第 10 章　词 法 分 析

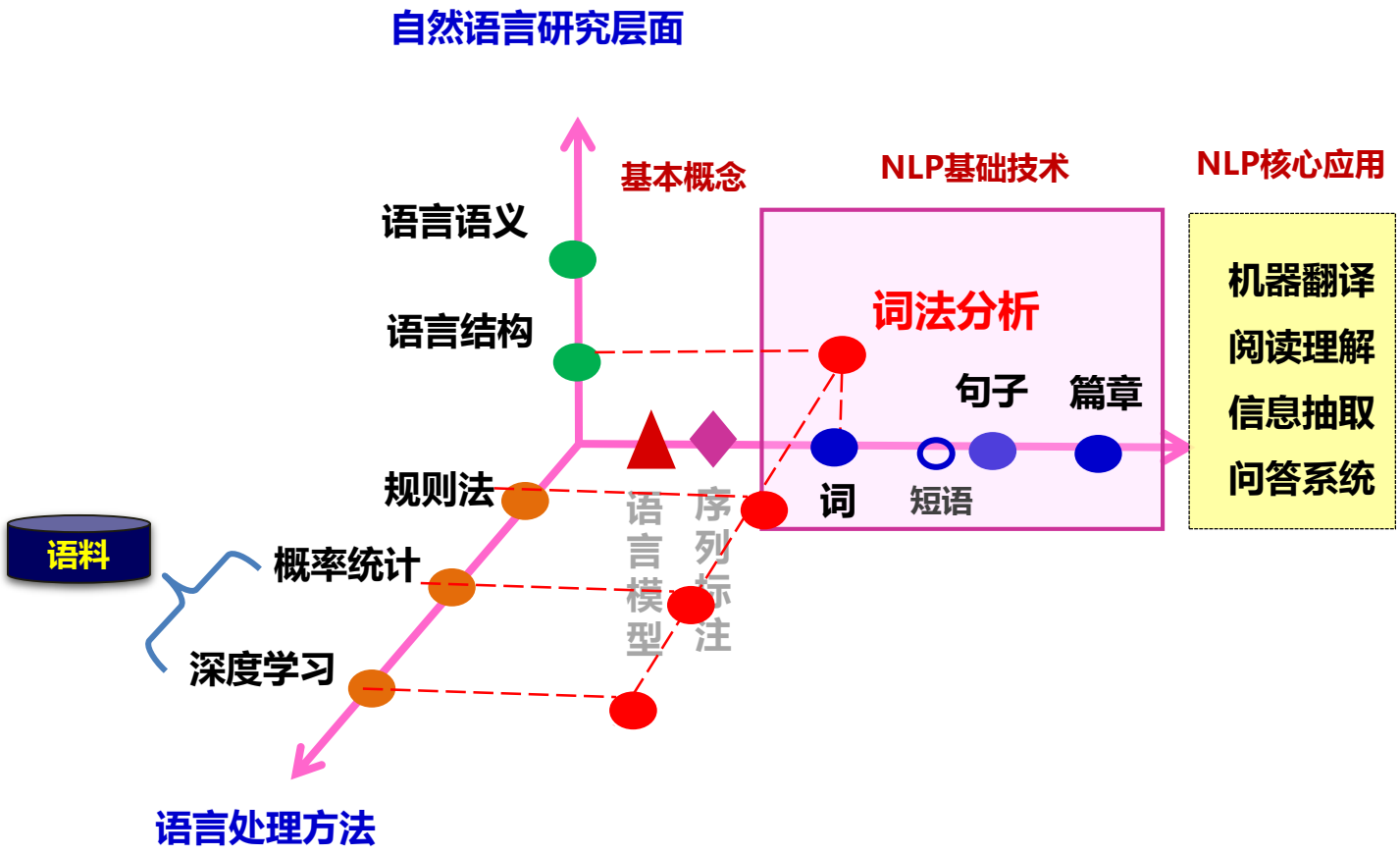
中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

自然语言处理课程内容及安排

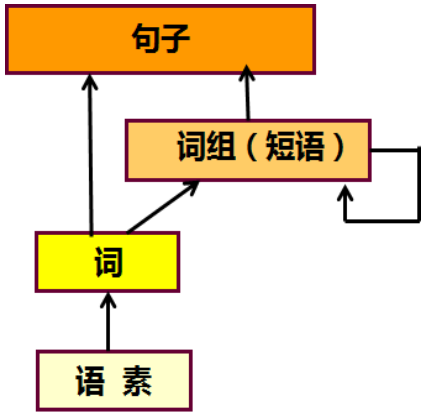
✧ 课程内容：



内 容 提 要

10.1 词法分析概述

词：



词 是最小的能够独立运用的语言的单位

词法分析 是句子分析、语义分析，文本分类，信息检索，机器翻译，机器问答等问题的基础。会对后续问题产生影响。

词法分析任务：

将输入句子字串转换成 词序列 并标记出各词的词性。

这里所说的“字”不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在文本中的文字符号，所有这些字符都是构词词的基本单元。从形式上看，词是稳定的字的组合

10.1 词法分析概述

不同的语言词根据其语言自身特点，具体做法有所不同

如：

英语（曲折语）

- 特点
  - 用空格隔开，无需分词
  - 用词形态变化来表示语法关系
- 英文的词法分析
  - 英文词识别、词形还原
  - 未登录词识别
  - 词性标注

汉语（孤立语）

- 特点
  - 词与词紧密相连，没有明显的分界标志
  - 词形态变化少，靠词序或虚词来表示
- 中文的词法分析
  - 分词
  - 未登录词识别
  - 词性标注

中文分词是其他中文信息处理的基础

在很多领域有广泛的应用

10.1 词法分析概述

中文分词任务：

原始句子

警察正在详细调查事故原因

自动分词

分词结果

警察/ 正在/ 详细/ 调查/ 事故/ 原因

词性标注

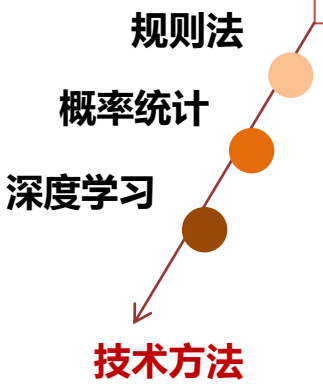
词性标注结果

警察/NN 正在/AD 详细/AD 调查/VV 事故/NN 原因/NN

10.1 词法分析概述

中文分词问题及处理技术：

任务	目标	需解决问题
自动分词	将输入汉字串切成词串	1. 歧义问题 2. 未登录词问题 3. 分词标准问题
词性标注	确定每个词的词性并加以标注	词性兼类歧义问题



一个成熟的分词系统，不可能单独依靠某一种算法来实现，都需要综合不同的算法来处理不同的问题。

内 容 提 要

10.2.1 自动分词

自动分词需要解决的问题：

任务	目标	需解决问题
自动分词	将输入汉字串切成词串	1. 歧义问题 2. 未登录词问题 3. 分词标准问题
词性标注	确定每个词的词性并加以标注	词性兼类歧义问题

规则法

概率统计

深度学习

技术方法

自动分词

原始句子

警察正在详细调查事故原因

分词结果

警察/ 正在/ 详细/ 调查/ 事故/ 原因

问题？

10.2.1 自动分词

1. 歧义：

切分歧义：对同一个待切分字符串存在多个分词结果。

分为交集型歧义、组合型歧义和混合歧义三种类型

- **交集型歧义**

字串abc 既可切分为ab/c,又可以切分为a/bc，其中a，ab，c和bc是词

如： “研究生命” => “研究/生命” 或 “研究生/命”

