中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

机器翻译原理与方法

第二讲 词法分析技术

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院计算技术研究所 2011 年秋季课程

内容提要

什么是词法分析

英语的词法分析

汉语的词法分析

语言的形态

- 形态：Morphology
  - The study of the internal structure of words, and of the rules by which words are formed, is called morphology. (from V. Fromkin & R. Roman: An Introduction to Language)
  - 单词的内部结构的研究，以及单词形成的规律，被称之为形态学。（选自V. Fromkin & R. Roman：语言介绍）

语言的形态

- 形态：又叫词形变化，同一个词在造句时，因其句法位置的差异而发生的不同变化，是表达语法意义的重要手段。这些不同的变化形成一个聚合。包括词尾，内部屈折，异根等方面。
- 语法范畴：词的变化形式所表示的意义方面的聚合。常见的语法范畴有：性、数、格、体、时态、人称、级等。
- 形态跟语法范畴有对应关系，但不是一回事。

形态分析

- 词法分析，又称形态分析
- 词法分析，指的是对词语的变化形式进行分析，从而得到词语的原形及其所发生的各种形态变化

为什么要做词法分析

- 词典中通常只保留词语的原形
  - 不管一个词语怎么变化，其基本意义和用法是一样的
  - 词典中每个词条，除了词形以外，还需要保存很多信息，如词性、语义、对译词、概率等等
  - 如果为词语的每个变化形式保留一个词典记录，会带来大量的冗余
- 分析出词语的形态变化有助于句子的进一步分析
  - 每一种形态变化都对应于相应的语法功能
  - 词语的形态变化可以有利于引导和约束句法分析
  - 词语的形态排歧有时也会帮助词义的判断

语言的分类

传统语言学根据词的形态把语言分为四大类：

- 分析语：每个词只有一个词素
  - 孤立语（词根语）：词基本上没有专门表示语法意义的附加成分，形态变化很少，语法关系靠词序和虚词来表示。
- 综合语：每个词有多个词素
  - 黏着语：词内有专门表示语法意义的附加成分，一个附加成分表达一种语法意义，一种语法意义也基本上由一个附加成分来表达，词根或词干跟附加成分的结合不紧密。如芬兰语、日语、蒙古语等。
  - 屈折语：用词的形态变化表示语法关系，一个形态成分可以表示若干种不同的语法意义，词根或词干跟词的附加成分结合得很紧密，往往不易截然分开。
  - 多式综合语（编插语）：最复杂的综合语，一个词语通常非常长，由很多词素组成。

语言的分类

- 屈折语和黏着语的联系是都有表示语法意义的附加语素。它们的区别是：
  - 第一，从附加语素形式表示的语法意义的关系来看，屈折语不是一对一的关系，黏着语是一对一的关系；
  - 第二，屈折语的附加语素与主体语素 (词根) 结合紧密，黏着语结合松散；
  - 第三，屈折语有少量的内部屈折变化形式，黏着语没有。

语言的分类

- 多式综合语跟黏着语不同：多式综合语中的附加语素的形式和意义不是一对一的关系，而且附加语素结合紧密。
- 多式综合语跟屈折语的不同是，多式综合语句子跟词不区分，屈折语区分。

孤立语

- 孤立语：汉语、泰语、越南语……
  - 词语只有一种形式，不发生形态变化
  - 汉语其实也有重叠等变化形式

屈折语

- 屈折语：英语、法语、德语、俄语……
  - 词语可以附加一层词缀
  - 词缀有多重语法意义

    英语动词 +s：第三人称、单数、现在时
  - 相对来说，英语词语的形态变化是比较简单的，动词只有四种变化形式，从形态变化复杂度上来说接近分析语
  - 有些屈折语的形态变化非常复杂

法语的形态变化

关于法语的动词形态变化， 以下举动词 **être**（/ɛtʁə/）“是”为例：

—— 摘自维基百科

时态 temps	现在时 présent				简单未来时 futur simple	未完成过去时 imparfait		简单过去时 passé simple
语式 modes	直陈式 indicatif	虚拟式 subjonctif	命令式 impératif	条件式 conditionnel	直陈式 indicatif	直陈式 indicatif	虚拟式 subjonctif	直陈式 indicatif
1人称单数 je 我	**suis** [sɥi]	**sois**	-	**serais**	**serai**	**étais**	**fusse**	**fus**
2人称单数 tu 你	**es** [e]	**sois**	**sois**	**serais**	**seras**	**étais**	**fusses**	**fus**
3人称单数 il/elle 他/她	**est** [e]	**soit**	-	**serait**	**sera**	**était**	**fût**	**fut**
1人称复数 nous 我们	**sommes** [sɔm]	**soyons**	**soyons**	**serions**	**serons**	**étions**	**fussions**	**fûmes**
2人称复数 vous 你们	**êtes** [ɛtə]	**soyez**	**soyez**	**seriez**	**serez**	**étiez**	**fussiez**	**fûtes**
3人称复数 ils/elles 他们/她们	**sont** [sɔ̃]	**soient**	-	**seraient**	**seront**	**étaient**	**fussent**	**furent**

粘着语

- 粘着语：主要包括阿尔泰语系、乌拉尔语系等语系的许多语言。例如维吾尔语、蒙古语、土耳其语、芬兰语、匈牙利语，还有日语、朝鲜语，非洲班图语族的许多语言。
- 粘着语的一个词语后面可以附加多个后缀，每个后缀都有独立的语法意义
- 粘着语的一个词语可能的形态变化非常丰富，有些语言词语可能的形态变化达数千种

土耳其语形态变化

- 在土耳其语中“ sev ”表示“爱”，是一个动词的主体语素。
- 在它的后面可以有下面的附加语素：
  - “ dir ”表示“第三人称”
  - “ ler”表示“复数”
  - “ mis ”表示“过去时”
  - “ erek ”表示“将来时”
- 由它们组合形成的：
  - “sev mi s dir ler”就是“他们从前爱”的意思
  - “sev erek dir ler”就是“他们将要爱”的意思

内容提要

什么是词法分析

英语的词法分析

汉语的词法分析

英语的词法分析

- Tokenization：把字符串变成词串（tokens）
  I'm a student. → I 'm a student .
- Stemming：对词的内部结构进行分析，并还原到词典形式。实际包括两个层次
  - 是对屈折进行还原。

    takes → take + ~s

    took → take + ~ed
  - 对派生进行还原。

    tokenization → token + ~ize + ~tion
- Stemming 也称为 Lemmatization 。
- POS-Tagging: 词性标注

Tokenization

- 数字： 123,456.78 90.7% 3/8 11/20/2000
- 缩略（包含不同的情况）：
  - 字母－点号－字母－点号组成的序列，比如： U.S. i.e. 等等；
  - 字母开头，最后以点号结束，比如： A. b. Mr. eds.prof. ；
- 包含非字母字符，比如： AT&T Micro$oft
- 带杠的词串，比如： three-year-old ， one-third ， so-called
- 带瞥号的词串，比如： I'm can't dog's let's
- 带空格的词串，比如： "and so on" ， "ad hoc“
- 其他：如网址（ http://ict.ac.cn ）、公式等

Tokenization 问题

- 例外较多，跟文本来源有关
- 歧义现象（如点号的句子边界歧义）

数字的识别

数词的识别一般可以用有限状态自动机来实现

- 识别分数的正则表达式：
  - [0-9]+ / [0-9]+
  - e.g. 12/21
- 识别百分数的正则表达式：
  - ([+ | -]) ? [0-9]+ (. [0-9]*) ? %
  - e.g. -5.9% 91%
- 识别十进制数字的正则表达式：
  - ([0-9]+(,)?)+ (. [0-9]+)?
  - e.g. 12,345

Tokenization 算法

- 输入：一段文本
- 输出：单词串
- 算法：（略）

Stemming

屈折型语言的词语变化形式：

- 屈折变化：即由于单词在句子中所起的语法作用的不同而发生的词的形态变化，而单词的词性基本不变的现象，如（ take, took, takes ）。识别这种变化是词法分析的最基本的任务。
- 派生变化：即一个单词从另外一个不同类单词或词干衍生过来，如 morphological ← morphology，英语中派生变化主要通过加前缀或后缀的形式构成；在其他语言中，如德语和俄语中，同时还伴有音的变化。
- 复合变化：两个或更多个单词以一定的方式组合成一个新的单词。这种变化形式比较灵活，如 well-formed, 6-year-old 等等。

Stemming 的目的：将上述变化还原

Stemming 常见的问题

- 半规则变化
  - flied → fly + ~ed
  - rebelled → rebel + ~ed
- 不规则变化
  - good, better, best
  - child, children
- 歧义现象
  - better → good + ~er or well + ~er ?
  - works → work + ~s or works ?

Stemming 规则示例 (1)

- 名词复数

  ***s → *, (PLUR)**

  ***es → *, (PLUR)**

  ***ies → *y, (PLUR)**

- 动词第三人称单数

  ***s → * (SINGULAR) (THIRDPERSON) (PRESENT)**

  ***es → * (SINGULAR) (THIRDPERSON) (PRESENT)**

  ***ies → *y (SINGULAR) (THIRDPERSON) (PRESENT)**

Stemming 规则示例 (2)

- 动词现在分词
  - *ing → * (VING)
  - *ing → *e (VING)
  - *ying→ *ie (VING)
  - *??ing →*? (VING)
- 动词过去分词、过去式
  - *ed → * (PAST,VEN)
  - *ed → *e (PAST,VEN)
  - *ied → *y (PAST,VEN)
  - *??ed → *? (PAST,VEN)

Stemming 算法

- 输入：一个单词
- 输出：一个或多个单词，其中每个单词还原为原形加前后缀（可以有多个）
- 算法：（略）

基于有限状态自动机的 Stemming

- 有限状态自动机是 Stemming 中的常用算法
- 有限状态自动机的优点是表现形式直观，效率高

Stemming 要做到何种程度

- 词干层。如：impossibilities→impossibility+ies
- 词根层。如：impossibilities→im+poss+ibil+it+ies
- 分析程度取决于自然语言处理系统的深度：
  - 不解决未定义词，分析到词干层
  - 解决未定义词，要分析到词根层。

内容提要

什么是词法分析

英语的词法分析

汉语的词法分析

汉语词法分析

- 汉语词法分析所面临的问题
- 基于词典的汉语词语机械切分算法
- 基于语言模型的汉语词语切分算法
- 基于隐马尔科夫模型的词性标注算法
- 基于字标注的汉语词语切分标注一体化算法

汉语词法分析所面临的问题

- 重叠词、离合词、词缀
- 汉语词语的切分歧义
- 汉语未定义词
- 词性标注

汉语双字形容词的重叠形式

形容词 (AB)	ABAB 式	AABB 式	A 里 AB 式
高兴	高兴高兴	高高兴兴	
明白	明白明白	明明白白	
热闹	热闹热闹	热热闹闹	
潇洒	潇洒潇洒	潇潇洒洒	
糊涂		糊糊涂涂	糊里糊涂
流气			流里流气
粘乎	粘乎粘乎	粘粘乎乎	
凉快	凉快凉快	凉凉快快	

汉语单字形容词的重叠形式

形容词（A）	AA 式	ABB 式	ABCD 式
黑	黑黑	黑压压	黑不溜秋
白	白白	白花花	白不呲咧
红	红红	红彤彤	
亮	亮亮	亮晶晶	
恶		恶狠狠	
香	香香	香喷喷	
滑	滑滑	滑溜溜	

汉语离合词

- 汉语动词存在离合词现象
  - 游泳：游了一会儿泳
  - 理发：发理了没有
  - 担心：担什么心
  - 洗澡：洗了个热水澡

汉语缩合词

- 二十五六
- 星期三四
- 男女生，进出口
- ……

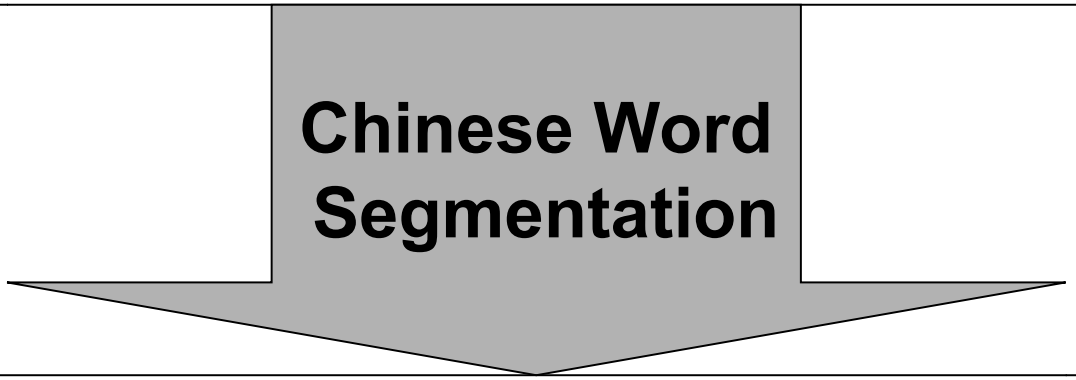
四字格

- 千辛万苦、千山万水、千奇百怪……
- 七上八下、七零八落、
- 龙腾虎跃、龙飞凤舞、虎啸龙吟……

汉语的词语切分

物理学很有趣

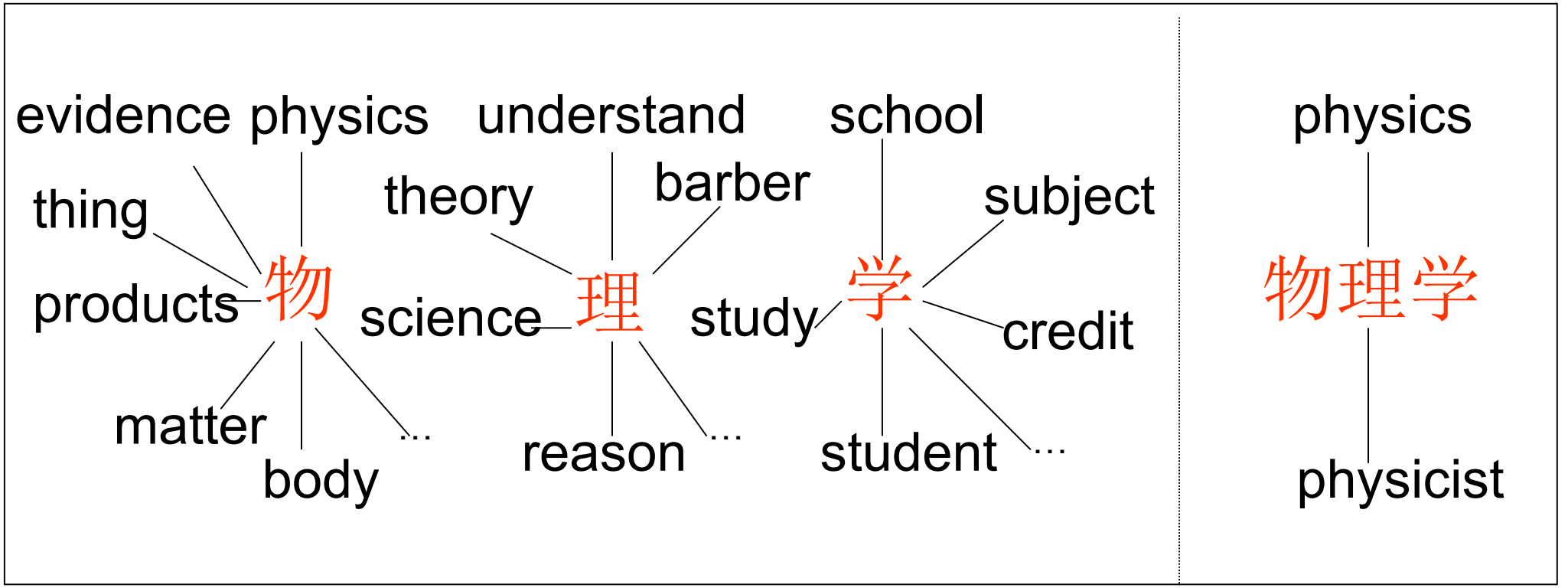
(Physics is very interesting)



物理学 / 很 / 有趣

Physics Very Interesting

为什么要做词语切分


evidence
physics
thing
products
物
matter
body
…
understand
theory
barber
science
理
reason
…
school
subject
study
学
credit
student
…
physics
物理学
physicist

为什么要做词语切分

物理学很有趣

(Physics is very interesting)

