考试安排

◆ 考试科目

自然语言处理

◆ 考试时间

拟定2022年12月19日(17周周一)08:00-10:00

◆ 考试地点

线上

◆ 考试形式

待定

第七章 词法分析与词性标注

7.1 概述

7.1 概述

词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。

自动词法分析就是利用计算机对自然语言的形态(morphology) 进行分析，判断词的结构和类别等。

词性或称词类(Part-of-Speech，POS)是词汇最重要的特性，是连接词汇到句法的桥梁。

7.2 分词与词性标注结果评价

7.2 分词与词性标注结果评价

- 两种测试：
  - 封闭测试 / 开放测试
  - 专项测试 / 总体测试

7.2 分词与词性标注结果评价

◆ 评价指标

正确率(Correct ratio/Precision, P)：

测试结果中正确切分或标注的个数占系统所有输出结果的比例。假设系统输出 N 个，其中，正确的结果为 n 个，那么

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

7.2 分词与词性标注结果评价

◆ 评价指标

召回率(找回率) (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。假设系统输出 N 个结果，其中，正确的结果为 n 个，而标准答案的个数为 M 个，那么，

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

两种标记： R_{OOV} 指集外词的召回率；

R_{IV} 指集内词的召回率。

7.2 分词与词性标注结果评价

◆ 评价指标

F-测度值(F-Measure)：

测试正确率与找回率的综合值。计算公式为：

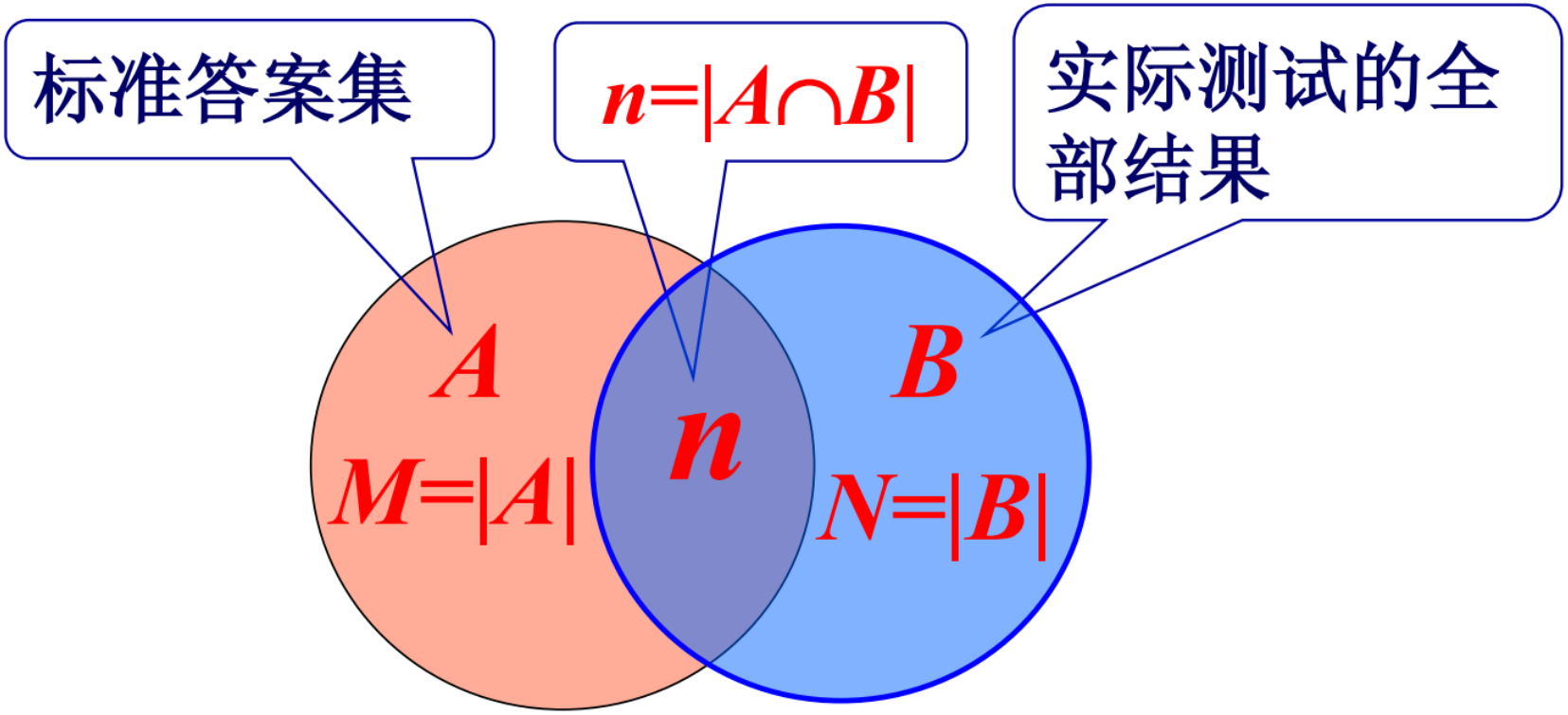
$$PF-measure=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\times 100\%$$