“白天鹅” => “白天/鹅”或“白/天鹅”

到底取哪一个，则要靠语境或上下文的意思来决定

链长： 交集型切分歧义所拥有的交集串的个数称为链长

如: (1) “中国产品质量” {国，产，品，质}，歧义字段的链长为 4；

(2) “部分居民生活水平” {分，居，民，生，活，水}，链长为 6

10.2.1 自动分词

- **组合型歧义**

若ab为词，而a和b在句子中又可分别单独成词。

如：

门把手弄坏了 =>“门/ 把/ 手/ 弄/ 坏/ 了”

或 “门/ 把手/ 弄/ 坏/ 了”

- **混合歧义**

以上两种情况通过嵌套、交叉组合等而产生的歧义。

如：

这篇文章写的太平淡了=>

“太平淡”是交集型歧义，而“太平”是组合型歧义

切分歧义是影响分词系统切分正确率的重要因素约占分词错误的10% 左右

10.2.1 自动分词

2.未登录词

未登录词：词典中没有收录过的人名、地名、机构名、专业术语、译名、新术语等。该问题在文本中出现频度远远高于歧义问题。

未登录词的类型：

- 实体名称（Named Entity）
  - 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
  - 机构名：方正公司 联想集团 国际卫生组织 外贸部
- 数字、日期词、货币等
- 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

未登录词问题是分词错误的主要来源

10.2.1 自动分词

3.分词标准

◆ 汉语中什么是词？不仅普通人有词语认识上的偏差，即使是语言专家，在这个问题上依然有不小的差异。

- **相关的标准**
  - 《信息处理用汉语分词规范》
    - GB/T13715-92，中国标准出版社，1993
  - 《资讯处理用中文分词规范》台湾中研院
  - 《人民日报》语料库词语切分规范
  - ……

“缺乏统一的分词规范和标准”这种问题反映在分词语料库上。不同语料库的数据无法直接拿过来混合训练。为克服此问题大家在统一的平台上进行实验和比较。

10.2.1 自动分词

- 不同的应用对分词粒度的需求不同。如，一般在搜索引擎中，构建索引时和查询时会使用不同的分词粒度。在索引的时候常使用细粒度的分词以保证召回，在查询的时候常使用粗粒度的分词以保证精度；在问答系统中，需要对文本实现较为深入的理解，对分词和实体识别的准确性要求很高。

从已有工程经验来看，几乎不存在通用而且效果非常好的分词系统

10.2.1 自动分词

自动分词技术方法：

任务	目标	需解决问题
自动分词	将输入汉字串切成词串	1. 歧义问题 2. 未登录词问题 3. 分词标注问题
词性标注	确定每个词的词性并加以标注	词性兼类歧义问题

规则法

概率统计

深度学习

技术方法

自动分词

原始句子

警察正在详细调查事故原因

分词结果

警察/ 正在/ 详细/ 调查/ 事故/ 原因

技术方法？

10.2.1 自动分词

1. 基于字典、词库匹配的分词方法（机械分词法）

该类算法是按照一定的策略将待匹配的字符串和一个已建立好的“充分大的”词典中的词进行匹配，若找到某个词条，则说明匹配成功，识别了该词。

主要有

- 正向最大匹配法（由左到右的方向）
- 逆向最大匹配法（由右到左的方向）
- 最少切分（使每一句中切出的词数最小）
- 双向最大匹配法（进行由左到右、由右到左两次描）

基于词典的分词算法在传统分词算法中是应用最广泛、分词速度最快的一类算法。其优点是实现简单，算法运行速度快

缺点是严重依赖词典，无法很好的处理分词歧义和未登录词

10.2.1 自动分词

- **最大匹配法 (Maximum Matching, MM)**

基本思想：先建立一个最长词条字数为L的词典, 然后取从前（后）按正向（逆向）取句子前L个字查词典，如查不到, 则去掉最后一个字继续查, 一直到找着一个词为止。最大匹配算法以及其改进方案是基于词典和规则的。其优点是实现简单，算法运行速度快，缺点是严重依赖词典，无法很好的处理分词歧义和未登录词。

如：假设词典中最长单词的字数为 7。

输入字串：他是研究生物化学的。

正向最大匹配切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/。

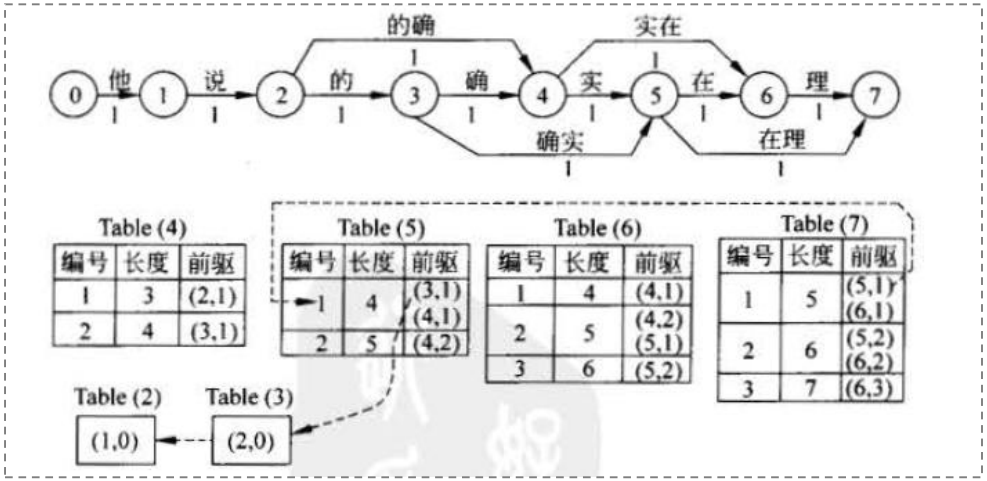
逆向最大匹配 切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/。

10.2.1 自动分词

- **最少分词法(最短路径法)**

基本思想:设待切分字串 S=c_1 c_2...c_n，其中c_i (i =1, 2, ..., n) 为单个的字， n 为串的长度，n≥1。建立一个节点数为n+1的切分有向无环图G，如果 $w= c_ic_{i+1}...c_j$ ($0<i<j\leq n$) 是一个词，则节点v_{i-1}, v_j之间建立有向边 ，从产生的所有路径中，选择路径最短的(词数最少的)作为最终分词结果。

例： 输入字串: 他说的确实在理。



解：使用dijkstra 贪心算法

可能输出:

他/ 说/ 的/ 确实/ 在理/ 。(5)

他/ 说/ 的确/ 实在/ 理/ 。 (5)

10.2.1 自动分词

优点：需要的语言资源（词表）不多。

弱点：对许多歧义字段难以区分，最短路径有多条时，选择最终的输出结果缺乏应有的标准；字串长度较大和选取的最短路径数增大时，长度相同的路径数急剧增加，选择最终正确的结果困难越来越越大。

10.2.1 自动分词

2. 基于统计的方法

统计方法具有较强的歧义区分能力，但需要大规模标注 (或预处理) 语料库的支持，需要的系统开销也较大。

2.1 基于词的分词方法

基于词 的生成模型主要考虑词汇之间以及词汇内部字与字之间的依存关系。大部分基于词的分词方法采用的是生成式模型 (Generative model)