物理学 / 很 / 有趣

Physics Very Interesting

物理 / 学 / 很 / 有趣

Physics Learn Very Interesting

汉语的切分歧义

- 交集型歧义（交叉型歧义）：如果字串 abc 既可切分为 ab/c，又可切分为 a/bc。其中 a， ab， c 和 bc 是词
  - 有意见： 我 对 他 有 意见。 总统 有意 见 他。
- 组合型歧义（覆盖型歧义）：若 ab 为词，而 a 和 b 在句子中又可分别单独成词
  - 马上： 我 马上 就 来。 他 从 马 上 下来。
  - 将来： 我 将来 要 上 大学。 我 将 来 上海。
- 混合型歧义：由交集型歧义和组合型歧义自身嵌套或两者交叉组合而产生的歧义
  - 人才能：这样 的 人才 能 经受 住 考验。
  - 人才能：这样 的 人 才能 经受 住 考验。
  - 人才能：这样 的 人 才 能 经受 住 考验。

交集型歧义字段的链长

- 链长：交集型歧义字段中含有交集字段的个数，称为链长。
  - 链长为 1 ：和尚未
  - 链长为 2 ：结合成分
  - 链长为 3 ：为人民工作
  - 链长为 4 ：中国产品质量　　结合成分子时
  - 链长为 6 ：努力学习语法规则
  - 链长为 8 ：治理解放大道路面积水

真歧义

中国 有

China have

中国有十几亿人。

中 国有

in state-owned

本地企业中国有企业占 30% 。

伪歧义

尚未	来
not yet	come

他尚未来过此地。

尚	未来
still	future

不存在这样的切分实例

未定义词识别的困难

- 未定义词没有明确边界
- 未定义词的构成单元（汉字）本身都可以独立成词

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：周鹏和同学，周鹏和同学
- 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
  - [王国]维、[高峰]、[汪洋]、张[朝阳]
- 人名与其上下文组合成词
  - 这里[有关]天培的壮烈；
  - 费孝[通向]人大常委会提交书面报告
- 人名地名冲突
  - 河北省刘庄

未定义词的类型

- 汉语人名：李素丽 老张 李四 王二麻子
- 汉语地名：定福庄 白沟 三义庙 韩村河 马甸
- 翻译人名：乔治·布什 叶利钦 包法利夫人
- 翻译地名：阿尔卑斯山 新奥尔良 约克郡
- 机构名：方正公司 联想集团 国际卫生组织 外贸部
- 商标字号：非常可乐 乐凯 波导 杉杉 同仁堂
- 专业术语：万维网 主机板 模态逻辑 贝叶斯算法
- 缩略语：三个代表 五讲四美 打假 扫黄打非 计生办
- 新词语：卡拉 OK 波波族 美刀 港刀

汉语词法分析

汉语词法分析所面临的问题

基于词典的汉语词语机械切分算法

基于语言模型的汉语词语切分算法

基于隐马尔科夫模型的词性标注算法

基于字标注的汉语词语切分标注一体化算法

基于词典的词语机械切分方法

- 输入：
  - 一个词表（词典）
  - 一个待切分句子
- 输出：
  - 一个词语序列

基于词典的词语机械切分方法

- 全切分
- 最大匹配方法
- 最短路径方法
- 基于记忆的交叉歧义排除法
- 基于规则的切分算法

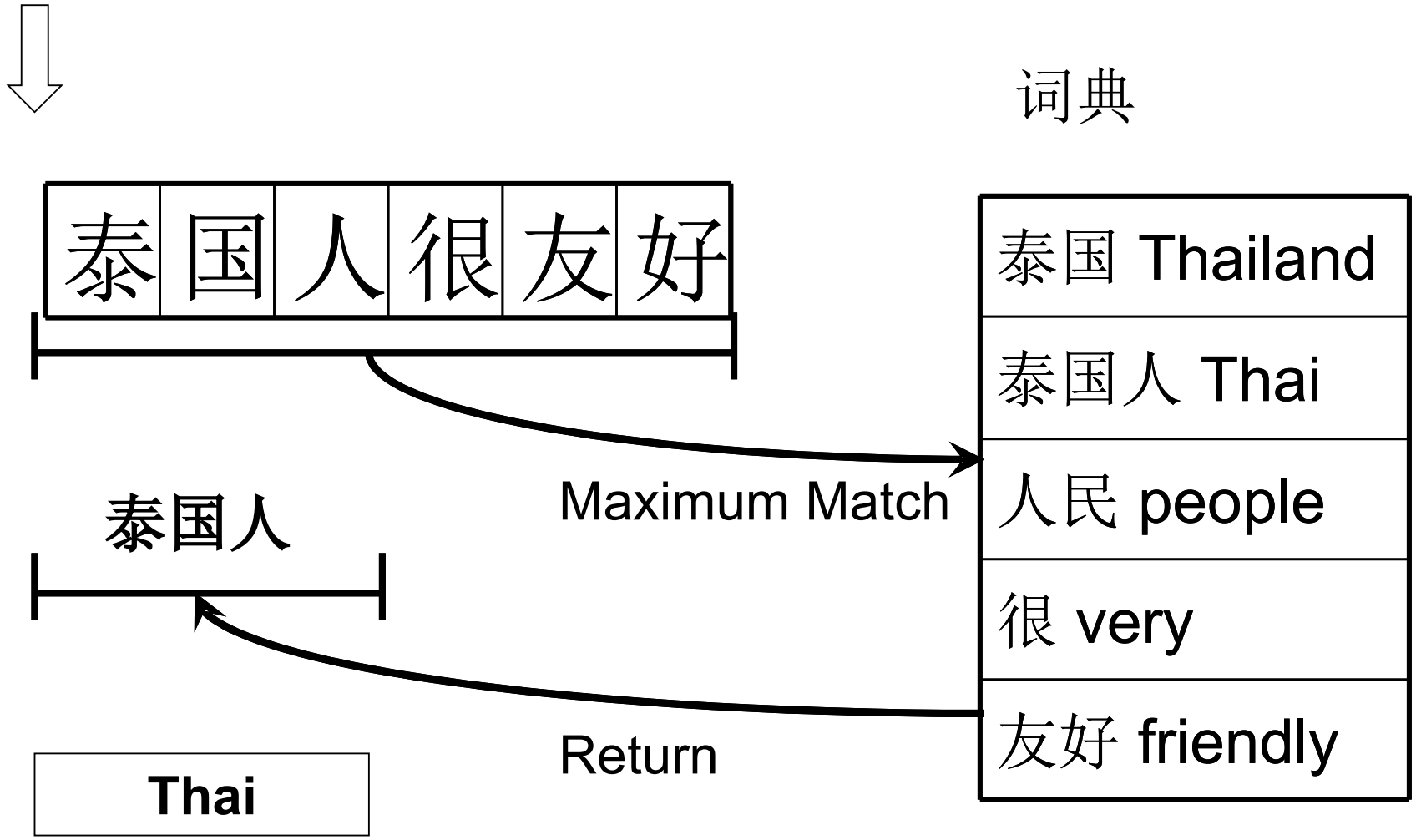
全切分方法

- 给出所有的切分结果
- 算法（略）（基本原理：递归算法）
- 算法的时间复杂度随着句子长度的增加呈指数增长

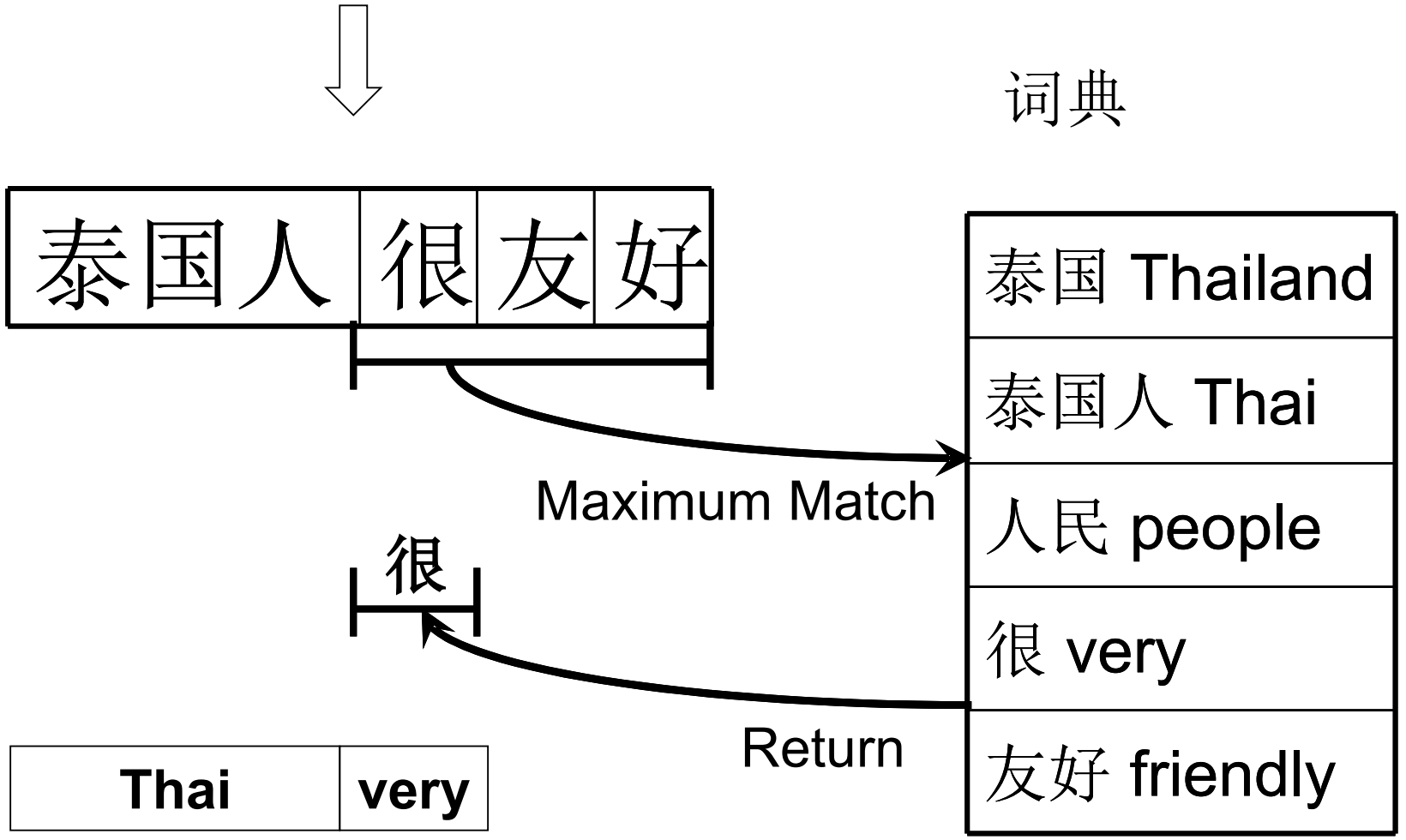
最大匹配方法

- 正向最大匹配（ MM ）
  - 自左往右
  - 每次取最长词
- 逆向最大匹配（ RMM)
  - 自右往左
  - 每次取最长词
- 双向最大匹配
  - 依次采用正向和逆向最大匹配
  - 如果结果一致则输出
  - 如果结果不一致再用其他方法排歧

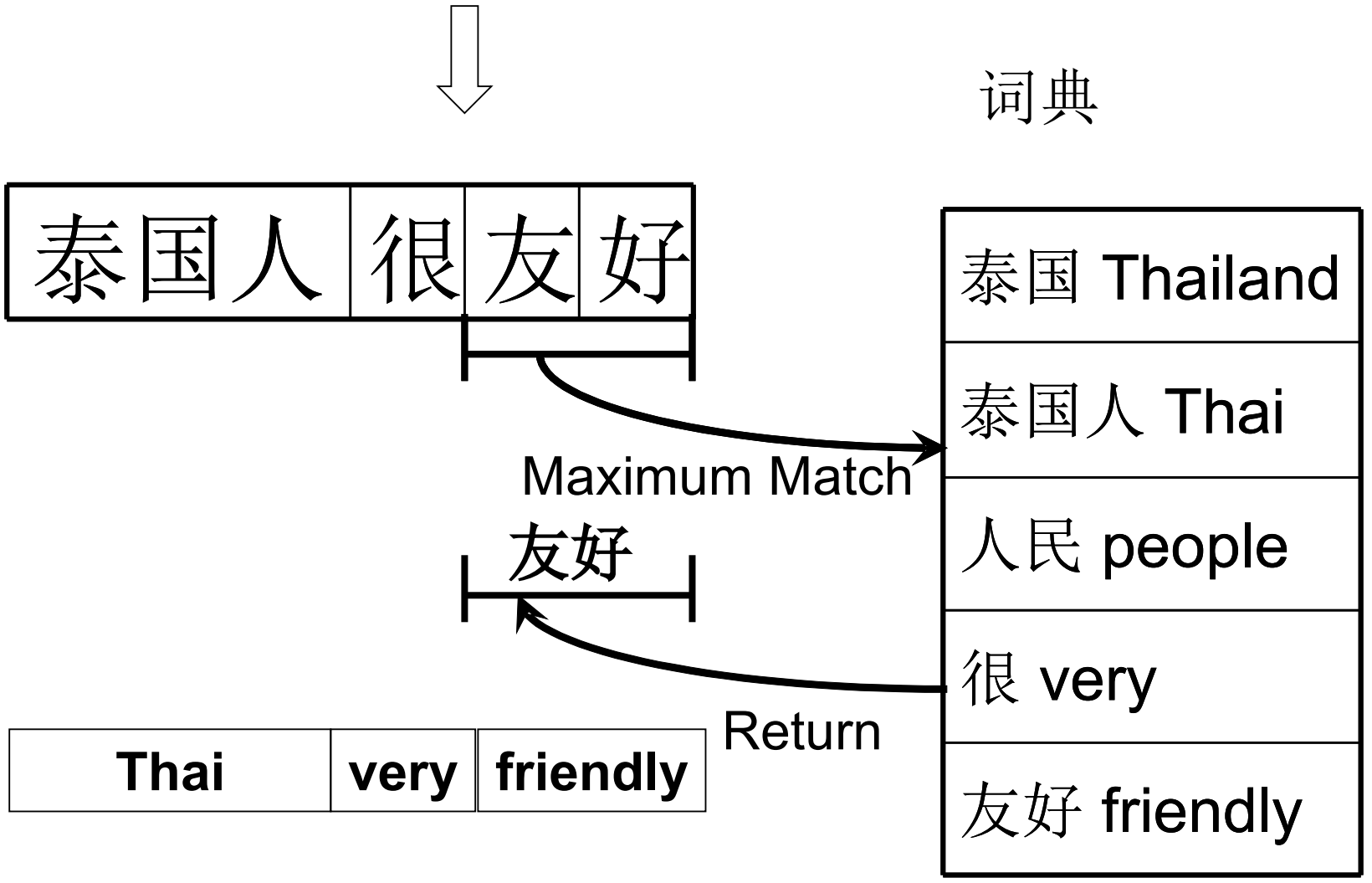
正向最大匹配


词典
泰 国 人 很 友 好
泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly
Maximum Match
泰国人
Return
Thai

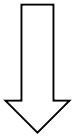
正向最大匹配


词典
泰国人 很 友 好
泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly
Maximum Match
很
Return
Thai very

正向最大匹配


词典
泰国人 很 友 好
泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly
Maximum Match
友好
Return
Thai very friendly

正向最大匹配



词典

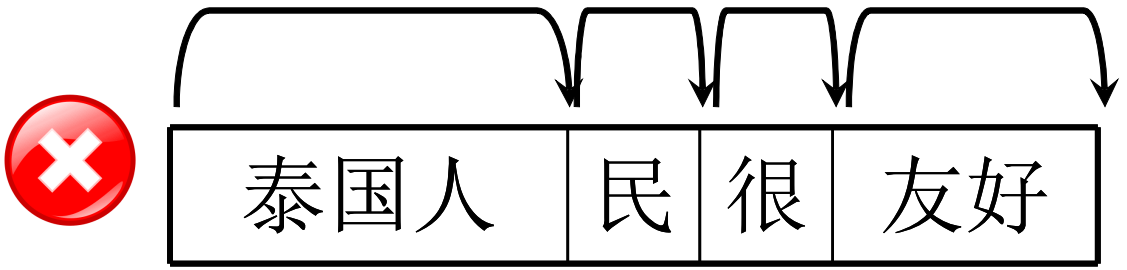
泰国人	很	友好

泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly

Thai are very friendly

正向最大匹配

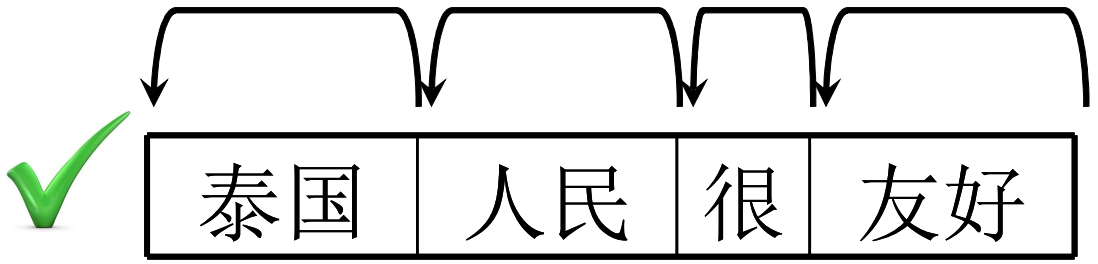
- 正向匹配算法显然很容易导致错误



泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly

逆向最大匹配

- 这个例子用逆向匹配算法切分正确 .