一般地，取 $\beta=1$，即

$$F_1=\frac{2\times P\times R}{P+R}\times 100\%$$

7.2 分词与词性标注结果评价

◆ 评价指标



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

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

7.2 分词与词性标注结果评价

◆ 评价指标

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：

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

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

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% = \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \times 100\% = 84.34\%$$

7.3 汉语自动分词基本算法

7.3 汉语自动分词基本算法

- 有词典切分/ 无词典切分
- 基于规则的方法/ 基于统计的方法

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

——有词典切分，机械切分

- 正向最大匹配算法 (Forward MM, FMM)
- 逆向最大匹配算法 (Backward MM, BMM)
- 双向最大匹配算法 (Bi-directional MM)

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

- 正向最大匹配算法(FMM)

基本思想：从左向右取待切分句子的 m (m 为词典中最长词的字数)个字符作为匹配字段，查找词典并进行匹配。若匹配成功，则将这个匹配字段作为一个词切分出来；若不成功，则去掉这个匹配字段的最后一个字，剩下的字符串作为新的匹配字段，进行再次匹配。重复以上过程，直到切分出所有词为止。

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

- 正向最大匹配算法描述：

假设有一个句子为：$S = c_1 c_2 \cdots c_n$ ，某一个词为：$W = c_1 c_2 \cdots c_m$ ，m 为词典中最长词的字数。

① 令 $i = 0$，当前指针 p_i 指向输入字串的初始位置，执行下面的操作：

② 计算当前指针 p_i 到字串末端的字数(即未被切分字串的长度) n，如果$n = 1$，转(4)，结束算法。否则，令 m =词典中最长单词的字数，如果$n < m$，令 $m = n$；

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

- 正向最大匹配算法描述：

③ 有从当前 p_i 起取 m 个汉字作为词 w_i，判断：

a) 如果 w_i 是词典中的词，则在 w_i 后添加一个切分标志，转(c)；

b) 如果 w_i 不是词典中的词且 w_i 的长度大于1，将 w_i 从右端去掉一个字，转(a)步；否则(w_i 的长度等于1)，则在 w_i 后添加一个切分标志，将 w_i 作为单字词添加到词典中，执行(c)步；

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

- 正向最大匹配算法描述：

c) 根据 w_i 的长度修改指针 p_i 的位置，如果 p_i 指向字串末端，转④，否则，$i = i + 1$，返回 ②；

④ 输出切分结果，结束分词程序。

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

例：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。

切分过程：他是研究生物化学的。

他/ 是研究生物化学的。

FMM切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/。

BMM切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/。

7.3 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

- 优点:
  - 程序简单易行，开发周期短;
  - 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源;
- 弱点:
  - 歧义消解的能力差;
  - 切分正确率不高，一般在95%左右。

