorname{argmax}\left[\delta_T(i)\right], \qquad 1 \le i \le N$$

(4) 通过回溯得到最优状态序列：

$$q_t = \varphi_{t+1}(q_{t+1}), \qquad t = T-1, T-2, \dots, 1$$

解决问题3：

- 问题3（模型参数估计）
  - 给定观察序列，通过计算确定模型λ，使$p(O|\lambda)$最大
  - 若状态序列Q已知，可直接采用最大似然估计来计算模型参数

$$\pi_i = \frac{|q_1 = S_i|}{\sum_i |q_1 = S_i|}$$

$$a_{ij} = \frac{|q_t = S_i, q_{t+1} = S_j, 1 \le t \le T-1|}{\sum_j |q_t = S_i, q_{t+1} = S_j, 1 \le t \le T-1|}$$

$$b_i(k) = \frac{|q_t = S_i, o_t = v_k, 1 \le t \le T|}{\sum_i |q_t = S_i, o_t = v_k, 1 \le t \le T|}$$

  - 若状态序列Q未知，则采用Baum-Welch算法

已知观测序列O，估计模型参数$\lambda=(A,B,\pi)$，使得观测序列概率$P(O|\lambda)$最大。

基于期望最大化算法进行求解：含有隐变量的参数估计问题

注：具体过程可参见《统计学习方法》李航著

本讲主要内容

- 自然语言理解概述
- 自然语言理解技术
- 两个重要模型及其应用
  - 语言模型
  - 隐马尔可夫模型

本讲主要内容

- 自然语言理解概述
- 自然语言理解技术

基本了解

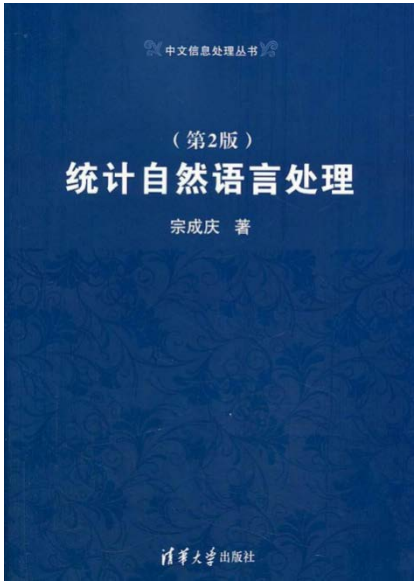
- 两个重要模型及其应用
  - 语言模型
  - 隐马尔可夫模型

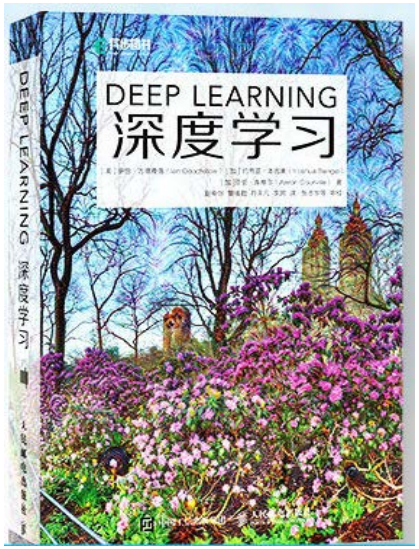
熟练掌握

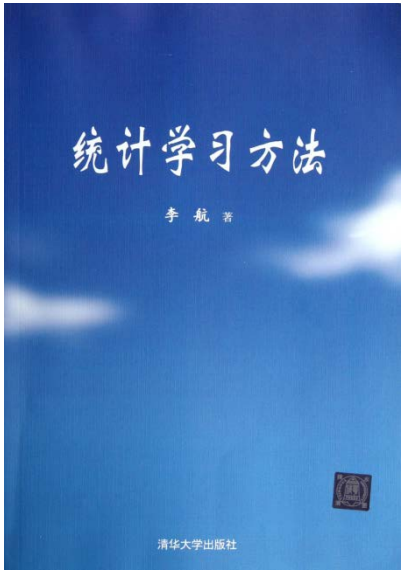
思考：RNN模型在机器翻译、语音识别等自然语言理解任务中的基本应用（序列数据的端到端学习）

参考文献

- 宗成庆，统计自然语言处理
- 李航，统计学习方法
- Yoshua Bengio，Ian J. Goodfellow and Aaron Courville：Deep Learning. http://www.iro.umontreal.ca/~bengioy/DLbook/







感谢大家聆听！