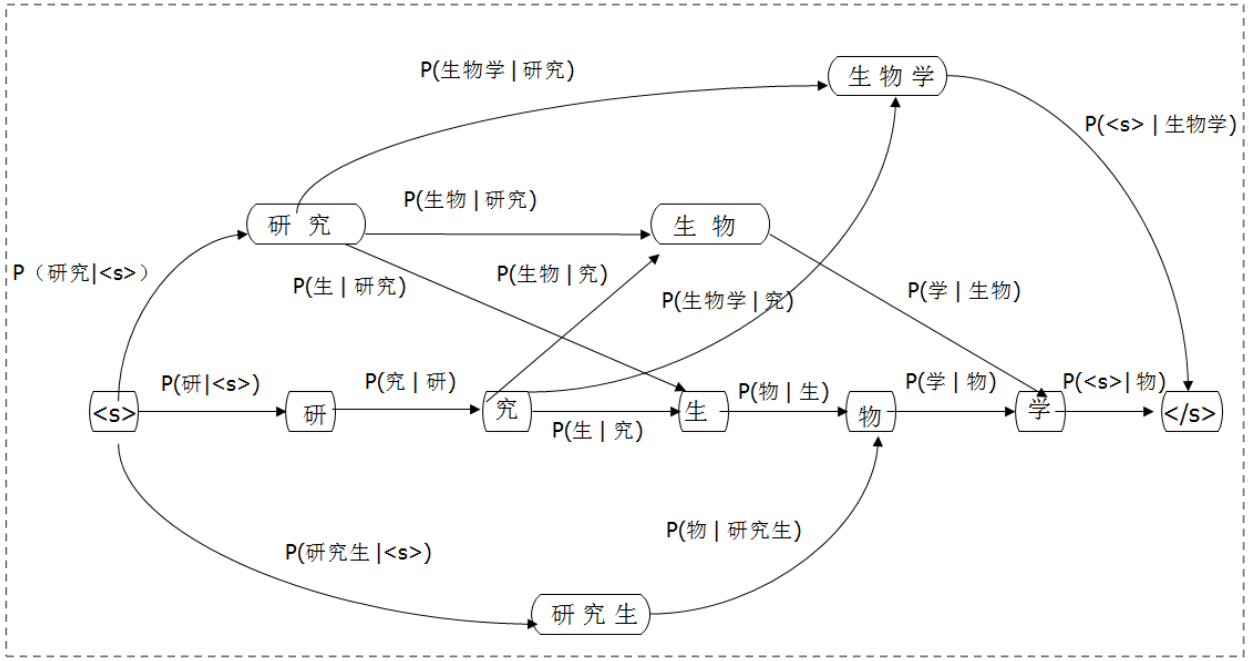
$$WSeq^{*} = \arg\max_{WSeq} p(WSeq \mid c_1^n) \quad = \arg\max_{WSeq} p(WSeq)$$

如， 3-gram: $p(w_1^m) = \prod_{i=1}^{m} p(w_i \mid w_1^{i-1}) \approx \prod_{i=1}^{m} p(w_i \mid w_{i-2}^{i-1})$

10.2.1 自动分词

- **n元语法模型方法**

例： 输入句子：研究生物学。



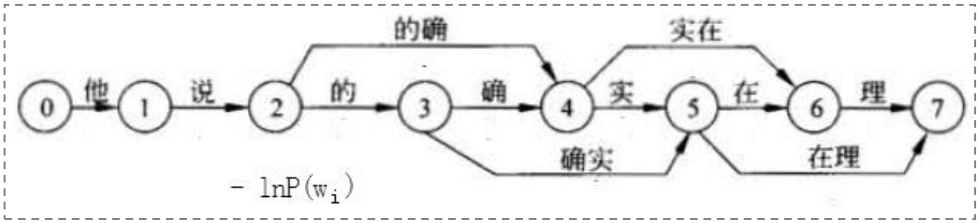
利用二元模型： $p(s) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_2) \times \ldots \times p(w_n|w_{n-1})$

可求得最大概率切分

10.2.1 自动分词

- **改进最短路径法（统计粗分模型）**

最短路径法



改进： 将边的权重改为该词出现频率P（w_i）

切分结果为： MAX $P(W) = \prod_{i=1}^{m} P(w_i)$

解：使用dijkstra 贪心算法 解法：

$$令\ P^*(W) = -\ln P(W) = \sum_{i=1}^{m}[-\ln P(w_i)]$$

MAX P(w) → Min P*(w)

将边的权重变为 $-\ln P(w_i)$ 即可求得最短路径（切分结果）

10.2.1 自动分词

2. 基于统计的方法

2.2 基于字的序列标注分词方法

Nianwen Xue 在其论文《Combining Classifiers for Chinese Word Segmentation》中首次提出对每个字符进行标注的方法。这类方法具有较强的歧义区分能力，和未登录词识别能力。但需要大规模标注 (或预处理) 语料库的支持，需要的系统开销也较大。训练文本的选择将影响分词结果

常用的是模型有 **HMM、CRF、SVM** 等

基于字的序列标注分词方法主要的优势在于能够平衡地看待词表词和未登录词的识别问题。单独用统计法缺点是学习算法的复杂度往往较高，计算代价较大，依赖手工定义的特征工程；目前方法是利用神经网络自动学习特征的优势将两者结合。

10.2.1 自动分词

- **用HMM实现简单的中文分词**

例 .　输入： 北京是中国的首都

输出： 词序列

解 ： 用单字序列标注方法

{ 词首/B，词内/I，词尾/E，单字词/O }

模型HMM：

S：状态集合， { B，I，E，O}

O：观察值集合，{单个汉字： 人、民、中.....}

A：状态转移概率矩阵

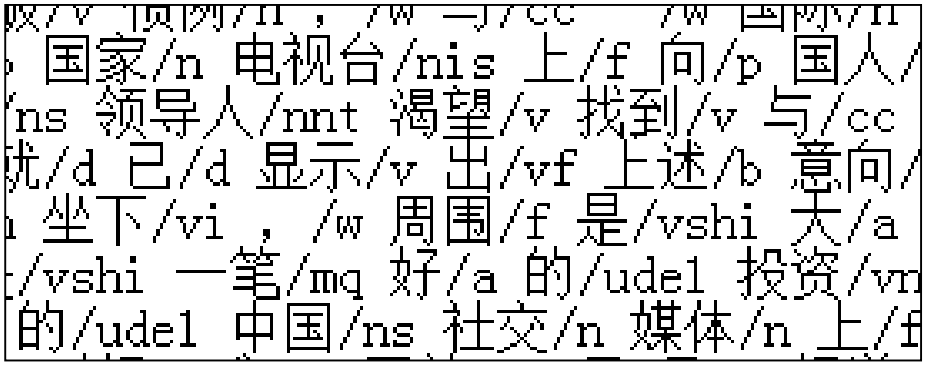
B：给定状态下，观察值的概率分布

π：初始状态空间的概率分布

10.2.1 自动分词

参数学习

语料：



训练语料：　国/B 家/E 电/B 视/I 台/E 上/O 向/O 国/B人/E 领/B 导/I 人/E....

10.2.1 自动分词

训练语料：

国/B 家/E 电/B 视/I 台/E 上/O 向/O
国/B人/E 领/B 导/I 人/E....