Thai people are very friendly

泰国 Thailand
泰国人 Thai
人民 people
很 very
友好 friendly

最大匹配方法的优缺点

- 优点
  - 简单、快速
  - 在某些应用场合已经足够
- 缺点
  - 单向最大匹配会忽略交集型歧义和组合型歧义

    幼儿园　地　节目　/　独立自主　和平　等　互利　的　原则
  - 双向最大匹配会忽略链长为偶数的交集型歧义和组合型歧义

    原子　结合　成分　子时　/　他　从　马上　下来

汉语词法分析

汉语词法分析所面临的问题

基于词典的汉语词语机械切分算法

基于语言模型的汉语词语切分算法

基于隐马尔科夫模型的词性标注算法

基于字标注的汉语词语切分标注一体化算法

什么是统计语言模型

- 语言模型给出任何一个句子的出现概率：

  $\Pr(Sentence=w_1 w_2 \ldots w_n)$

  归一化条件：$\sum_{Sentence}\Pr(Sentence)=1$

- 统计语言模型实际上就是一个概率分布，它给出了一种语言中所有可能的句子的出现概率
- 在统计语言模型看来，对于一种语言，任何一个句子都是可以接受的，只是接受的可能性（概率）不同
- 统计语言模型问题是一个典型的序列评估问题

语言模型的类型

- 理论上，单词串的任何一种概率分布，都是一个语言模型。
- 实际上， N 元语法模型是最简单也是最常见的语言模型。
- N 元语法模型由于没有考虑任何语言内部的结构信息，显然不是理想的语言模型。
- 其他语言模型：
  - 隐马尔科夫模型（ HMM ）（加入词性标记信息）
  - 概率上下文无关语法（ PCFG ）（加入短语结构信息）
  - 概率链语法（ Probabilistic Link Grammar ）（加入链语法的结构信息）
- 目前为止，其他形式的语言模型效果都不如 N 元语法模型
- 统计机器翻译研究中开始有人尝试基于句法的语言模型

N 元语法模型—概念辨析

- N 元语法模型： N-Gram Model 。
- 所谓 N-Gram ，指的是由 N 个词组成的串，可以称为“ N 元组”，或“ N 元词串”。
- 基于 N-Gram 建立的语言模型，称为“ N 元语法模型 (N-Gram Model)”。
- Gram 不是 Grammar 的简写。在英文中，并没有 N-Grammar 的说法。
- 在在汉语中，单独说“ N 元语法”的时候，有时指“ N 元组 (N-Gram)”，有时指“ N 元语法模型 (N-Gram Model)”，请注意根据上下文加以辨别。

N 元语法模型—定义

- N 元语法模型（ N-gram Model ）

$$P(w)=\prod_{i=1}^{n} p(w_i|w_1 w_2 \ldots w_{i-1})$$

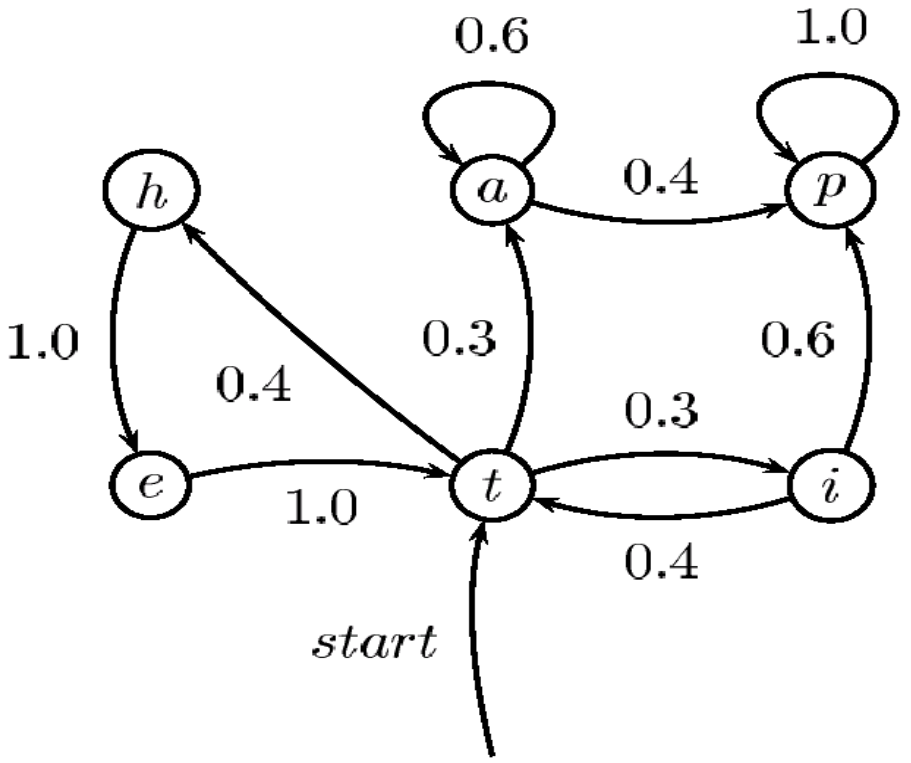
$$\approx\prod_{i=1}^{n} p(w_i|w_{i-N+1} w_{i-N+2} \cdots w_{i-1})$$

- 假设：单词 w_i 出现的概率只与其前面的 N-1 个单词有关

N 元语法模型—举例

- N=1 时： 一元语法模型
  - 相当于词频表，给出所有词出现的频率
- N=2 时： 二元语法模型
  - 相当于一个转移矩阵，给出每一个词后面出现另一个词的概率
- N=3 时： 三元语法模型
  - 相当于一个三维转移矩阵，给出每一个词对儿后面出现另一个词的概率
- 在自然语言处理中， N 元语法模型可以在汉字层面，也可以在单词层面，还可以在概念层面……

二元语法模型—图示



$$P(t\text{-}i\text{-}p) = p(X_1 = t)p(X_2 = i|X_1 = t)p(X_3 = p|X_2 = i)$$

$$= 1.0 \times 0.3 \times 0.6 = 0.18$$

袋子模型 Bag Model (1)

- 将一个英语句子中所有的单词放入一个袋子中
- 用 N 元语法模型试图将其还原
  - 对于这些单词的任何一种排列顺序根据 N 元语法模型计算其出现概率
  - 取概率最大的排列方式

袋子模型 Bag Model (2)

- 实验：取 38 个长度小于 11 个单词的英语句子，实验结果如下：

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

In our organization research has two missions.
⇒ In our missions research organization has two.

N 元语法模型的参数估计

- 最大似然估计：
  选择参数，使得训练语料出现的概率最大

$$p(w_n|w_1 w_2 \ldots w_{n-1}) = \frac{f(w_{1..} w_n)}{f(w_{1..} w_{n-1})}$$

用实际样本中事件出现的频率来估计该事件的概率

数据平滑

- 数据稀疏问题
  - 如果 $f(w_1 \ldots w_n) = 0$，那么出现零概率，导致整个文本的出现概率为零
- 解决办法：劫富济贫
- 约束：概率的归一性

$$\sum_{w_n} p(w_n | w_1 w_2 \ldots w_{n-1}) = 1$$

句子首尾标记处理

- 在语言模型训练时，要注意给每一个句子加上句子首尾标记，通常用 <s> 和 </s> 来表示，应用语言模型时，也要把句子首尾标记考虑进来，否则会影响模型应用的效果。
- 在考虑句子首尾标记的情况下，用三元语法模型计算一个句子 w1w2w3 的概率应该是：

$$P(< s > w_1 w_2 w_3 < / s >) = P(w_1 |< s >) \times P(w_2 |< s > w_1)$$
$$\times P(w_3 | w_1 w_2) \times P(< / s > | w_2 w_3) \times P(< / s > | w_3)$$

N元语法模型工具

- 开源工具：
  - SRI Language Model
  - IRST Language Model (in Moses)

基于 N 元语法的词语切分

- 对于每一个切分结果，采用 n 元语法模型计算其概率，并输出概率最大的切分结果

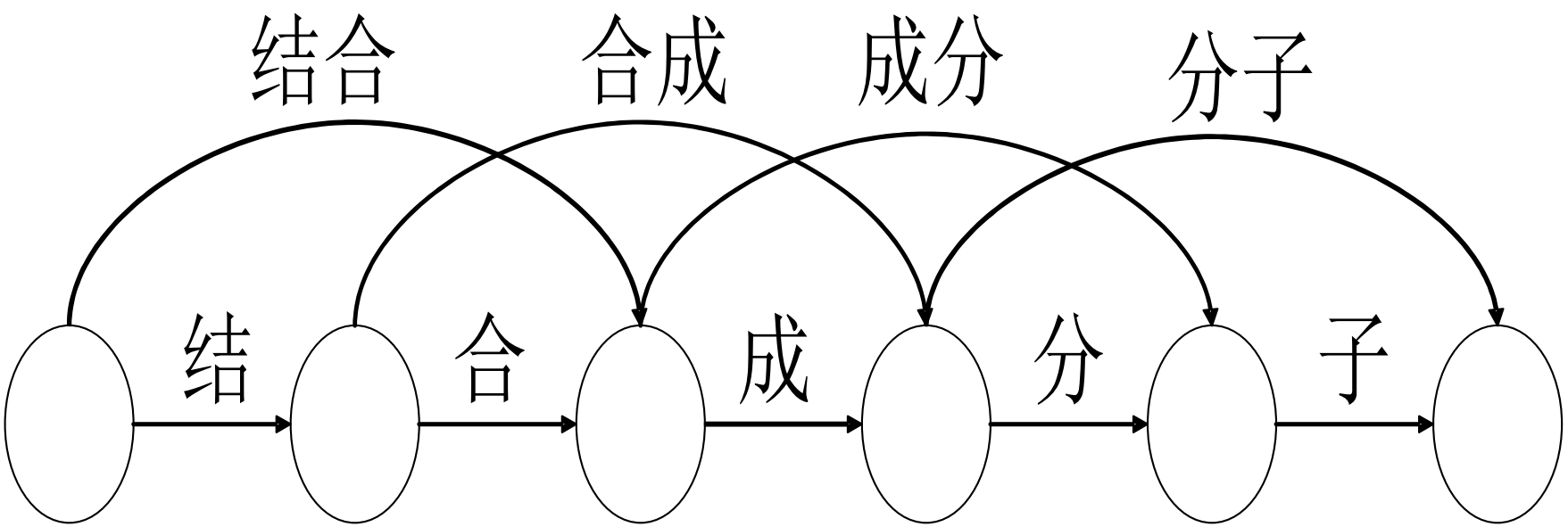
$$W^* = \arg\max_W P(W)$$

$$= \arg\max_W p(w_1 \ldots w_{N-1}) \prod_{i=N}^{l} p(w_i \mid w_{i-N+1} \ldots w_{i-1})$$

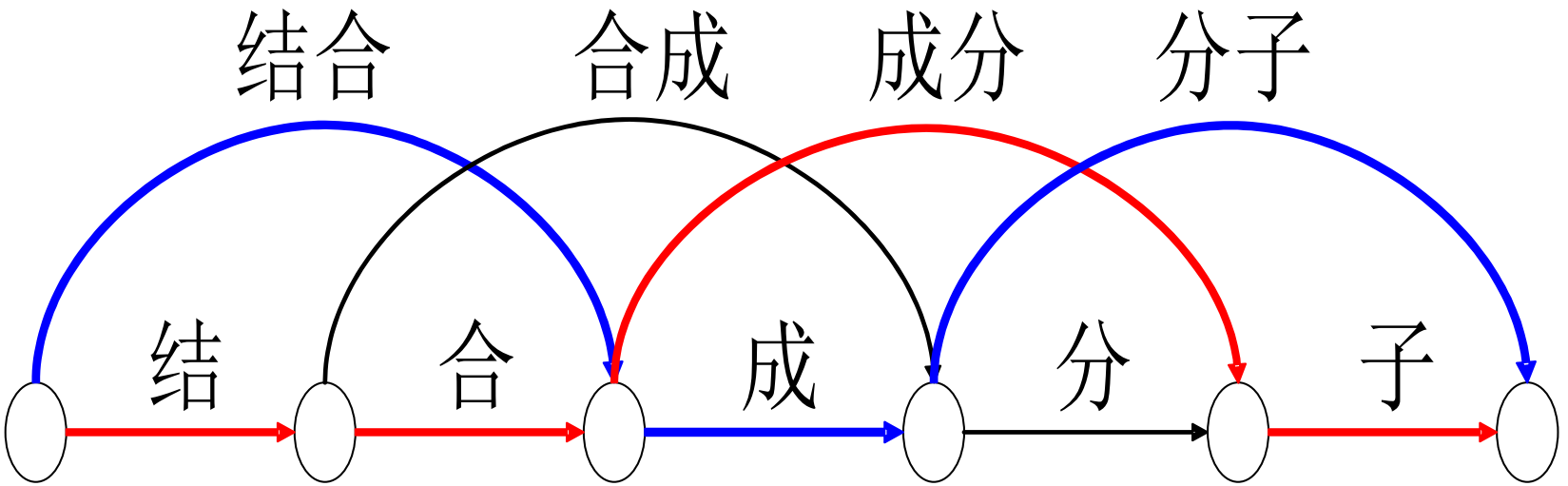
一元切分词图

- 对任何一个未切分句子，可以构造一个一元切分词图
- 一元切分词图为一个有向图：
  - 结点：相邻两个汉字之间的间隔
  - 边：一个候选的词语
- 在一元切分词图上，从句子起始位置开始，到句子结束位置中止的任何一条路径，对应着一种可能的词语切分结果。