7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

基本思想：

给定一待处理字串，根据词典，找出词典中所有可能的词，构造出字串的一个有向无环图，算出从开始到结束所有路径中最短的前N条路径。

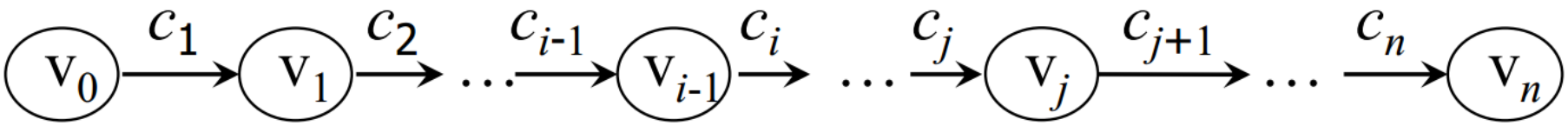
因为允许相等长度的路径并列，故最终的结果集合会大于或等于N。

7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

基本思想：

设待切分字串 $S = c_1 c_2 \cdots c_n$，其中 $c_i (i = 1,2,\cdots,n)$ 为单个的字，$n(n \geq 1)$ 为串的长度。建立一个节点数为 $n+1$ 的切分有向无环图 G，各节点编号依次为 $V_1, V_2, \cdots, V_n$。



求最短路径：贪心法或简单扩展法。

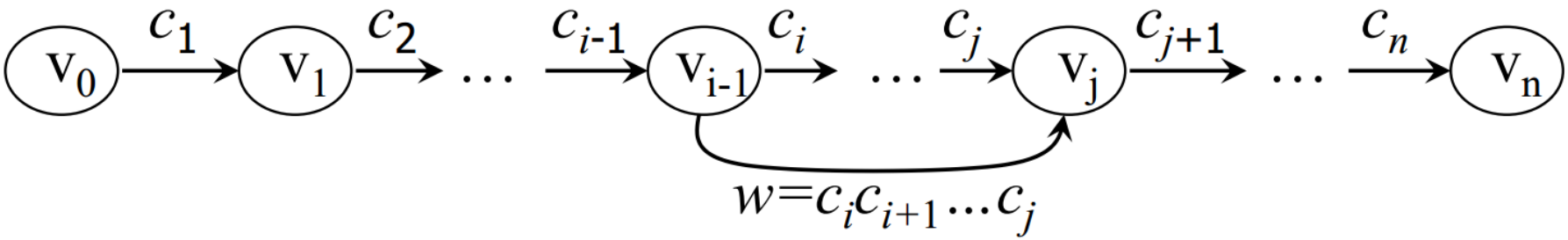
7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

算法描述：

① 相邻节点 v_{k-1}，v_k 之间建立有向边 $< v_{k-1}, v_k >$，边对应的词默认为 $c_k(k = 1,2,\cdots,n)$。

② 如果 $w = c_i c_{i+1} \cdots c_j (0 < i < j \leq n)$ 是一个词，则节点v_{i-1}，v_j 之间建立有向边 $< v_{i-1}, v_j >$，边对应的词为 w。



7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

算法描述：

③ 重复步骤②，直到没有新路径(词序列)产生。

④ 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。

7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

例： (1) 输入字串：他只会诊断一般的疾病。

可能输出：他/ 只会/ 诊断/ 一般/ 的/ 疾病/。 (7)

他/ 只/ 会诊/ 断/ 一般/ 的/ 疾病/。 (8)

… …

最终结果：他/ 只会/ 诊断/ 一般/ 的/ 疾病/ 。

(2) 输入字串：他说的确实在理。

可能输出：他/ 说/ 的/ 确实/ 在理/ 。 (6)

他/ 说/ 的确/ 实在/ 理/ 。 (6)

… …

7.3 汉语自动分词基本算法

2. N-最短路径法 (最少分词法)

- 优点:
  - 切分原则符合汉语自身规律;
  - 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。