假设，语料中不重复的中文单字共8000个

- $A =$

	B	I	E	O
B	0	0.3	0.7	0
I	0	0.4	0.6	0
E	0.4	0	0	0.6
O	0.5	0	0	0.5

$A \in R^{4\times4}$，每行元素之和为1

- $B =$

	国	家	电	视	台	上	向	国	人	....
B	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	
I	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	....
E	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	....
O	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	....

$B \in R^{4\times8000}$，每行元素之和为1

- $\pi = [XXX, 0, 0, XXX]^T$ $\pi \in R^4$，元素之和为1

10.2.1 自动分词

用最大似然估计学习参数：

有观察序列$O=O_1O_2...O_T$ 和 状态序列$Q=q_1q_2.....q_T$

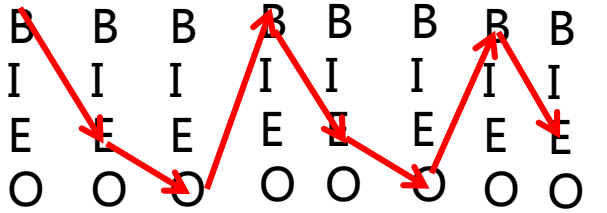
用极大似然估计

- $\pi_i = \frac{\sum_{t=1}^{T} \delta(q_t, S_i)}{T}$, $(S_0=\mathrm{B}, S_1=\mathrm{I}, S_2=\mathrm{E}, S_3=\mathrm{O})$
- $a_{ij} = \frac{\sum_{t=1}^{T-1} \delta(q_t, S_i) \times \delta(q_{t+1}, S_j)}{\sum_{t=1}^{T-1} \delta(q_t, S_i)}$
- $b_{jk} = \frac{\sum_{t=1}^{T} \delta(q_t, S_j) \times \delta(o_t, v_k)}{\sum_{t=1}^{T} \delta(q_t, S_j)}$

10.2.1 自动分词

预测-分词

Viterbi算法



输入： 北 京 是 中 国 的 首 都

输出： B E O B E O B E

分词结果： 北京/ 是/ 中国/ 的/ 首都

- $A =$

	B	I	E	O
B	0	0.3	0.7	0
I	0	0.4	0.6	0
E	0.4	0	0	0.6
O	0.5	0	0	0.5

- $B =$

	国	家	电	视	台	上	向	国	人	……
B	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	
I	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	……
E	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	……
O	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	……

- $\pi = [XXX, 0, 0, XXX]^T$

注意： 分词和词性标注虽均用HMM模型， 但 状态集 观察集 不同，训练语料标注不同，模型参数不同

10.2.1 自动分词

- **用CRF实现简单的中文分词**

关键问题：选取特征 和 构造特征函数

对于含有n个字的汉语句子 $c_1^n = c_1\ c_2 \ldots\ldots\ c_n$

$$P(t_1^n \mid c_1^n) = \prod_{k=1}^{n} P(t_k \mid t_1^{k-1}, c_1^n) \approx \prod_{k=1}^{n} P(t_k \mid c_{k-2}^{k+2})$$

其中：t_k表示第K 个字的词位， $t_k \in$ {B, M, S, E}

特征模板：

（a） c_k(k= −2,−1,0,1,2)

（b） $c_k\ c_{k+1}$ (k= −2,−1,0,1,)

（c） $c_{-1}\ c_1$

（d） $T(c_{-2})\ T(c_{-1})\ T(c_0)\ T(c_1)\ T(c_2)$

前三类是窗口内字及组合特征；

$T(c_i)$ 是指字c_i的字符类别。如，阿拉伯数字，数字中文，英文字母等

10.2.1 自动分词

评价：该方法采用基于字 的区分式模型，其主要优势在于，它能够平衡地看待词表词和未登录词的识别问题，文本中的词表词和未登录词都是用统一的字标注过程来实现。在学习构架上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块，因此，大大地简化了分词系统的设计

10.2.1 自动分词

2. 基于统计的方法

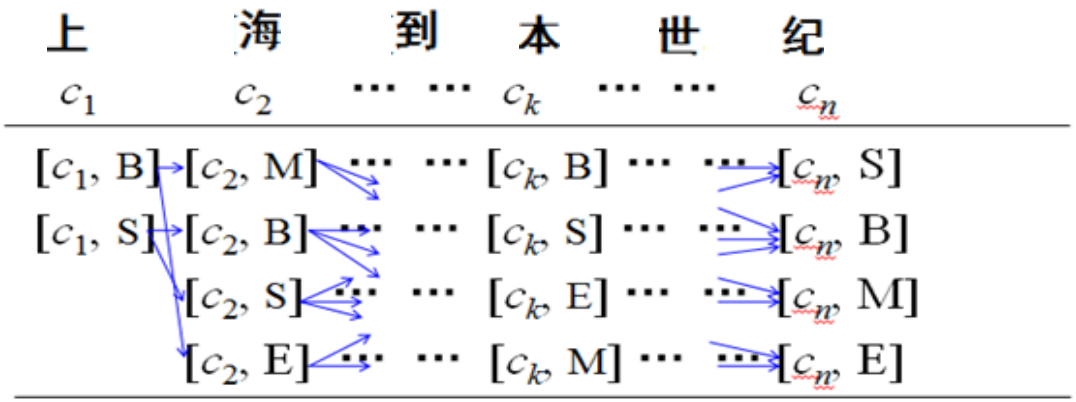
2.3生成式方法与区分式方法的结合

将基于字模型的有利于处理集外词的优势和基于词模型的考虑了词汇之间以及词汇内部字与字之间依存关系的长处结合起来。

将待切分字串的每个汉字用$[c, t]_i$ 替代，以$[c, t]_i$ 作为基元，利用语言模型选取全局最优(生成式模型)。

10.2.1 自动分词

如： **上海到本世纪**



标签t∈ {B,M,E,S}

语言模型：(3-gram)

$$p(w_1^m) = \prod_{i=1}^{m} p(w_i \mid w_1^{i-1}) \approx \prod_{i=1}^{m} p(w_i \mid w_{i-2}^{i-1}) \Rightarrow P([c,\ t]_1^n) \approx \prod_{i=1}^{n} P([c,\ t]_i \mid [c,\ t]_{i-2}^{i-1})$$

P(上海到本世纪) = P ([上,t] [海,t] [到,t] [本,t] [世,t] [纪,t])

分词结果： MAX P ([上,t] [海,t] [到,t] [本,t] [世,t] [纪,t])

= [上,B] [海,E] [到,S] [本,S] [世,B] [纪,E]

10.2.1 自动分词

2. 基于统计的方法

2.4未登录词识别方法

在统计方法中，各种不同类型的未登录词识别方法思想大同小异，通常的做法就是将识别问题转化成序列标注问题

关键问题：特征选取 即 找各种未登录词的构成规律

构成规律

- 内部构成规律（用字规律）
- 外部环境（上下文）
- 重复出现规律

早期的命名实体识别大都用基于规则的方法； 系统实现代价高，可移植性差。20世纪90年代后期以来，基于统计的机器学习方法被成功的应用。目前要用机器学习方法.