一元切分词图



一元切分词图

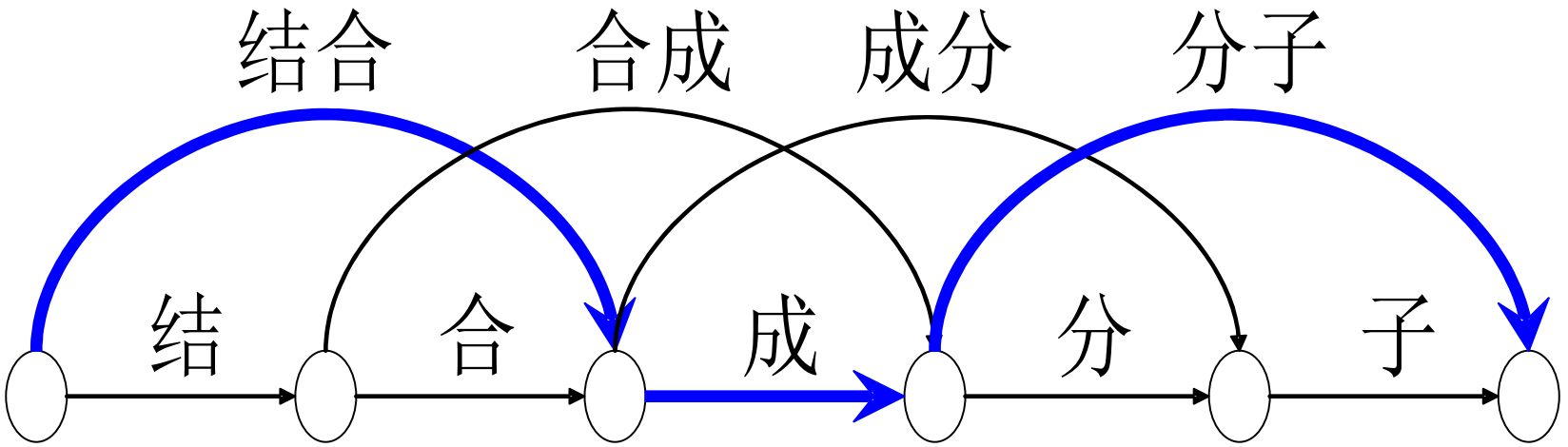


路径即切分：

- **红色路径： 结 . 合 . 成分 . 子**
- **蓝色路径： 结合 . 成 . 分子**

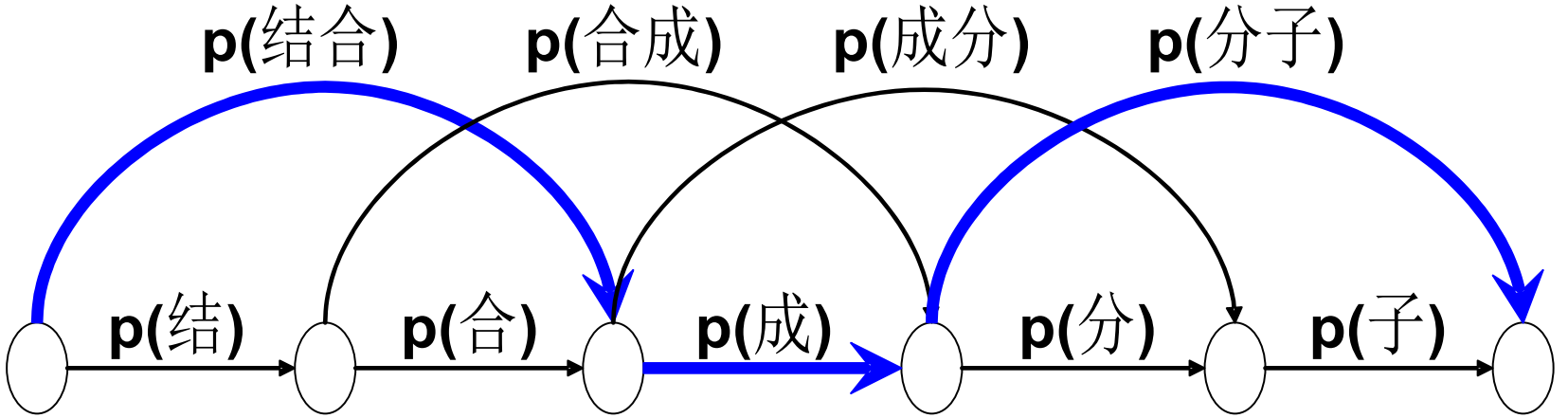
切分表示为词图的路径

词语切分可以转化为在切分词图上寻找概率最大的最优路径问题。



基于一元语法模型的词语切分

给词图上每一条边赋予相应词语的一元语法模型概率，一条切分路径的概率表示为路径上所有边的概率乘积。

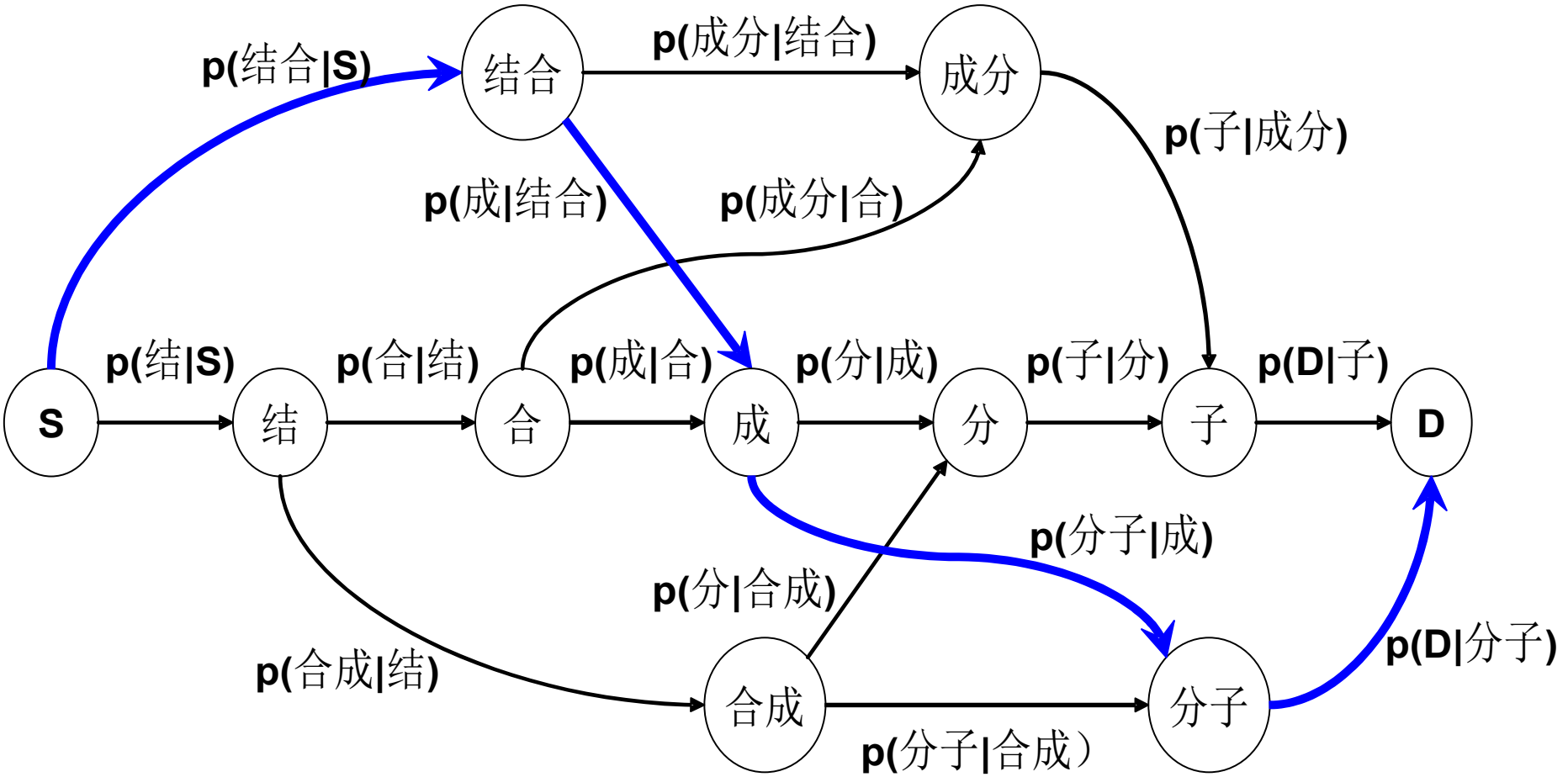


p(结合·成·分子)=p(结合)p(成)p(分子)

二元切分词图

- 一元切分词图无法表示二元语法模型所需的二元词语转移概率
- 二元切分词图定义为如下的有向图：
  - 结点：任何一个可能的候选词语 (W_i)
  - 边：相邻两个词语的接续关系 (W_{i-1})→(W_i)
- 在二元切分词图的每一条边上标记二元词语转移概率 $P(W_i|W_{i-1})$
- 任何一个词语切分可以表示为二元切分词图上的一条起始结点到结束结点的路径
- 路径上所有边的概率之积就是该切分结果对应的二元语法模型概率。

二元切分词图

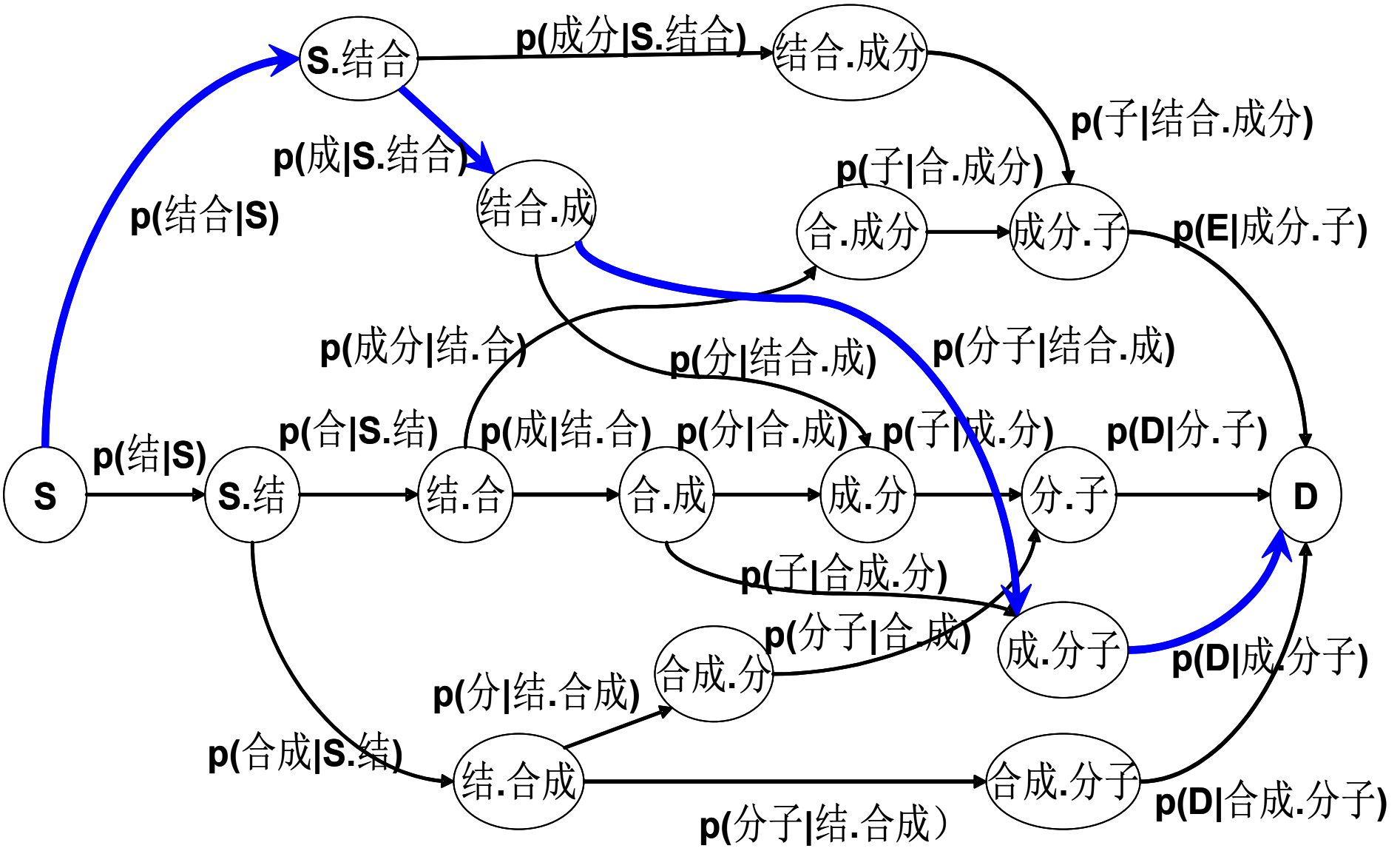


p(结合·成·分子)=p(结合|S)p(成|结合)p(分子|分子)p(D|分子)

三元切分词图

- 三元切分词图
  - 结点：任何一个可能的词语对 $(W_{i-1}W_i)$
  - 边：两个结点之间的有向边，有向边的起点的后一个词语与终点的前一个词语相同 $(W_{i-2}W_{i-1})\rightarrow(W_{i-1}W_i)$
- 在三元切分词图的每一条边上标记三元词语转移概率 $P(W_i|W_{i-2}W_{i-1})$
- 任何一个词语切分可以表示为三元切分词图上的一条起始结点到结束结点的路径
- 路径上所有边的概率之积就是该切分结果对应的三元语法模型概率。

三元切分词图


S.结合
p(成分|S.结合)
结合.成分
p(子|结合.成分)
p(成|S.结合)
结合.成
p(子|合.成分)
合.成分
成分.子
p(E|成分.子)
p(结合|S)
p(成分|结.合)
p(分|结合.成)
p(分子|结合.成)
p(合|S.结)
p(成|结.合)
p(分|合.成)
p(子|成.分)
p(D|分.子)
p(结|S)
S
S.结
结.合
合.成
成.分
分.子
D
p(子|合成.分)
p(分子|合.成)
成.分子
p(D|成.分子)
合成.分
p(分|结.合成)
p(合成|S.结)
结.合成
合成.分子
p(D|合成.分子)
p(分子|结.合成)

N 元切分词图

- N 元切分词图
  - 结点：任何一个可能的 N-1 词语对 ($W_{i-N+2}\cdots\cdots W_i$)
  - 边：两个结点之间的有向边，有向边的起点的后 N-2 个词语与终点的前 N-2 个词语相同 ($W_{i-N+1}\cdots\cdots W_{i-1}$)→($W_{i-N+2}\cdots\cdots W_i$)
- 在 N 元切分词图的每一条边上标记 N 元词语转移概率 $P(W_i|W_{i-N+1}\cdots\cdots W_{i-1})$
- 任何一个词语切分可以表示为 N 元切分词图上的一条起始结点到结束结点的路径
- 路径上所有边的概率之积就是该切分结果对应的 N 元语法模型概率。

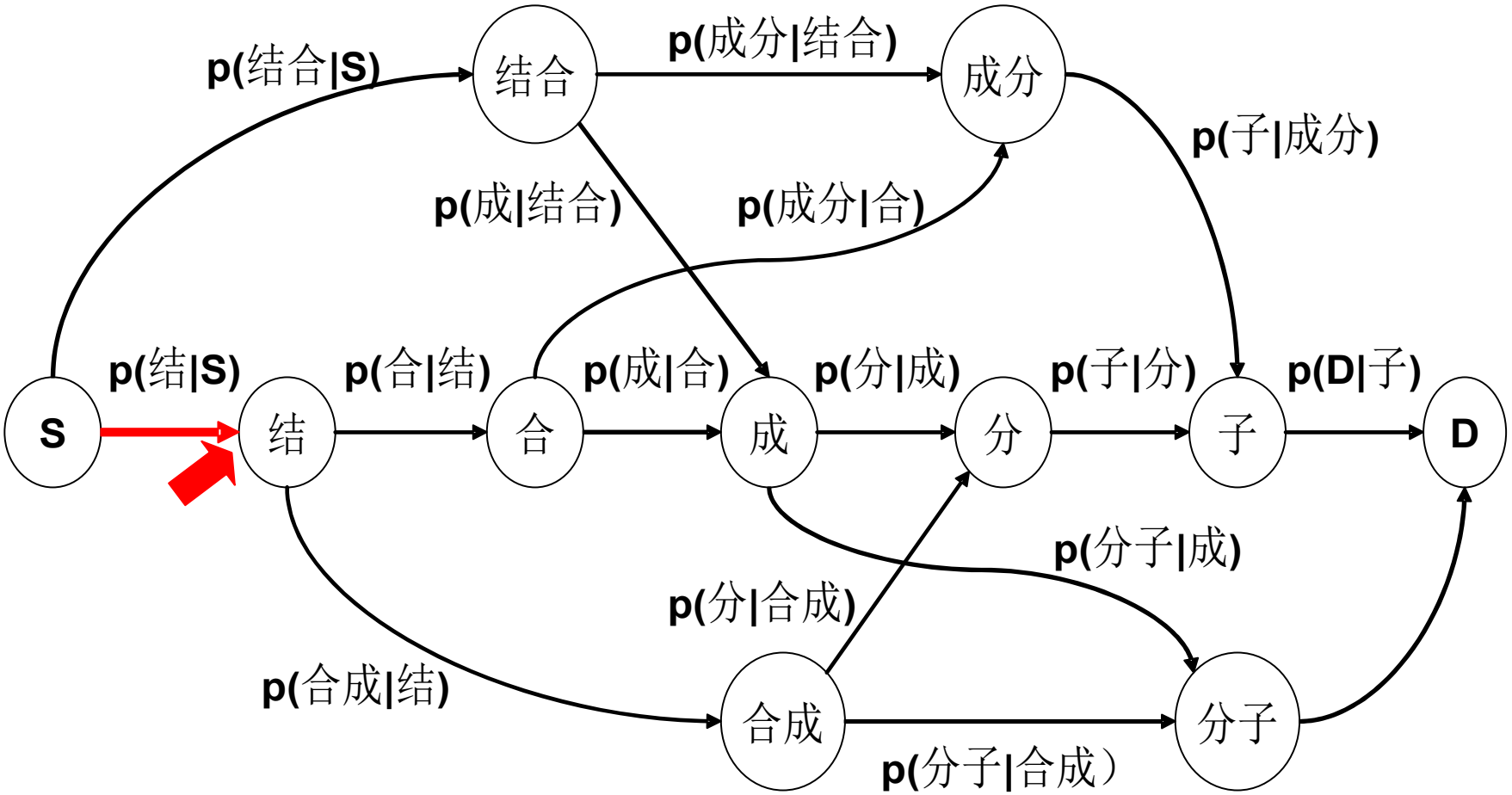
N 元切分词图的构造

- 先构造 N-1 元词图
- 对于 N-1 元词图上每一条边：在 N 元词图上添加一个结点；
- 对于 N-1 元词图上每一个结点：假设该结点有 S 条入边 $(e_{i1},\ldots,e_{iS})$ 和 T 条出边 $(e_{o1},\ldots,e_{oT})$ ，那么对于对于该结点的每一对入边和出边的组合 (e_{is},e_{ot}) ，在 N 元词图上增加一条边，该边的起点和终点分别是 e_{is} 和 e_{ot} 在 N 元词图上对应的结点