7.3 汉语自动分词基本算法

3. 基于语言模型的分词方法

方法描述：

设对于待切分的句子 S，$W = w_1 w_2 \cdots w_k (1 \leq k \leq n)$ 是一种可能的切分。

$$W^* = \underset{W}{\operatorname{argmax}}\, p(W|S)$$

$$= \underset{W}{\operatorname{argmax}}\, p(W) \times p(S|W)$$

详见第5章举例。

语言模型

生成模型

7.3 汉语自动分词基本算法

3. 基于语言模型的分词方法

- 优点：
  - 减少了很多手工标注的工作；
  - 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率。
- 弱点：
  - 训练语料的规模和覆盖领域不好把握；
  - 计算量较大。

7.3 汉语自动分词基本算法

4. 基于HMM的分词方法

基本思想：

输入字串(句子) S 作为 HMM μ 的输入；切分后的单词串 S_w 作为HMM状态的输出，即观察序列 $S_w = w_1 w_2 \cdots w_n \ (n \geq 1)$。词性序列 S_c 为状态序列，每个词性标记 c_i 对应 HMM 中的一个状态 q_i，$S_c = c_1 c_2 \cdots c_n$。

$$\hat{S}_w = \underset{S_w}{\operatorname{argmax}}\, p(S_w|\mu)$$

详见第6章举例。

7.3 汉语自动分词基本算法

4. 基于HMM的分词方法

- 优点：
  - 可以减少很多手工标注的工作量；
  - 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率。
- 弱点：
  - 训练语料的规模和覆盖领域不好把握；
  - 模型实现复杂、计算量较大。

7.3 汉语自动分词基本算法

5. 由字构词(基于字标注)的分词方法 (Character-based tagging)

第一篇由字构词的汉语分词方法的论文[Xue，2002]发表在2002年第一届国际计算语言学学会(ACL)汉语特别兴趣小组SIGHAN（http://www.sighan.org/）组织的汉语分词评测(Bakeoff)研讨会上。该方法在2005年和2006年的两次 Bakeoff 评测中取得好成绩。

7.3 汉语自动分词基本算法

5. 由字构词(基于字标注)的分词方法

基本思想：将分词过程看作是字的分类问题。该方法认为，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

这里的“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汉语文本中的文字符号，所有这些字符都是由字构词的基本单元。

7.3 汉语自动分词基本算法

5. 由字构词(基于字标注)的分词方法

例如:

(1) 上海/ 计划/ 到/ 本/ 世纪/ 末/ 实现/ 人均/ 国内/ 生产/ 总值/ 五千美元/ 。/

(2) 上/B 海/E 计/B 划/E 到/S 本/S 世/B 纪/E 末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总/B 值/E 五/B 千/M 美/M 元/E 。/S

7.3 汉语自动分词基本算法

5. 由字构词(基于字标注)的分词方法

在字标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型，然后在待切分字串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果，最后根据词位定义直接获得最终的分词结果。

- 工具：
  - 支持向量机 (SVM)
  - 条件随机场 (CRF)

最常用的两类特征是：
- 字本身
- 词位(状态)的转移概率

7.3 汉语自动分词基本算法

5. 由字构词(基于字标注)的分词方法

评价：

该方法的重要优势在于，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的。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[黄昌宁，2006]

7.3 汉语自动分词基本算法

6. 其他分词方法

- 全切分方法
- 串频统计和词形匹配相结合的分词方法
- 规则方法与统计方法相结合
- 多重扫描法

……

推荐 Urheen 汉语自动分词系统:
http://www.nlpr.ia.ac.cn/cip/software.htm

7.3 汉语自动分词基本算法

◆方法比较

(1) 最大匹配分词算法是一种简单的基于词表的分词方法，有着非常广泛的应用。这种方法只需要最少的语言资源（仅需要一个词表，不需要任何词法、句法、语义知识），程序实现简单，开发周期短，是一个简单实用的方法，但对歧义字段的处理能力不够强大。

7.3 汉语自动分词基本算法

◆方法比较

(2) 最短路径分词方法的切分原则是使切分出来的词数最少。这种切分原则多数情况下符合汉语的语言规律，但无法处理例外的情况，而且如果最短路径不止一条时，系统往往不能确定最优解。

(3) 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。

谢谢!