10.2.1 自动分词

- **中国人名识别**

中国人名的内部构成规律

中国人名组成：
- 姓：张、王、李、刘、诸葛、西门、范徐丽泰
- 名：李素丽，张华平，王杰、诸葛亮

身份词：
- 前：工人、教师、影星、犯人
- 后：先生、同志
- 前后：女士、教授、经理、总理

地名或机构名：
- 前：信工所张三

的字结构：
- 前：年过七旬的李四

动作词：
- 前：批评，逮捕，选举
- 后：说，表示，吃，结婚

10.2.1 自动分词

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

人名地名冲突

- 河北省刘庄

中国目前仍使用的姓氏共737 个，其中，单姓729 个，复姓8 个。

10.2.1 自动分词

- **中文机构名识别**

中文机构名称内部构成规律

中文机构名称的构成

- 词法角度：偏正式 {名词|形容词|数量词|动词} + 名词
- 句法角度：“定语 + 名词性中心语”型的名词短语(定名型短语)
- 中心语：机构称呼词，如：大学，学院，研究所，学会，公司等。

中文机构名称的类型

- 地名，如：北京大学，武汉大学
- 人名，如：中山大学，哈佛大学
- 学科、专业和部门系统，如：公安部，教育委员会
- 研究、生产或经营等活动的对象，如：软件研究所，卫星制造厂
- 上述情况的综合，如：白求恩医科大学

10.2.1 自动分词

中文机构名称识别困难

- 中文机构名用词非常广泛
- 机构名长度极其不固定
- 机构名很不稳定。随着社会发展，新机构不断涌现，旧机构不断被淘汰、改组或更名。
- 公司名涉及到公司全称和公司简称的识别

10.2.1 自动分词

- **中国地名的识别**

中文地名组成

地名可以分为典型地名和非典型地名，典型地名如国、省、市、县、乡、村等；非典型地名还包括路、居委会、大厦商场、门牌单元、图书馆、门面等。理论上，只要是有经纬度坐标的实体，都可以纳入地名识别范畴。

中文地名识别困难

– 地名数量大，缺乏明确、规范的定义。《中华人民共和国地名录》（1994）收集88026个，不包括相当一部分街道、胡同、村庄等小地方名称。

– 真实语料中地名出现情况复杂。如地名简称、地名用词与其它普通词冲突、地名是其它专用名词的一部分，地名长度不一等。

10.2.1 自动分词

- **未登录词识别的研究进展**
  - 很成熟：
    - 数字、日期、货币词
  - 较成熟
    - 中国人名、译名
    - 中国地名
  - 较困难
    - 商标字号
    - 机构名
  - 很困难
    - 专业术语
    - 缩略语
    - 新词语

10.2.1 自动分词

3. 深度学习方法

近些年有些学者提出用深度网络模型来对中文进行分词

主要分为两类：一是基于字符的中文分词，另一类是基于词的中文分词。其方法主要用到了RNN，CNN，GNN等深度神经网络来自动地获取特征，从而代替传统方法中手工定义的特征。从句子中获取简单的特征改为获取复杂的特征，从单一语料库单一标准的模型改进为可以使用多语料进行分词等等，文献见如下列表：

10.2.1 自动分词

第一类：基于字符的分词方法：基本思想是根据字所在词的位置，对每个字打上标签

1. Deep Learning for Chinese Word Segmentation and POS Tagging
2. Long Short-Term Memory Neural Networks for Chinese Word Segmentation
3. Gated Recursive Neural Network for Chinese Word Segmentation
4. Adversarial Multi-Criteria Learning for Chinese Word Segmentation
5 . Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation
6 . A Feature-Enriched Neural Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging

10.2.1 自动分词

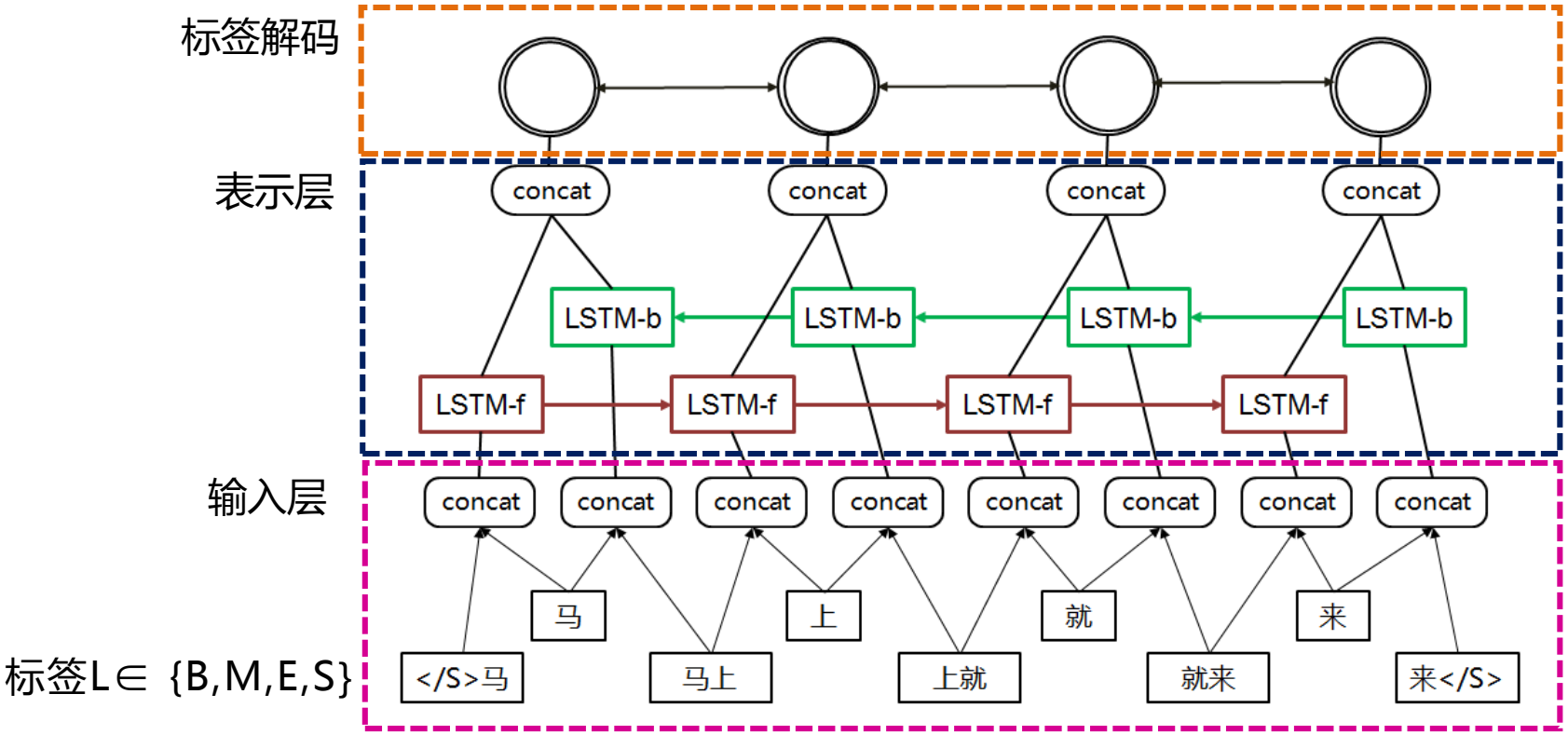
第二类：基于词的分词方法

7. Neural Word Segmentation Learning for Chinese
8 . Fast and Accurate Neural Word Segmentation for Chinese
9. Chinese Parsing Exploiting Characters
10. Transition-based: Transition-based Neural Word Segmentation
11. Segmenting Chinese Microtext: Joint Informal-Word Detection and Segmentation with Neural Networks