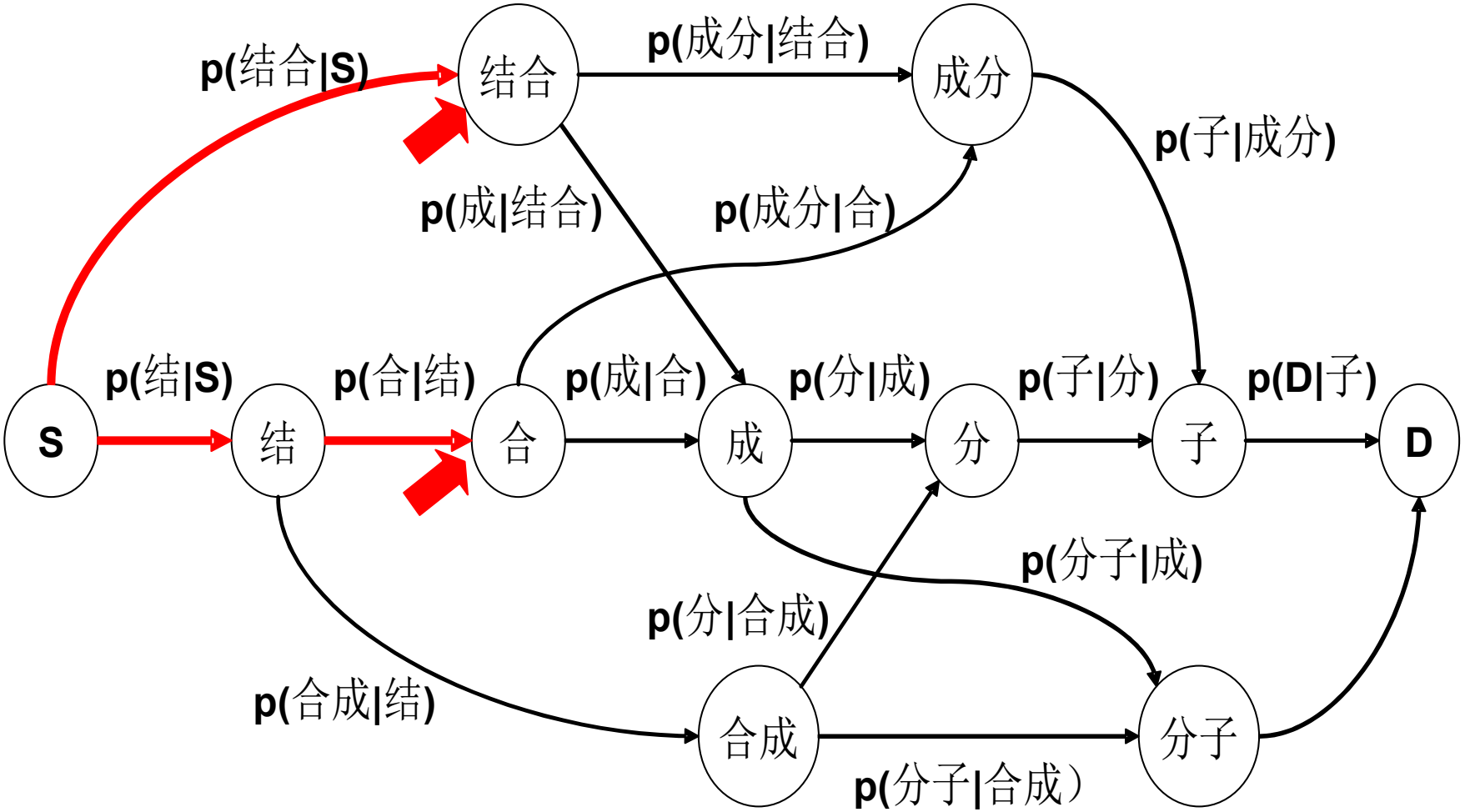
在词图上搜索最优路径： Viterbi 算法

- Viterbi Search Algorithm
  - Assign p(source_node)=1
  - For each node ***n*** in the word lattice from source to destination in a **topological order**
    - p(***n***)=0, previous_edge(***n***)=ø
    - For each edge ***e*** directed to ***n*** from ***n***'
      - p'(***n***)=p(***n***')*p(***e***)
      - If p'(***n***) > p(***n***) then p(***n***)=p'(***n***), previous_edge(***n***)=***e***
  - Let ***best_segmentation*** is a empty array of edges
  - Let node ***n*** is the destination node
  - Repeat until ***n*** is the source node
    - Push previous_edge(***n***) to the head of ***best_segmentation***
    - Let ***n*** be the node where ***e*** start from
  - Return ***best_segmentation***