10.2.1 自动分词

- **BI-LSTM+CRF 中文分词** （注：为便于讲解用标准RNN代替LSTM）

模型结构



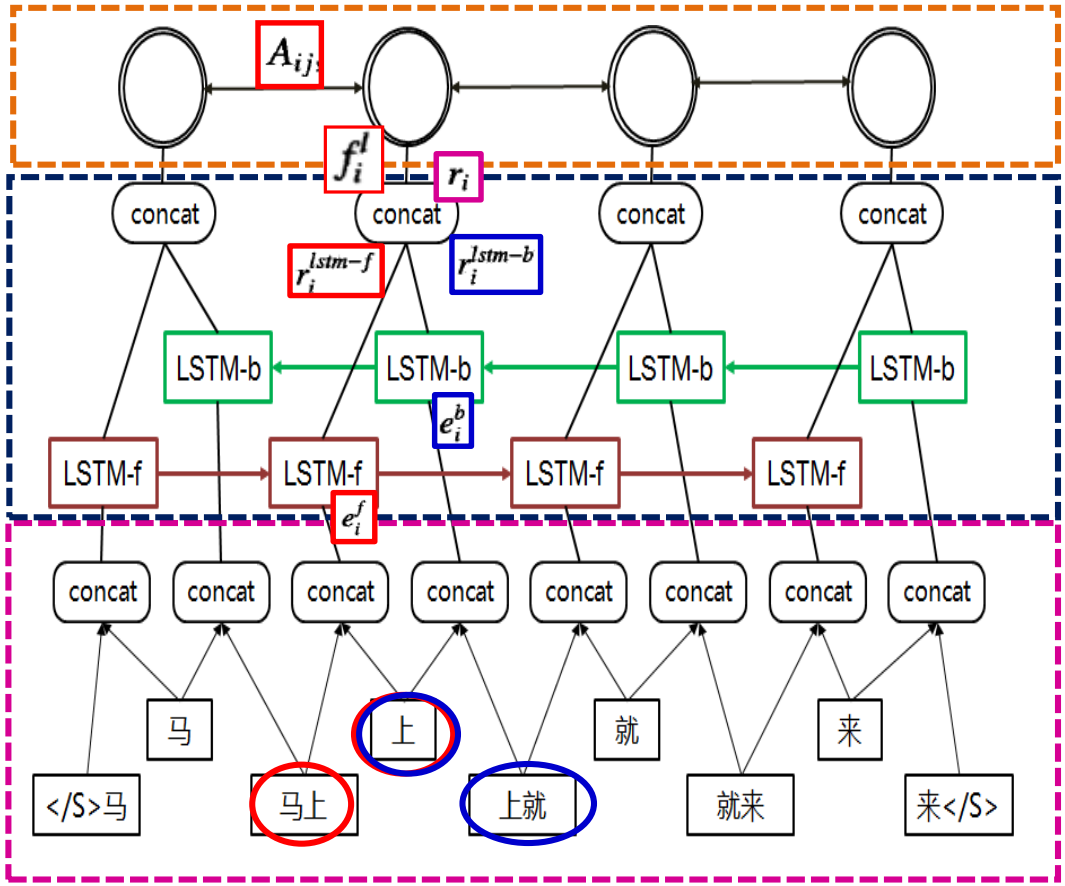
郁振庭，基于神经网络的词法分析研究，硕士论文，2017

转移得分矩阵A可以对应条件随机场中的转移特征函数，f 对应条件随机场中的状态特征函数。

注，：e_{ci}代表当前字的单字词嵌入 $e_{ci-1\ ci}$ 代表前一个字和当前字组合的双字词嵌入

n 句子的长度，θ 网络的参数

输入层　表示层　标签解码层



输入：</S> 马 上 就 来 </S>

Look up in Pre-trained Embeddings

$$s(c_1^n, l_1^n, \theta, A) = \sum_{t=1}^{N} (A_{l_{t-1}, l_t} + f(t, c_1^n, l_t, \theta))$$

$$f_i^l = W_l r_i \qquad p_i^l = \frac{e^{f_i^l}}{\sum_k e^{f_i^k}}$$

$$r_i^{lstm-f} \leftarrow e_i^f \qquad r_i^{lstm-b} \leftarrow e_i^b$$

$$r_i = concat2(r_i^{lstm-f}, r_i^{lstm-b})$$

$$concat2 = tanh(W_2[r_i^{lstm-f}; r_i^{lstm-b}] + b_2)$$

$$e_i^f = concat1(e_{c_i}, e_{c_{i-1}c_i})$$

$$concat1(e_{c_i}, e_{c_{i-1}c_i}) = tanh(W_1[e_{c_i}; e_{c_{i-1}c_i}] + b_1)$$

$$e_i^b = concat1(e_{c_i}, e_{c_i c_{i+1}})$$

$$concat1(e_{c_i}, e_{c_i c_{i+1}}) = tanh(W_1[e_{c_i}; e_{c_i c_{i+1}}] + b_1)$$

输入：句子{e_1, e_2, e_3, ..., e_n } ,其中e_i 表示句子中的第i个字，n是句子的长度

10.2.1 自动分词

模型学习

训练目标

假设当前的训练数据集为D=x_1, ...x_i, ..., x_n，对于句子 $x_i \in D$，其正确的标注序列为y i，当前模型的网络参数为θ，那么最终的优化目标可以表达为最大化下式中的对数似然函数。

$$\theta, A = \underset{\theta,A}{\operatorname{argmax}} L(\theta, A) = \underset{\theta,A}{\operatorname{argmax}} \sum_{i=1}^{|D|} \log P(y_i|x_i, \theta, A)$$

其中，

$$P(y_i|x_i, \theta, A) = \frac{\exp s(x_i, y_i, \theta, A)}{\sum_{y_i^*} \exp s(x_i, y_i^*, \theta, A)}$$

$$s(c_1^n, l_1^n, \theta, A) = \sum_{t=1}^{N} (A_{l_{t-1}, l_t} + f(t, c_1^n, l_t, \theta))$$

采用梯度下降法训练模型参数 θ，A

模型预测

$$y^* = \operatorname{argmax}\ s(c_1^n, l_1^n, \theta, A)$$

内 容 提 要

10.2.2 词性标注

词性标注需要解决的问题：

词性（part-of-speech）是词汇的基本语法属性，通常称为**词类**

任务	目标	需解决问题
自动分词	将输入汉字串切成词串	1. 歧义问题 2. 未登录词问题 3. 分词标准问题
词性标注	确定每个词的词性并加以标注	词性兼类歧义问题

规则法

概率统计

深度学习

技术方法

问题？

10.2.2 词性标注

汉语词性兼类问题： （在任何一种自然语言中，词性兼类问题都普遍存在）

1. 汉语缺乏词形态变化, 无法通过词形变化判别词类

2. 常用词兼类现象严重（兼类占 22.5%）

(1) 形同音不同，如：“好(hao3，形容词)、好(hao4，动词)”

这个人什么都**好**，就是**好**酗酒。

(2) 同形、同音，但意义毫不相干，

如：“会(会议，名词)、会(能够、动词)”

每次他都**会**在**会**上制造点新闻。

(3) 组合情况，如：“行(xing2，动词/形容词；hang2，名词/量词)”

每当他走过那**行**白杨树时，他都感觉好像每一棵树都在向他**行**注目礼

10.2.2 词性标注

3. 没有统一的汉语词类划分标准

- 有的语料用比较粗糙的标记集，例如：N, V, A, Aux, ...
- 有的语料用更细致的分类：(例如： Penn Treebank)

如： **北大计算语言学研究所的词性标注集**

26个基本词类代码，74个扩充代码，标记集中共有106个代码

名词(n)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、代词(r)、动词(v)、形容词(a)、状态词(z)、副词(d)、介词(p)、连词(c)、助词(u)、语气词(y)、叹词(e)、拟声词(o)、成语(i)、习用语(l)、简称(j)、前接成分(h)、后接成分(k)、语素(g)、非语素字(x)、标点符号(w)。

10.2.2 词性标注

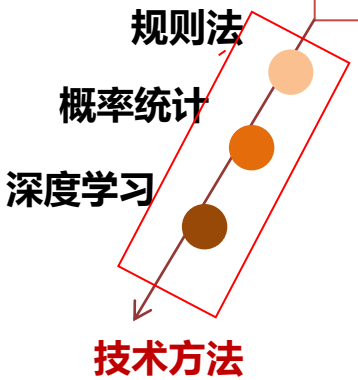
UPenn Treebank 的词性标注集（33 类）

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WPS	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PPS	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

10.2.2 词性标注

词性标注技术方法：

任务	目标	需解决问题
自动分词	将输入汉字串切成词串	1. 歧义问题 2. 未登录词问题 3. 分词标准问题
词性标注	确定每个词的词性并加以标注	词性兼类歧义问题



技术方法

技术方法？

10.2.2 词性标注

词性标注方法：

1. 基于规则的词性标注方法
2. 基于统计模型的词性标注方法
3. 规则和统计方法相结合的词性标注方法
4. 基于神经网络的词性标注方法

在统计和神经网络方法中，词性标注为序列标注问题，解决这类问题的模型有 HMM ，CRF，RNN+CRF 等。（参见第 9 章）

内 容 提 要

10.1　词法分析概述

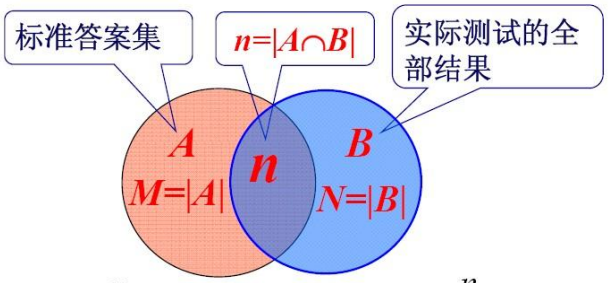
10.2　词法分析任务

10.3　词法分析评价

10.4　词法分析系统概述

10.3 词法分析评价

评价指标



$$P = \frac{n}{N} \times 100\% \qquad R = \frac{n}{M} \times 100\,\%$$

正确率 (precision, **P**) ：测试结果中正确切分的个数占系统所有输出结果的比例

$$P = \frac{n}{N} \times 100\ \%$$

召回率(Recall ratio, **R**)：测试结果中正确结果的个数占标准答案总数的比例

$$R = \frac{n}{M} \times 100\ \%$$

F-度量值：正确率和召回率的综合值

$$F - measure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \times 100\ \%$$

一般地，取 β =1，即

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\ \%$$

10.3 词法分析评价

例：假设某个汉语分词系统在一测试集上输出**5260** 个分词结果，而标准答案是**4510** 个词语，根据这个答案，系统切分出来的结果中有**4120** 个是正确的。那么：

正确率 $$P = \frac{4120}{5260} \times 100\% = 78.33\%$$

召回率 $$R = \frac{4120}{4510} \times 100\% = 91.35\%$$

F-度量值

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

$$= \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \times 100\%$$

$$= 84.34\%$$

10.3 词法分析评价

说明：

如果汉语自动分词与词性标注一体化进行，对于词性标注来说，可以用“召回率”衡量词性标注系统的性能，但是，如果不是分词与词性标注一体化进行，而是词性标注系统对已经切分好的汉语词汇进行词性标注，那么，一般不采用“召回率”指标衡量词性标注系统的性能。

内 容 提 要

10.1　词法分析概述

10.2　词法分析任务

10.3　词法分析评价

10.4　词法分析系统概述

10.4 词法分析系统概述

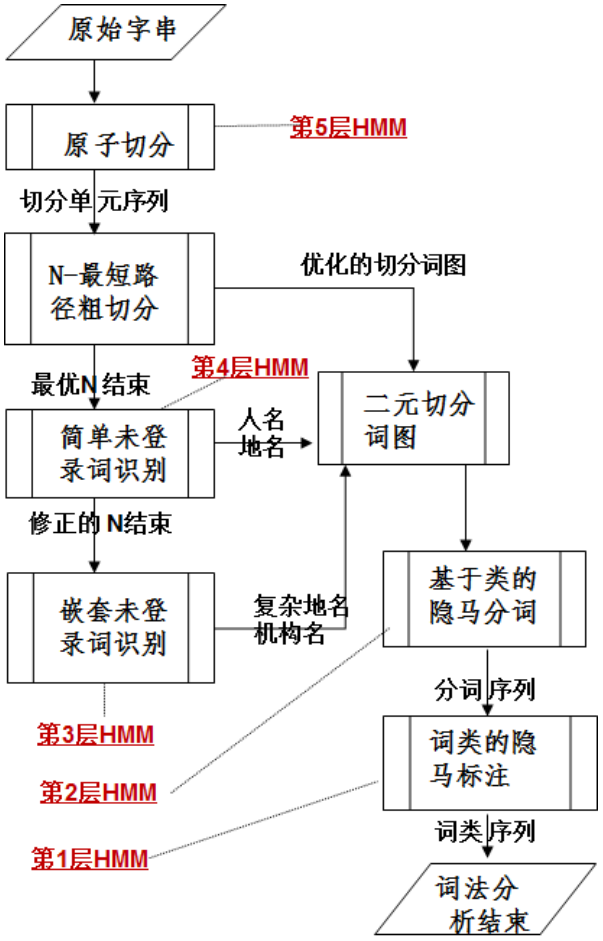
分词系统：

汉语分词系统 需将切分、未登录词处理，词性标注 集成到一起 。一个成熟的分词系统，不可能单独依靠某一种算法来实现，都需要综合不同的算法来处理不同的问题

如，实现系统时，对于未登录词识别可使用CRF 模型；对于人名识别，可使用最大熵模型；对于数字、日期等，可使用有限状态自动机；对于词性标注，可使用 HMM 模型。

10.4 词法分析系统概述

早期-计算所分析系统工程ICTCLAS （基于HMM的汉语词法分析框架）



N-最短路径粗切分：

基本思想是在初始阶段保留切分概率P(W)最大的N个结果，作为分词结果的候选集合，以最大限度地保留歧义字段和未登录词。

10.4 词法分析系统概述

常用的公开提供服务的分词系统：

	分词系统	标识
1	BosonNLP	BOSON
2	IKAnalyzer	中文分词库 IKAnalyzer
3	NLPIR	NLPIR 汉语分词系统 ICTCLAS 2015
4	SCWS	SCWS 中文分词
5	结巴分词	jieba
6	盘古分词	盘古分词-开源中文分词组件
7	庖丁解牛	paoding
8	搜狗分词	Sogou搜狗 云服务
9	腾讯文智	互联网+ 腾讯云
10	新浪云	新浪云
11	语言云	語言