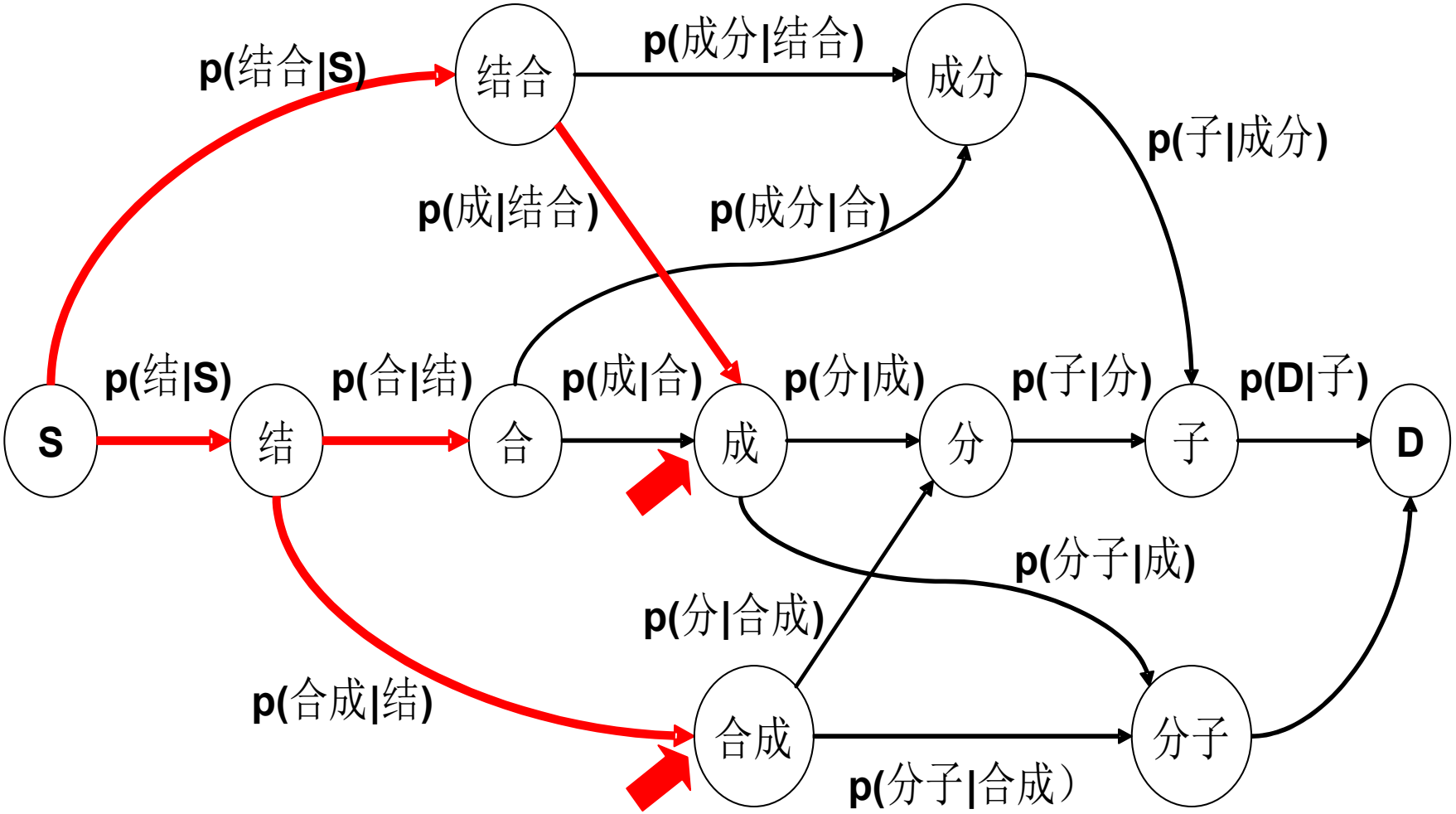
在词图上搜索最优路径： Viterbi 算法



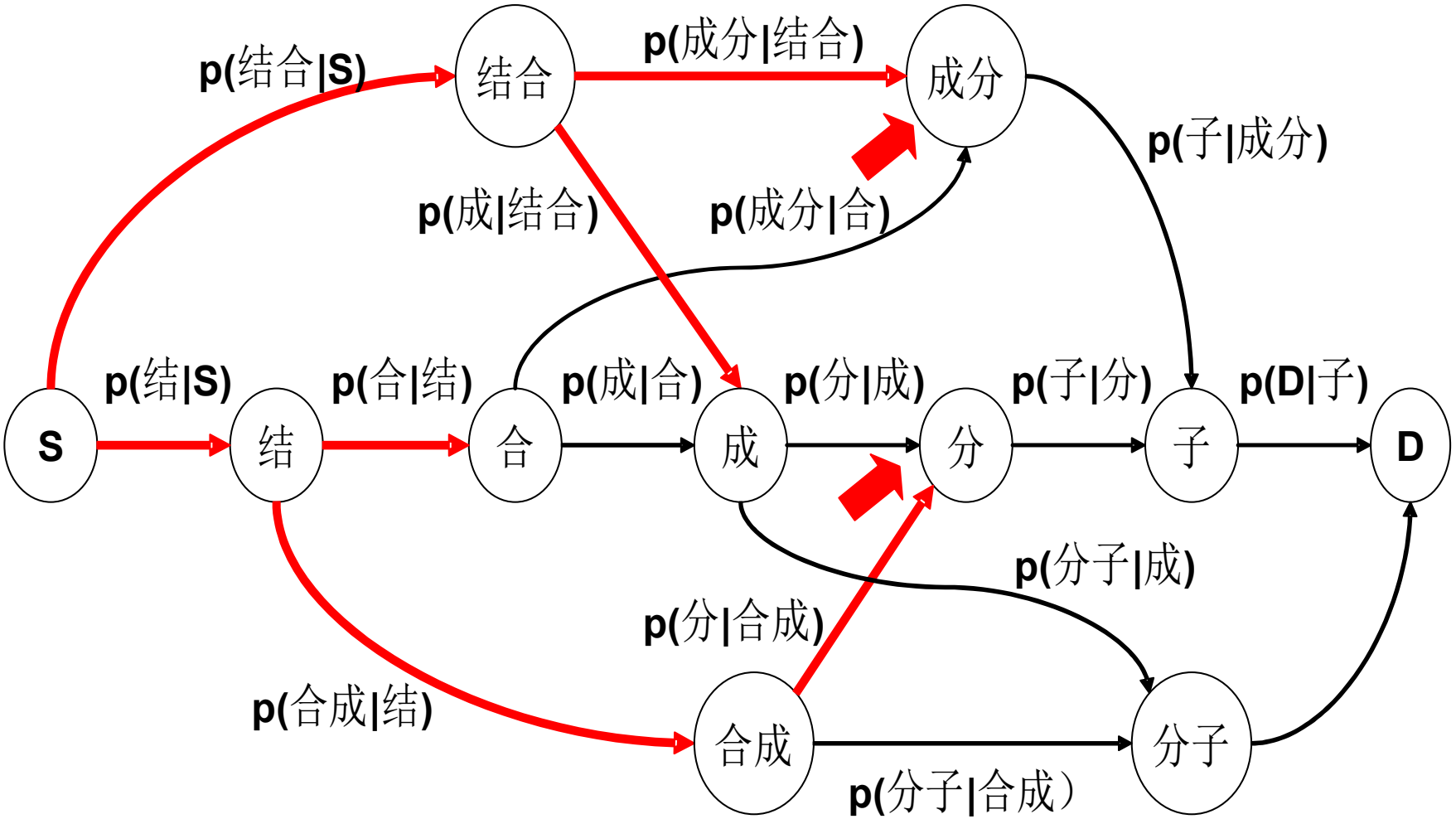
在词图上搜索最优路径： Viterbi 算法



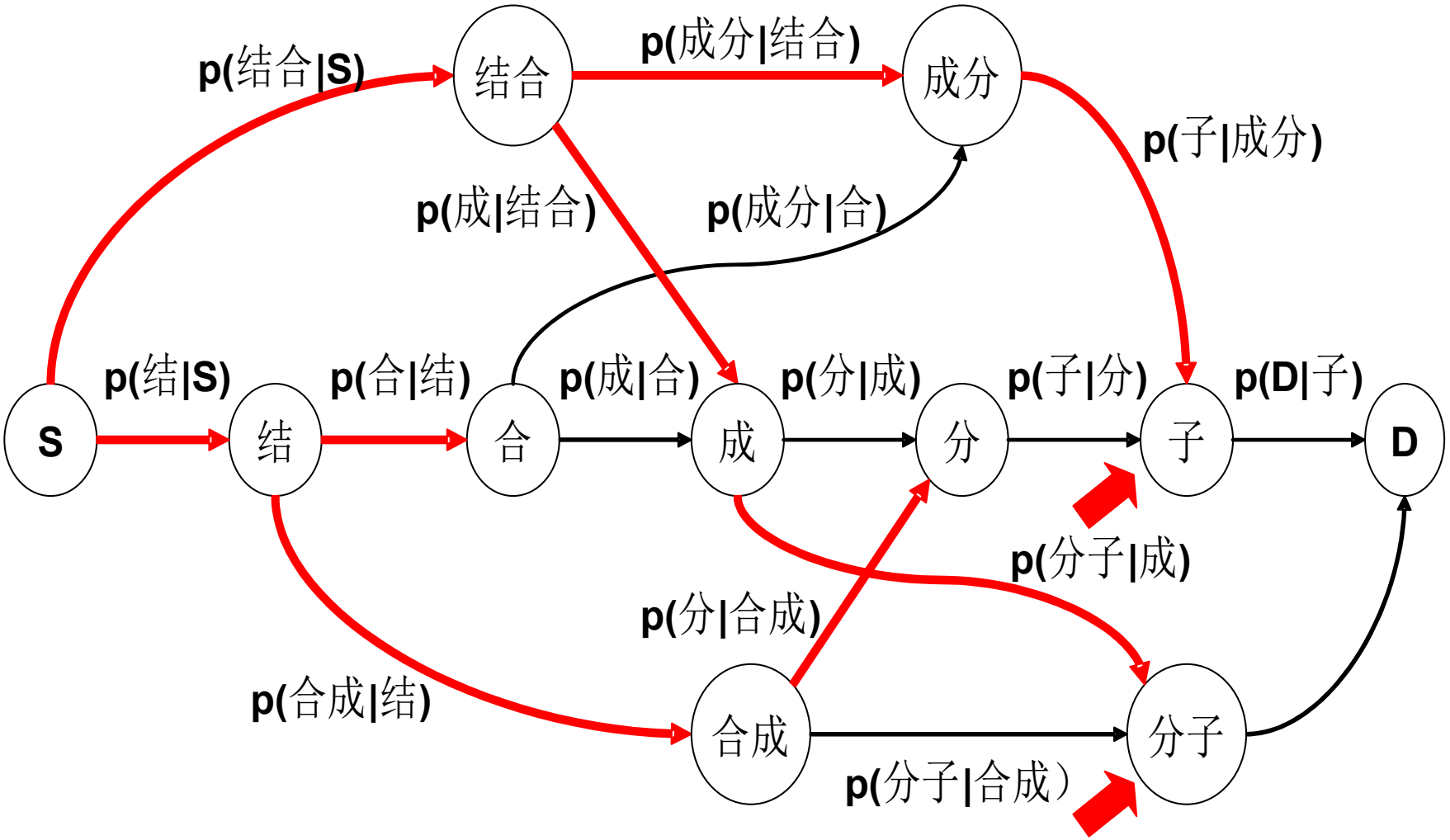
在词图上搜索最优路径： Viterbi 算法



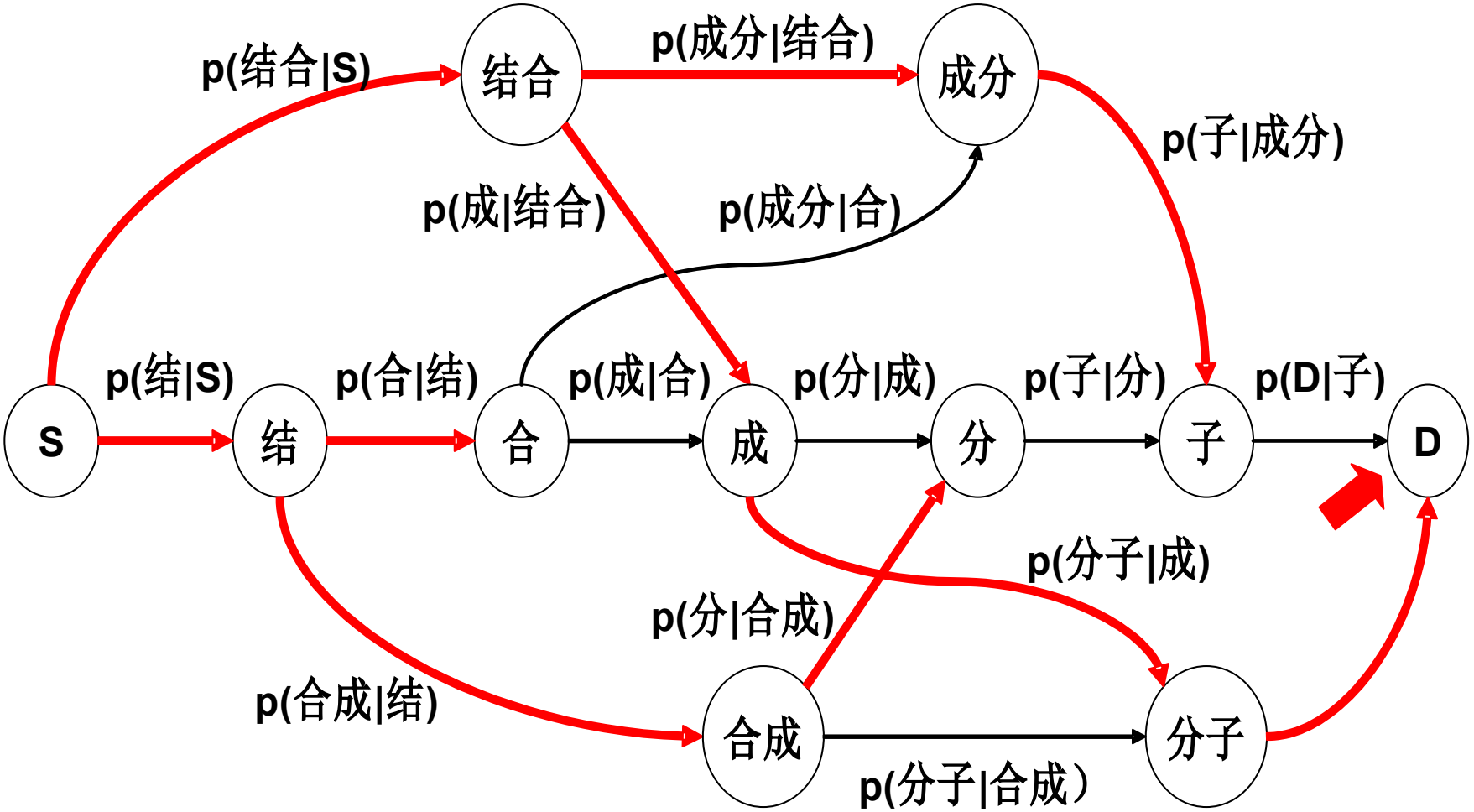
在词图上搜索最优路径： Viterbi 算法



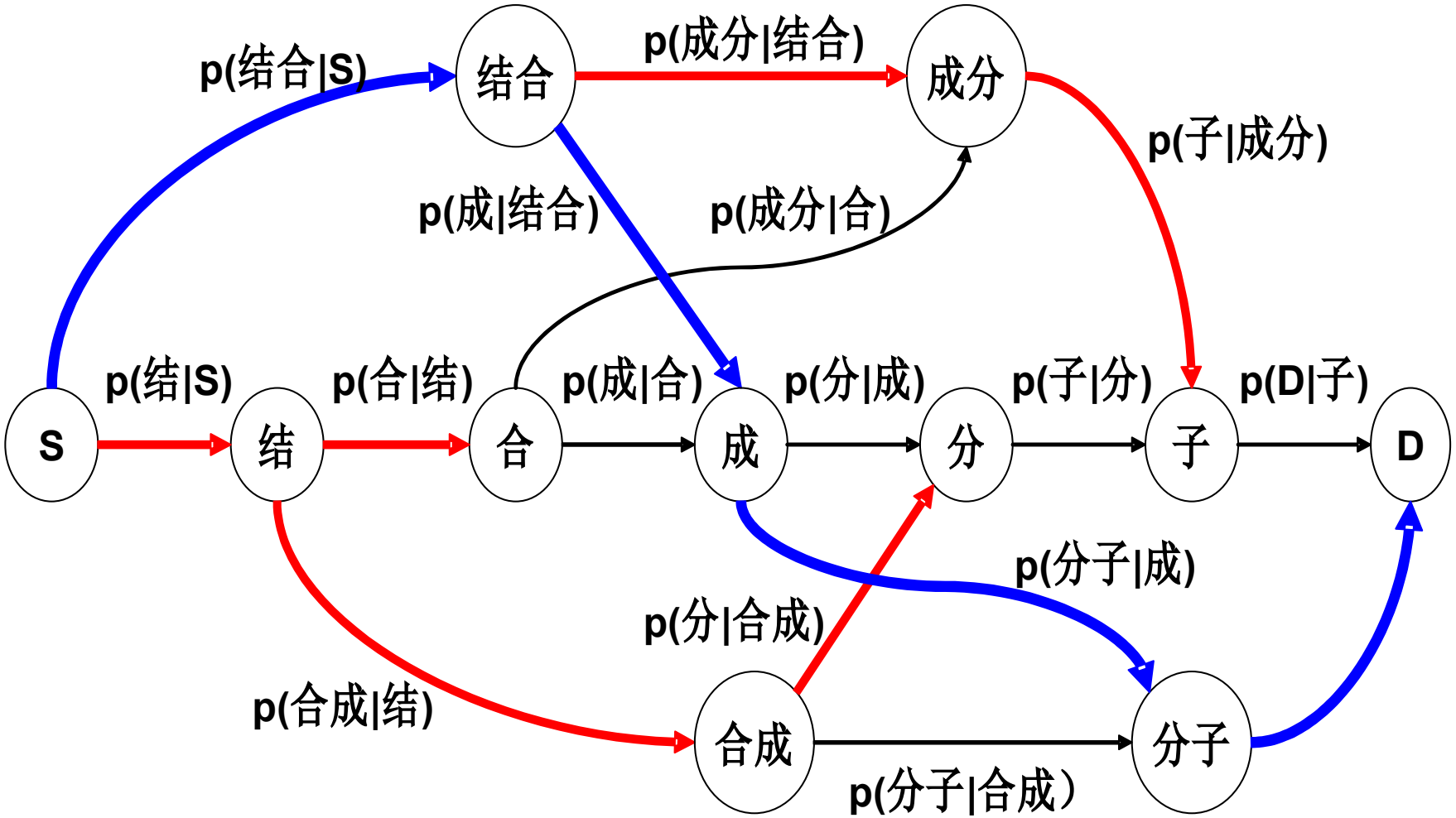
在词图上搜索最优路径： Viterbi 算法



在词图上搜索最优路径： Viterbi 算法



在词图上搜索最优路径： Viterbi 算法



基于 n 元语法模型的词语切分

- 采用一元语法
  - 把词频的负对数理解成“代价”，这种方法也可以理解为最短路径法的一种扩充
  - 正确率可达到 92%
  - 简便易行，效果一般好于基于词表的方法
- 采用三元语法
  - 实验表明，在较大规模数据上，采用三元语法进行词语切分正确率可以达到 98% 以上
- 缺点：无法识别未定义词

汉语词法分析

汉语词法分析所面临的问题

基于词典的汉语词语机械切分算法

基于语言模型的汉语词语切分算法

基于隐马尔科夫模型的词性标注算法

基于字标注的汉语词语切分标注一体化算法

词性标注（ POS-Tagging ）

- 词性标注：为句子中每一个词语加上词性标记
- 词性标记（ Part-of-Speech ， POS ）
  - 实词：名词 n 、动词 v 、形容词 a 、副词 d 、代词 r
  - 虚词：介词 p 、助词 u 、语气词 o 、感叹词
- 英语的词性
  - 英语的词性主要是由词的变化形式决定的
  - 英语的词性与词的句法功能存在着比较明确的一一对应关系
- 汉语的词性
  - 汉语词几乎没有形态变化
  - 汉语词性与所充当的语法功能不存在明确的一一对应关系

词性标注（POS-Tagging）

- 语法体系 —— 词性标记集的确定
- 一词多类现象
  - Time flies like an arrow.

    Time/n-v-a flies/v-n like/p-v an/Det arrow/n

  - 把这篇报道编辑一下

    把/q-p-v-n 这/r 篇/q 报道/v-n 编辑/v-n 一/m-c 下/f-q-v

汉语词性标记集

- 几个典型的词性标记集
  - 北京大学《人民日报》语料库标记集
  - 清华大学《汉语树库》词性标记集
  - 语用所《信息处理用现代汉语词类及词性标记集规范》
  - 宾州树库规范
  - 计算所词性标记集（V3.0）
- 参考：词性标记集对照表

词的兼类现象 (1)

兼类数	兼类词数	百分比	例词及词性标记
5	3	0.01%	和 c-n-p-q-v
4	20	0.04%	光 a-d-n-v
3	126	0.23%	画 n-q-v
2	1475	2.67%	锁 n-v
合计	1624	2.94%	总词数： 55191

数据来源：北大计算语言所《现代汉语语法信息词典》 1997 年版

词的兼类现象 (2)

兼类	词数	百分比	例词
n-v	613	42%	爱好，把握，报道
a-n	74	5%	本分，标准，典型
a-v	217	15%	安慰，保守，抽象
b-d	103	7%	长期，成批，初步
n-q	64	4%	笔，刀，口
a-d	30	2%	大，老，真
合计	1101	75%	兼两类词数： 1475

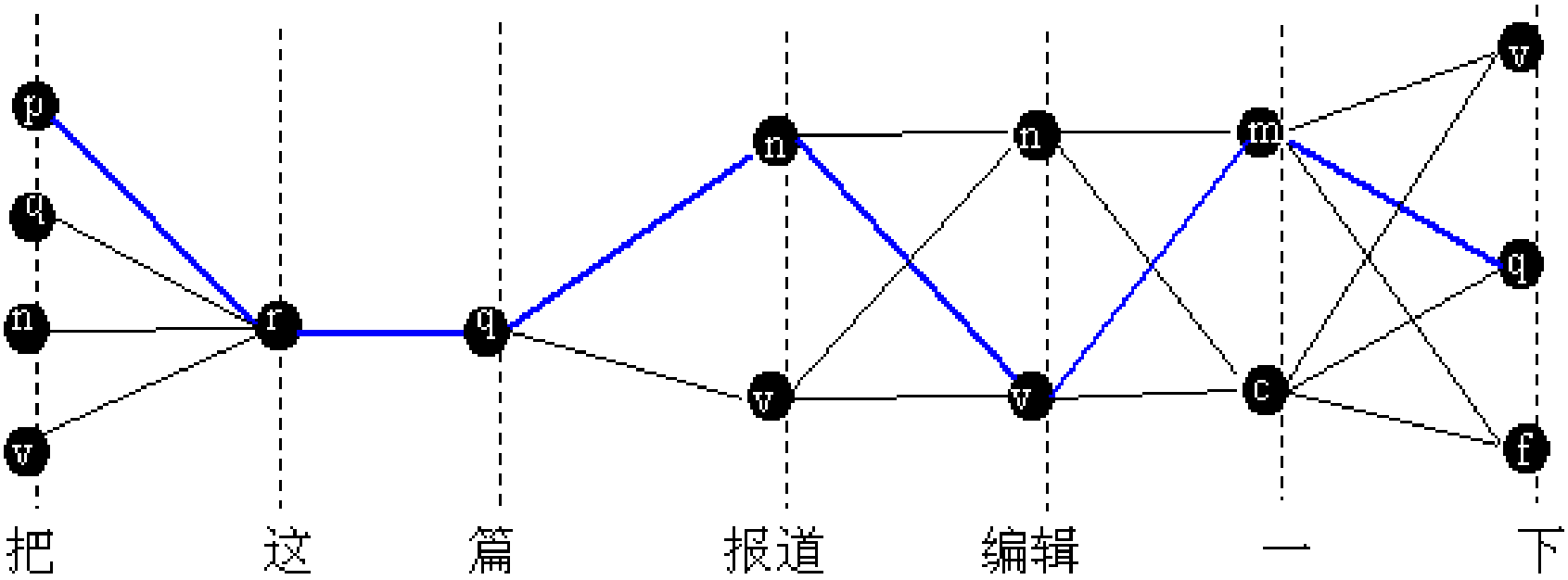
词的兼类现象 (3)
（English data, from Brown corpus）

引自： http://www.cs.columbia.edu/~becky/cs4999/04mar.html

10.4 percent of the lexicon is ambiguous as to part-of-speech (types)
40 percent of the words in the Brown corpus are ambiguous (tokens)

Degree of ambiguity	Total frequency (39,440)
1 tag	35,340
2-7 tags	4,100
2	3,760
3	264
4	61
5	12
6	2
7	1

词性标注：寻找最优路径



4×1×1×2×2×2×3 = 96 种可能性，哪种可能性最大？

词性标注方法回顾

序号	作者 / 标注项目	标记集	方法，特点	处理语料规模	精确率
1	Klein&Simmons (1963)	30	手工规则	百科全书小样本	90%
2	TAGGIT (Greene&Rubin, 1971)	86	人工规则 (3300 条)	Brown 语料库	77 %
3	CLAWS (Marshall,1983; Booth, 1985)	130	概率方法 效率低	LOB 语料库	96 %
4	VOLSUNGA (DeRose,1988)	97	概率方法 效率高	Brown 语料库	96 %
5	Eric Brill's tagger (1992-94)	48	机器规则 (447 条) 效率高	UPenn WSJ 语料库	97 %