各家分词系统链接地址

BosonNLP：http://bosonnlp.com/dev/center

IKAnalyzer：http://www.oschina.net/p/ikanalyzer

NLPIR：http://ictclas.nlpir.org/docs

SCWS中文分http://www.xunsearch.com/scws/docs.php

结巴分词：https://github.com/fxsjy/jieba

盘古分词：http://pangusegment.codeplex.com/

庖丁解牛：https://code.google.com/p/paoding/

搜狗分词：http://www.sogou.com/labs/webservice/

腾讯文智：

http://www.qcloud.com/wiki/API%E8%AF%B4%E6%98%8E%E6%96%87%E6%A1%A3

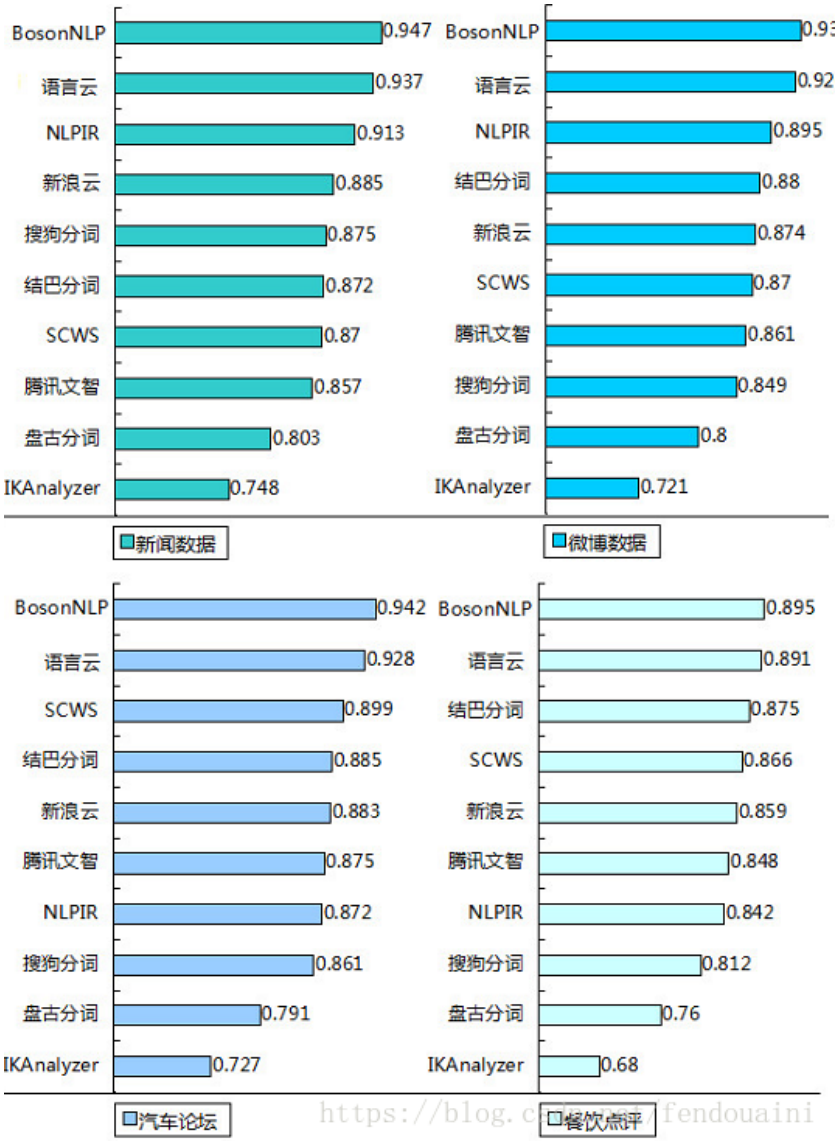
新浪云：

http://www.sinacloud.com/doc/sae/python/segment.html

语言云：http://www.ltp-cloud.com/document

10.4 词法分析系统概述

准确度对比：



10.4 词法分析系统概述

应用情况对比：

分词服务	分词粒度	出错情况	支持处理字符	新词识别	词性标注	认证方法	接口
BosonNLP	多选择	无	识别繁体字	有	有	Token	REST API
IKAnalyzer	多选择	无	兼容韩文、日文字符	有	无	无	jar包
NLPIR	多选择	中文间隔符，返回局部乱码	未知	有	有	无	多语言接口
SCWS	多选择	无	未知	有	有	无	PHP库|命令行工具
结巴分词	多选择	无	识别繁体字	有	有	无	python库
盘古分词	多选择	无	识别繁体字并自动转换	有	无	无	无
庖丁解牛	多选择	无	是	有	无	无	jar包
搜狗分词	小	放弃超过一定字符长度的句子	识别繁体字并自动转换	未知	有	无	支持上传文档，但是一直失败
腾讯文智	小	空白字符、中文间隔符，整段返回错误码	未知	有	返回中文词性	Signature	REST API
新浪云	大	无	未知	有	有	需要在新浪有一个仓库	REST API
语言云	适中	无	识别繁体字	有	有	Token	REST API

10.4 词法分析系统概述

分词系统存在的主要问题

- **语料问题**

基于机器学习模型的分词系统，依赖训练语料，标注语料的词语使用无法覆盖实际语言现象，因此基于标注语料训练的分词系统在差异较大的领域会出现准确率降低的情况；还有很多分词系统没有对训练数据进行一致性校验，在实际情况中训练数据包含了不少标注不一致的情况。主要有：切分不一致；词性标注不一致；各分词语料之间在词语颗粒度上有一定差异。如：承租人、承租者 (北大) | 承租 商 (微软)，高 清晰度 彩电 (北大) | 高清晰度电视 (微软)。往往导致系统在不同的测试集上差异较大。

从已有工程经验来看，几乎不存在通用而且效果非常好的分词系统，

10.4 词法分析系统概述

- **理论方案问题**

采用什么样的模型？常见的有隐马尔科夫模型、最大熵模型、条件随机场模型、结构感知机模型、RNN 模型等。

- 最大熵模型、条件随机场模型等模型特征工程会对最终分词结果产生很大影响。

采用基于字还是基于词的方案？从实践来看，基于字的模型对未登录词识别能力较强，但基于词的模型很少会出现切分“离谱”的情况。采用什么颗粒度单元，取决于具体任务。

10.4 词法分析系统概述

- **工程问题**

- 针对有规律的部分，可以利用规则或者正则表达式来识别，例如数字、标点、时间、日期、重叠式等，如笑一笑。
- 增加词表：是提高切分准确率“立竿见影”的办法。在自然语言处理中，只要是封闭集合的词语或实体，可以考虑利用词表来切分，例如成语、地名、公交站名、路名等。该方法简单有效。
- 扩大训练语料：的一种办法是购买更多语料；另外一种办法是利用其它分词系统来切分数据，对数据进行清洗，形成新数据。

经验：神经网络在 NLP 上的成功应用的领域往往是准确率不高或者运行效率很低的场合，例如问答系统、机器翻译、句法分析。在准确率比较高或者运行效率不错的场景下，利用深度学习会得不偿失。

参考文献：

宗成庆，统计自然语言处理（第2版）课件

https://blog.csdn.net/a101330107/article/details/78638146?utm_source=blogxgwz4

https://blog.csdn.net/chivalrousli/article/details/41987081?utm_source=blogxgwz0

https://www.cnblogs.com/croso/p/5349517.html

http://wenku.baidu.com/view/a77ec0e4f8c75fbfc77db268.html?from=search

在此表示感谢！

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

谢谢各位！