隐马尔科夫模型—假设

对于一个随机事件，有一个观察值序列： $O_1,\ldots,O_T$

该事件隐含着一个状态序列： $X_1,\ldots,X_T$

假设 1 ：马尔可夫假设（状态构成一阶马尔可夫链）

$$p(X_i|X_{i-1}\ldots X_1) = p(X_i|X_{i-1})$$

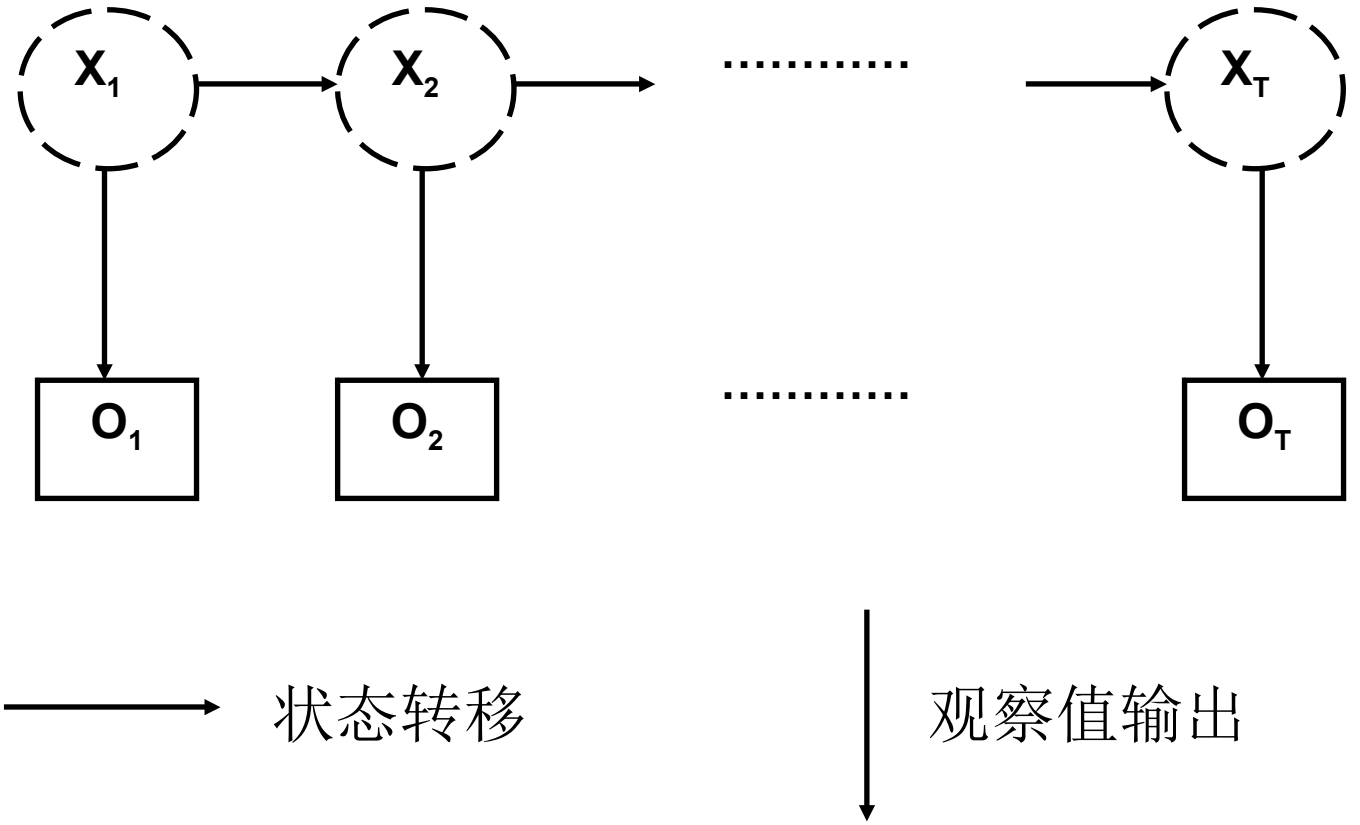
假设 2 ：不动性假设（状态与具体时间无关）

$p(X_{i+1}|X_i) = p(X_{j+1}|X_j)$ ，对任意 i,j 成立

假设 3 ：输出独立性假设（输出仅与当前状态有关）

$$p(O_1,\ldots,O_T \mid X_1,\ldots,X_T) = \Pi\, p(O_t \mid X_t)$$

隐马尔科夫模型—图示



隐马尔科夫模型—定义

一个隐马尔可夫模型 (HMM) 是一个五元组：

$$(\Omega_X , \Omega_O, A, B, \pi)$$

其中：

$\Omega_X = \{q_1,...q_N\}$：状态的有限集合

$\Omega_O = \{v_1,...,v_M\}$：观察值的有限集合

$A = \{a_{ij}\}$， $a_{ij} = p(X_{t+1} = q_j \mid X_t = q_i)$：转移概率

$B = \{b_{ik}\}$， $b_{ik} = p(O_t = v_k \mid X_t = q_i)$：输出概率

$\pi = \{\pi_i\}$， $\pi_i = p(X_1 = q_i)$：初始状态分布

隐马尔科夫模型—问题

令 $\lambda = \{A,B,\pi\}$ 为给定 HMM 的参数，

令 $\sigma = O_1,\ldots,O_T$ 为观察值序列，

隐马尔可夫模型（ HMM ）的三个基本问题：

1. 评估问题：对于给定模型，求某个观察值序列的概率 $p(\sigma|\lambda)$ ；（语言模型）
2. 解码问题：对于给定模型和观察值序列，求可能性最大的状态序列；
3. 学习问题：对于给定的一个观察值序列，调整参数 λ，使得观察值出现的概率 $p(\sigma|\lambda)$ 最大。

隐马尔科夫模型—算法

- 评估问题：向前算法
  - 定义向前变量
  - 采用动态规划算法，复杂度 $O(N^2T)$
- 解码问题：韦特比（ Viterbi ）算法
  - 采用动态规划算法，复杂度 $O(N^2T)$
- 学习问题：向前向后算法
  - EM 算法

隐马尔科夫模型—例子

- 假设：某一时刻只有一种疾病，且只依赖于上一时刻疾病
  一种疾病只有一种症状，且只依赖于当时的疾病
- 症状 (观察值)：发烧，咳嗽，咽喉肿痛，流涕
- 疾病 (状态值)：感冒，肺炎，扁桃体炎
- 转移概率：从一种疾病转变到另一种疾病的概率
- 输出概率：某一疾病呈现出某一症状的概率
- 初始分布：初始疾病的概率
- 解码问题：某人症状为：咳嗽→咽喉痛→流涕→发烧
  请问：其疾病转化的最大可能性如何？

隐马尔科夫模型—例子（续）

- 转移概率

	感冒	肺炎	扁桃体炎
感冒	0.4	0.3	0.3
肺炎	0.2	0.6	0.2
扁桃体炎	0.1	0.1	0.8

- 输出概率

	发烧	咳嗽	咽喉痛	流涕
感冒	0.4	0.3	0.1	0.2
肺炎	0.3	0.5	0.1	0.1
扁桃体炎	0.2	0.1	0.6	0.1

- 初始分布

感冒	肺炎	扁桃体炎
0.5	0.2	0.3

HMM 学习问题—最大似然估计

已知观察序列 O 对应的状态序列为（有指导学习）：

$$X = X_1 X_2 \ldots X_T$$

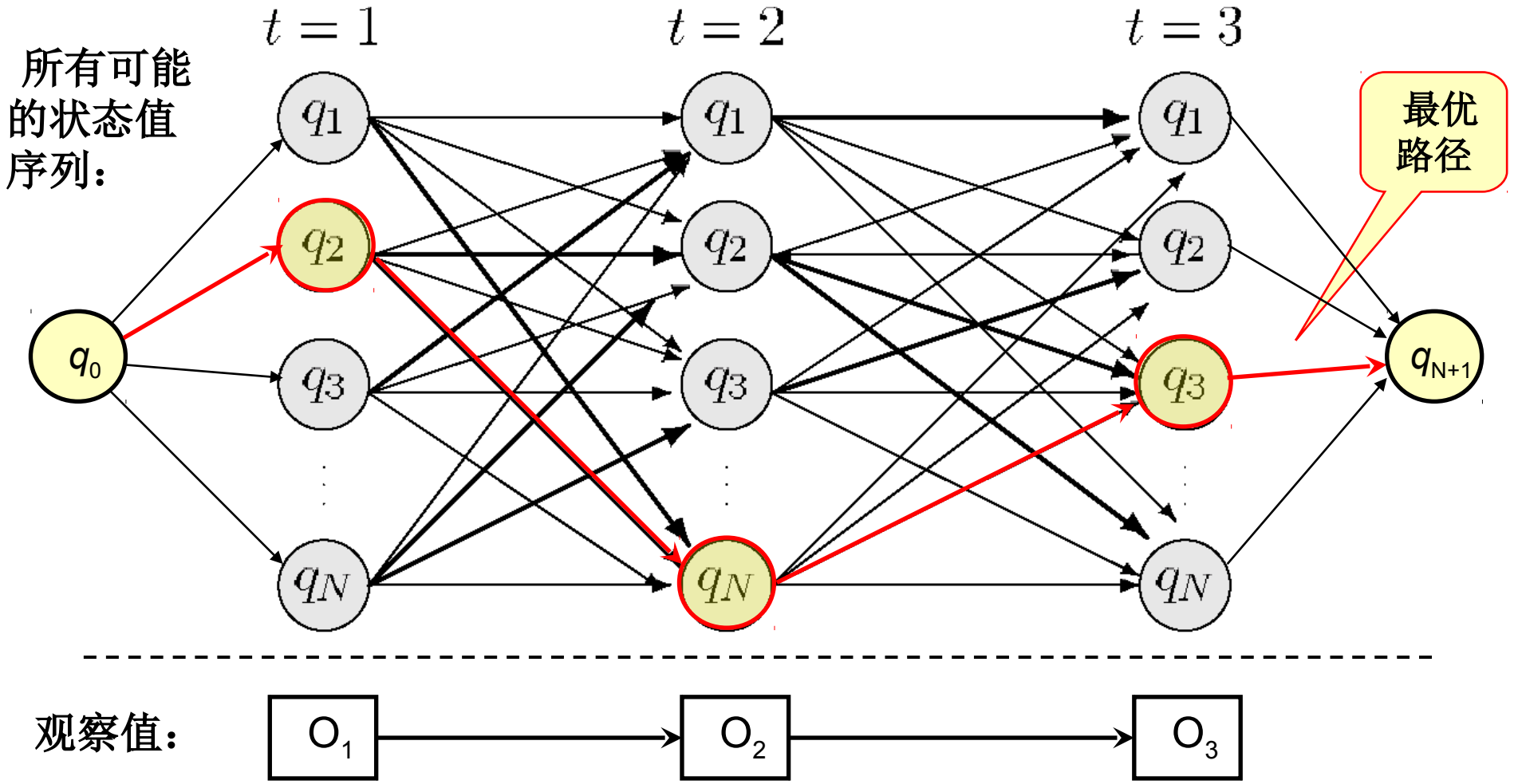
采用最大似然估计：

$$\bar{\pi}_i = \delta(X_1, q_i), \text{其中} \delta(x, y) = \begin{cases} 1, \text{如果} x = y \\ 0, \text{如果} x \neq y \end{cases}$$

$$\bar{a}_{ij} = \frac{X \text{中从状态} q_i \text{转移到状态} q_j \text{的次数}}{X \text{中从状态} q_i \text{转移到另一状态(含} q_j \text{)的次数}} = \frac{\sum_{t=1}^{T-1} \delta(X_t, q_i) \times \delta(X_{t+1}, q_j)}{\sum_{t=1}^{T-1} \delta(X_t, q_j)}$$

$$\bar{b}_{jk} = \frac{X \text{中从状态} q_j \text{输出到观察值} v_k \text{的次数}}{X \text{中到达状态} q_j \text{的次数}} = \frac{\sum_{t=1}^{T} \delta(X_t, q_j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \delta(X_t, q_j)}$$

HMM 解码问题



可能的状态序列有 N^T 种

HMM 解码问题— Viterbi 算法 (1)

定义 Viterbi 变量为 HMM 在时间 t 沿着某一条路径到达状态 q_i，且输出观察值 $O_1O_2\ldots O_i$ 的最大概率

$$\delta_t(i) = \max_{X_1 X_2 \ldots X_{i-1}} P(X_1 X_2 \ldots X_t = q_i, O_1 O_2 \ldots O_t | \lambda)$$

HMM 解码问题— Viterbi 算法 (2)

- 初始化

$$\delta_t(i)=\pi_i b_{iO_1}$$

- 迭代计算

$2\le t\le T$

$1\le j\le N$

$$\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_{jO_t}$$

$$\Psi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_{jO_t}$$

- 取最优

$$p^*=\max_{1\le i\le N}[\delta_T(i)]$$

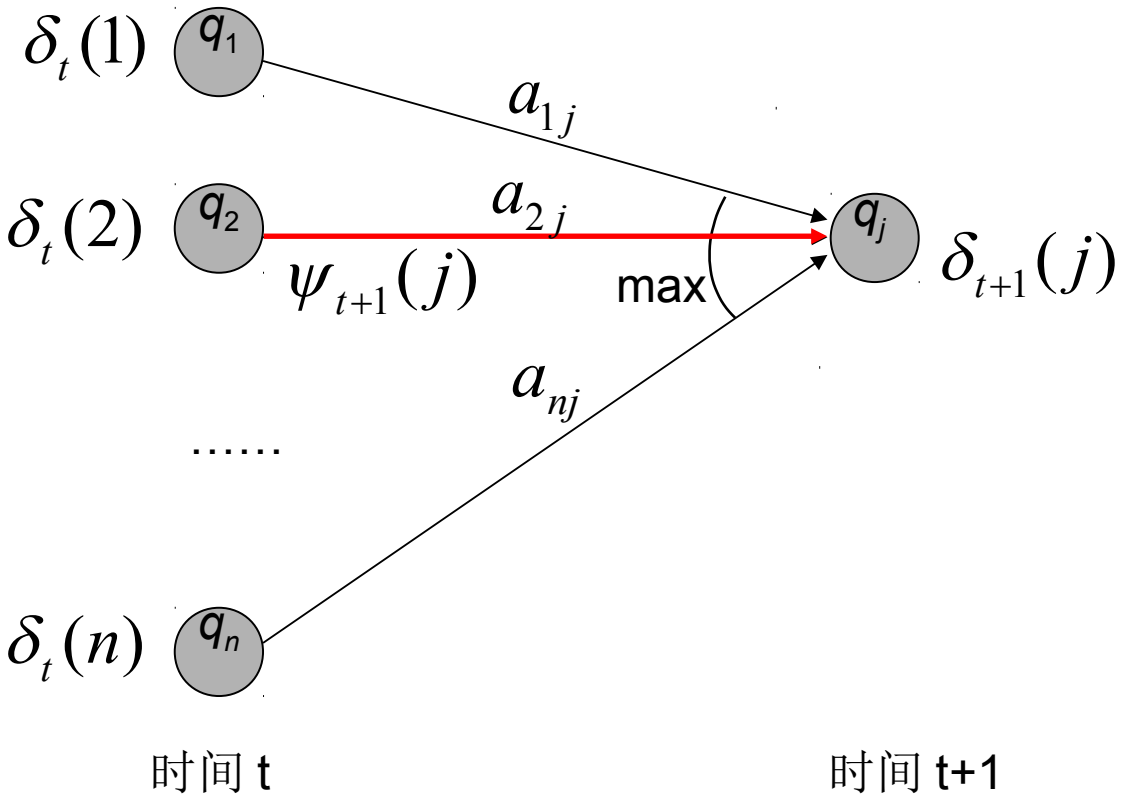
$$q_T^*=\arg\max_{1\le i\le N}[\delta_T(i)]$$

- 路径回溯

$2\le t\le T$

$$q_t^*=\Psi_{t+1}(q_{t+1}^*)$$

HMM 解码问题— Viterbi 算法 (3)



Viterbi 算法的时间复杂度： $O(N^2T)$

HMM 解码问题— Viterbi 算法 (4)

- **Viterbi Algorithm**
  - **Assign p(source_state)=1**
  - **For each observation *o* from source to destination**
    - **For each possible state *n* of observation *o***
      - **p(*n*)=0, previous_state(*n*)=ø**
      - **For each edge e directed to *n* from *n*'**
        - » **p'(*n*)=p(*n*')×transition_probability(*n*'|*n*)×output_probability(*o*|*n*)**
        - » **If p'(*n*) > p(*n*) then p(*n*)=p'(*n*), previous_state(*n*)= *n*'**
  - **Let *best_tag_sequence* is a empty array of states**
  - **Let state *n* is the destination state**
  - **Repeat until *n* is the source state**
    - **Push previous_state(*n*) to the head of *best_tag_sequence***
    - **assign *n*= previous_state(*n*)**
  - **Return *best_tag_sequence***

基于 HMM 进行词性标注 (1)

- 把词汇序列（记做 $W=w_1w_2\ldots w_n$ ）理解为观察值
- 把词性标注序列（记做 $T=t_1t_2\ldots t_n$ ）理解为隐含的状态值
- 词性标注问题变成 HMM 中的解码问题
- 已知词串 W （观察序列）和模型参数 λ 情况下，求使得条件概率 $P(T|\Omega, \lambda)$ 值最大的那个 T’，一般记做：

$$T' = \arg\max_{T} P(T|W, \lambda)$$

基于 HMM 进行词性标注 (2)

利用 Bayes 公式，可以进一步分解为：

$$\arg\max_T P(T|W) = \arg\max_T P(T)P(W|T)$$

其中：

$$P(T) = P(t_1|t_0)P(t_2|t_1,t_0)\ldots P(t_i|t_{i-1},t_{i-2},\ldots)$$

根据 HMM 假设，可得

$$P(T) \approx P(t_1|t_0)P(t_2|t_1)\ldots P(t_i|t_{i-1})$$

词性之间的转移概率可以从语料库中估算得到：

$$P(t_i|t_{i-1}) \approx \frac{\text{训练语料中 } t_i \text{ 出现在 } t_{i-1} \text{ 之后的次数}}{\text{训练语料中 } t_i \text{ 出现的总次数}}$$

基于 HMM 进行词性标注 (3)

P(W|T) 是已知词性标记串，产生词串的条件概率：

$$P(W|T)=P(w_1|t_1)P(w_2|t_2,t_1,w_1)...P(w_i|t_i,t_{i-1},...,t_1,w_i,w_{i-1},...,w_1)$$

根据 HMM 假设，上面公式可简化为：

$$P(W|T)\approx P(w_1|t_1)P(w_2|t_2)...P(w_i|t_i)$$

已知词性标记下输出词语的概率可以从语料库中统计得到：

$$P(w_i|t_i)\approx\frac{\text{训练语料中}\,w_i\,\text{的词性被标记为}\,t_i\,\text{的次数}}{\text{训练语料中}\,t_i\,\text{出现的总次数}}$$

基于 HMM 进行词性标注示例

- 把 /?　这 /?　篇 /?　报道 /?　编辑 /?　一 /?　下 /?
  把 /q-p-v-n　这 /r　篇 /q　报道 /v-n　编辑 /v-n　一 /m-c　下 /f-q-v

P(T1|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(f|m)P(下 |f)

P(T2|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(q|m)P(下 |q)

P(T3|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(v|m)P(下 |v)

.....

P(T96|W) = P(n|$)P(把 |n)P(r|q)P(这 |r)...P(v|c)P(下 |v)

从中选一个最大值

词性转移概率

词语输出概率

基于 HMM 进行词性标注 (4)

- 基于 HMM 的词性标注实际上对应着 HMM 的解码问题
- 采用 Viterbi 算法即可解决

汉语词法分析

汉语词法分析所面临的问题

基于词典的汉语词语机械切分算法

基于语言模型的汉语词语切分算法

基于隐马尔科夫模型的词性标注算法

基于字标注的汉语词语切分标注一体化算法

基于字标注的中文词法分析

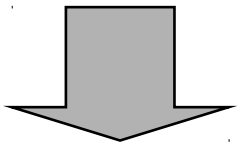
- Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03*, pages 176–179, Sapporo, Japan.

空挡标注

- 最简单的分词方案，可以理解为：

  对句子中每两个汉字之间的空挡判断是否进行切分

费 0 孝 0 通 1 向 1 人 0 大 1 报 0 告

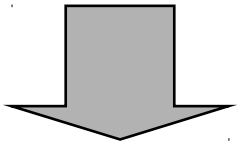


费孝通 向 人大 报告

字标注

- 对每一个汉字进行标注 {B,M,E,S}：
  - B：词首字
  - M：词中字
  - E：词尾字
  - S：单字词

费/B 孝/M 通/E 向/S 人/B 大/E 报/B 告/E



费孝通 向 人大 报告

空挡标注与字标注的转换

- 上述两种标注是可以转换的：
  - 字标注可以通过该字左右的空挡标注得到：
    - B➔10
    - M➔00
    - E➔01
    - S➔11

更复杂的字标注

- Hai Zhao, Chang-Ning Huang, and Mu Li, An Improved Chinese Word Segmentation System with Conditional Random Field, Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pp.162-165, Sydney, Australia, July 22-23, 2006
- 采用基于字的六标注集合： B 、 B_1 、 B_2 、 M 、 E 、 S
  - 单字词： S
  - 两字词： BE
  - 三字词： BB_1E
  - 四字词： BB_1B_2E
  - 五字词： BB_1B_2ME
  - 六字词： BB_1B_2MME
- 问题：六字标注集如何表示为空挡标注？

字标注模型

- 字标注（或空挡标注）都是序列标注问题
- 理论上，字标注问题也可以采用语言模型或者隐马尔科夫模型来解决
- 但由于标记集太小，采用语言模型和隐马尔科夫模型很难取得很好的效果：语言模型和隐马尔科夫模型的区分能力太弱

更复杂的字标注模型

- 最大熵模型
- 最大熵马尔科夫模型
- 条件随机场模型
- 感知机模型

最大熵原理

- Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., (1996), A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Volume 22, No. 1
- 自然语言处理的最大熵模型，常宝宝，北京大学
- 自然语言处理中的最大熵方法（ PPT 讲义），马金山，哈尔滨工业大学信息检索研究室

  （本讲义部分内容源自马金山 PPT ，特此感谢）

最大熵工具

- 最普遍使用的工具：
  - 名称：
    Maximum Entropy Modeling Toolkit for Python and C++
  - 作者：张乐（东北大学博士生）
  - 主页：
    http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

基于最大熵模型字标注的汉语词语切分

- 采用最大熵模型对每个汉字标注 BMES 标记
- 假设当前字 C_0 ， 当前标记是 T_0
- 常用最大熵特征模板（当前字是 C_0 ）：
  - $C_n T_0 (n = -2, -1, 0, 1, 2)$ ：汉字
  - $C_n C_{n+1} T_0 (n = -2, -1, 0, 1)$ ：两字组
  - $C_{-1} C_1 T_0$ ： 当前字左右两个字
  - $D(C_0) T_0$ ： 当前字是否数字
  - $A(C_0) T_0$ ： 当前字是否字母
  - $P(C_0) T_0$ ： 当前字是否标点

生成最大熵训练实例

- 假设给定训练样本：

  <s> 今天　是　星期三　。 </s>

- 生成对应标注序列：

  今 |B　天 |E　是 |S　星 |B　期 |M　三 |E　。 |S

- 对于每一个字生成一个训练实例。

生成最大熵训练实例

- 上例中“天”字生成的训练实例：

$C_{-2}T_0$=<s>E	$C_{-1}T_0$= 今 E	C_0T_0= 天 E	C_1T_0= 星 E
C_2T_0= 期 E	$C_{-2}C_{-1}T_0$=<s> 今 E	$C_{-1}C_0T_0$= 今天 E	
$C_0C_1T_0$= 天星 E	$C_1C_2T_0$= 星期 E	$C_{-1}C_1T_0$= 今是 E	
$D(C_0)T_0$= 否	$A(C_0)T_0$= 否	$P(C_0)T_0$= 否	

- 这里列出了该实例中被激活（值为 1 ）的所有特征，其他所有未列出的特征均未被激活（值为 0 ）

最大熵模型的参数训练

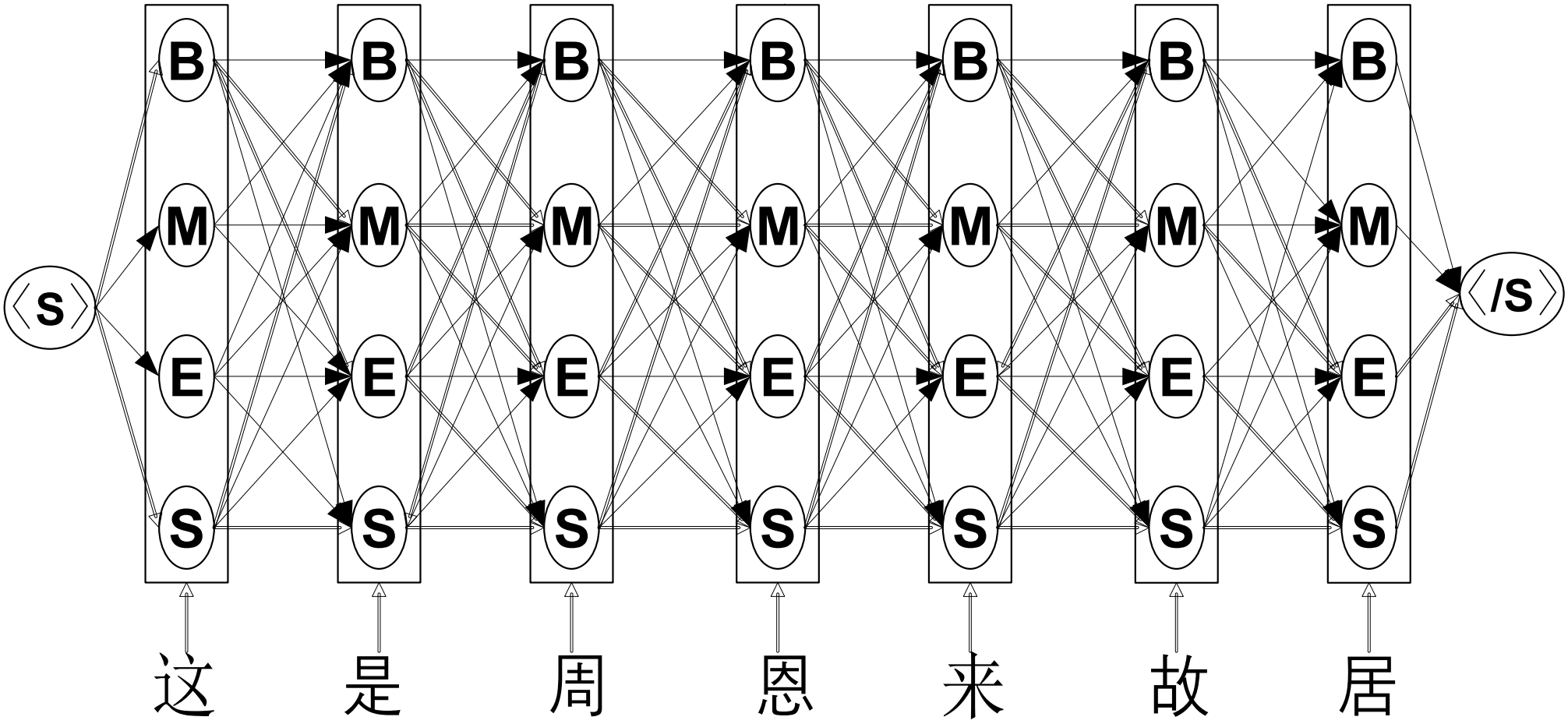
- 对训练语料库中每个汉字生成一个实例
- 把所有实例的列表送给最大熵模型的训练工具
- 最大熵模型的训练工具将为每一个训练实例生成一个参数 λ

最大熵模型的使用

- 输入的汉字串，如“昨天是星期五”
- 对于输入串中每一个汉字（这里假设是“天”字），以及该汉字的每一个可能的标记，生成一个实例，这样对每个汉字就生成了四个实例，如 E(天→B),E(天→M),E(天→E),E(天→S)。
- 对这每个实例 E(天→X)，生成其所有激活的特征，计算这些特征所对应的参数 λ 之和 N(天→X)
- 可能性最大的标记为：

$$X=\max_{X'\in\{B,M,E,S\}} N(\text{天}\rightarrow X')$$

搜索最优标注路径

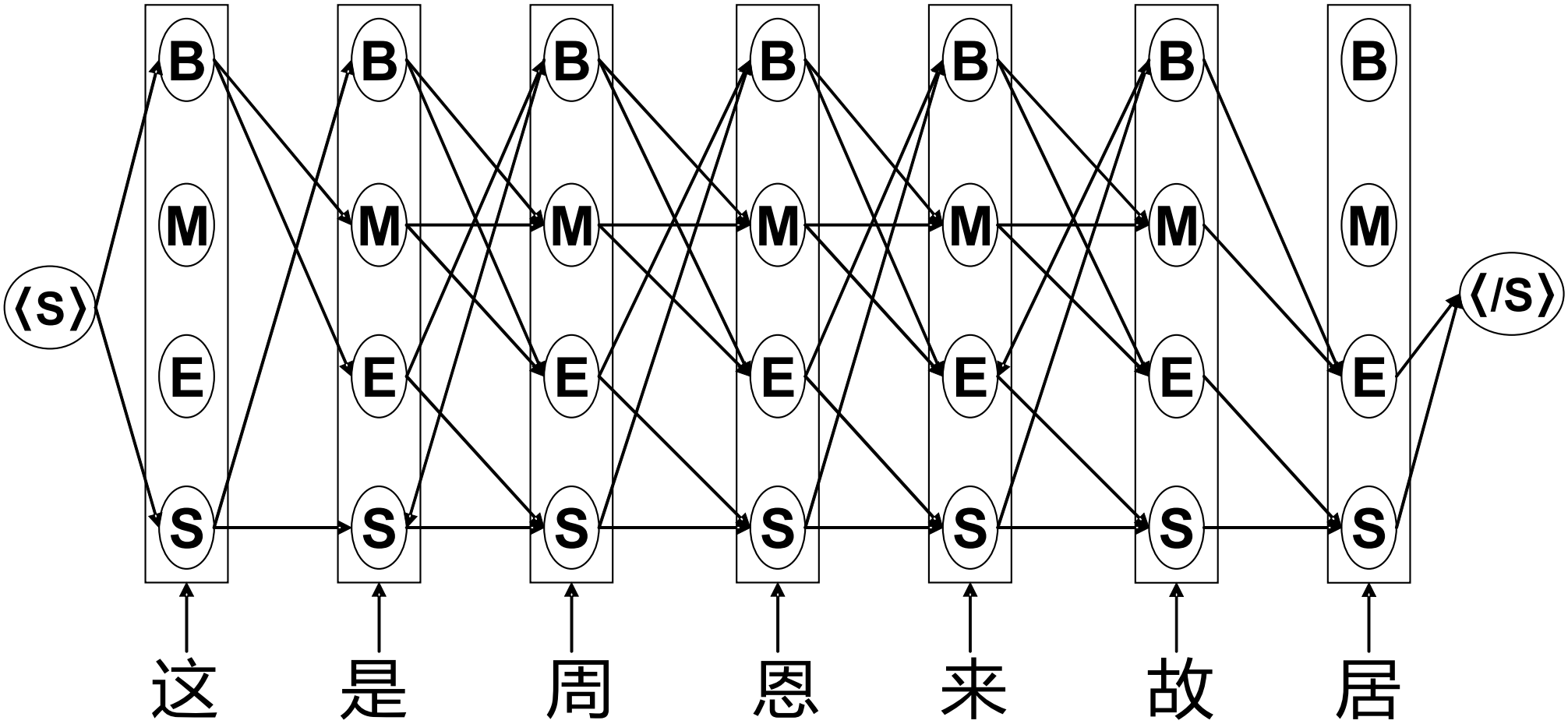


搜索最优标注路径

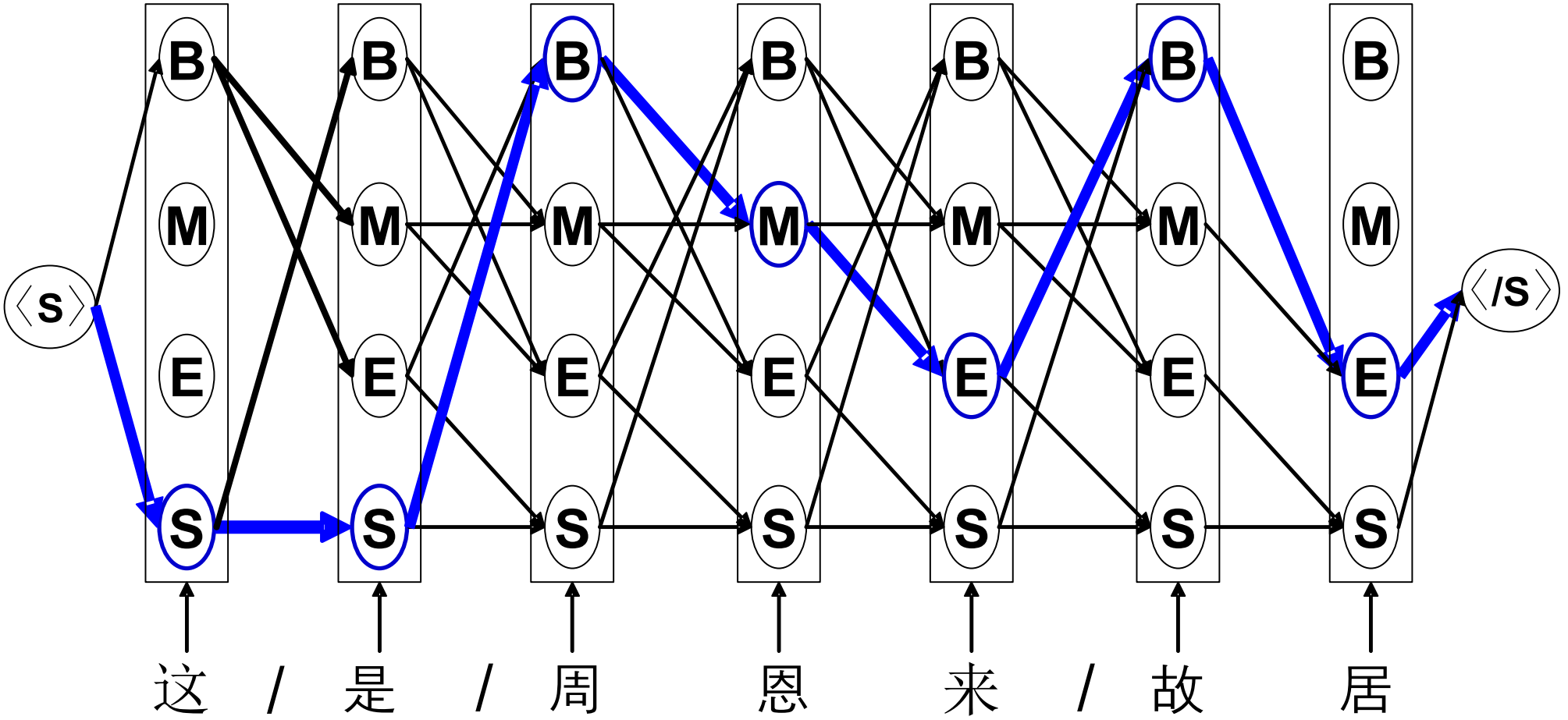
- 有些边是非法的，比如
  M→B， E→M， E→E， S→M， S→E，等等

  注意：这里的非法边和未定义词识别的 BMEO 标记转移图的非法边略有不同，想一想为什么
- 每条路径的概率是路径上各节点标注概率 P(标记 | 汉字) 之积，边上没有转移概率
- 搜索算法与 HMM 模型解码的 Viterbi 算法相同

搜索最优标注路径



搜索最优标注路径



基于字标注方法的汉语词语切分标注一体化方法

- 融入词性标注：
  - 扩充标记集
  - 为每一个词性定义 BMES 四个标记

小结

- 词法分析、形态分析
- 语言的分类
- 英语的 Tokenization
- 英语的形态分析
- 汉语词法分析面临的问题
- 汉语词语机械切分方法
- 基于语言模型的汉语词语切分方法
- 基于隐马尔科夫模型的词性标注方法
- 基于字标注的汉语切分标注一体化方法