

汉语词性标注中兼类词排歧算法探讨*

王素格 张永奎 刘开瑛

(山西大学计算机科学系, 山西太原 030006)

摘要: 本文对词性标注的几种算法: CLAWS 算法、VOLSUNGA 算法、遗传算法做了比较, 分析各自的时间复杂度, 并针对文本中的每个 SPAN 用遗传算法对其兼类词进行词性标注, 实验结果证明此方法是可行的。

关键词: 词性标注 同现概率矩阵 语料库 遗传算法

The Discussion about Ambiguous Algorithm to the Chinese Part-of-Speech Tagging

Wang Suge Zhang Yongkui Liu Kaiying

Department of Computer Science, Shanxi University, Taiyuan

Abstract: In this paper, some part-of-speech tagging algorithms, including CLAWS algorithm, VOLSUNGA algorithm and genetic algorithm, are compared. The time complexity of each algorithm is analyzed respectively. Then the genetic algorithm is used to tag the ambiguous words in every SPAN, and the experiment result is feasible.

keywords: part-of-speech tagging, co-concurrency frequency matrix, corpus, genetic algorithm

一、引言

词性标注是对文本中的每个词自动赋予一种确定的词性。随着计算机对大量真实文本处理的迫切需要, 对词性的要求也显得日益迫切, 由于它的研究结果直接影响到句法分析、语义分析、语音分析、机器翻译、信息检索等诸多研究, 因此, 一直引起人们的关注。词性自动标注的难点是兼类词的排歧。由于汉语缺乏词的形态变化, 常用兼类词占比重大, 致使兼类词在句中的频较高, 成为词性标注的主要困难。80 年代 Mashall 提出的 LOB 语料库标注算法 CLAWS, 首先, 将概率统计模型用于词类的自动标注, 正确率据称已达 97%, 但算法的时间复杂度为指数级。之后, DeRose 等又在 CLAWS 的基础上进行了改进, 提出 VOLSUNGA 算法, 利用了语料库中词与词的统计信息, 据称正确率达到 96%。本文试图使用一种基于进化过程的信息遗传机制和优胜劣汰的自然选择原则的搜索算法(以字符串表示状态空间), 遗传算法将对含有兼类词的一个 SPAN 进行词性标注, 目的是排除歧义, 通过实验证明此方法是可行的。

* 国家自然科学基金资助项目 69575011

二、统计模型

2.1 n 元模型

对于给定的一个词的序列 W ，推断最可能的标记序列 T ，即对每个标记序列 T 的后验概率

$$P(T/W) = \frac{P(T)P(W/T)}{P(W)}$$

其中 $W=w_1, w_2, \dots, w_n$ w_i 为第 i 个词

$T_s=t_{s,1}, t_{s,2}, \dots, t_{s,n}$ 是 W 的最可能的词性标记序列， $t_{s,i}$ 为第 i 个词的可能词性。

$T_s^*=t_{s,1}^*, t_{s,2}^*, \dots, t_{s,n}^*$ 是 W 的最佳词性标记序列， $t_{s,i}^*$ 为第 i 个词的最终词性。

$$\text{即 } T_s^* = \arg \max_{T_s} P(T_s)P(W/T_s)$$

$$\begin{aligned} &= \arg \max_{T_s} P(t_{s,1})P(w_1/t_{s,1}) \prod_{i=2}^n P(t_{s,i}/t_{s,1}, \dots, t_{s,i-1}) \\ &\quad \cdot P(w_i/t_{s,1}, w_1, t_{s,2}, w_2, \dots, t_{s,i-1}, w_{i-1}, t_{s,i}) \end{aligned}$$

2.2 二元模型

由于 n 元模型考虑了所有的上、下文因素，从理论上讲，基于 n 元模型的词性标注系统的标注正确率高，但实现起来时间复杂度太大，而据语言学知识，相距较远的词性联系不大，可以近似地不予考虑，因此，对于 n 元模型可以假设每个标记只与前面相邻的标记有关，这样， n 元模型可以变为二元模型

$$T_s^* = \arg \max_{T_s} P(t_{s,1})P(w_1/t_{s,1}) \prod_{i=2}^n P(t_{s,i}/t_{s,i-1})P(w_i/t_{s,i})$$

2.3 同现概率矩阵

$$P=(p_{ij})_{n \times n} \quad \text{其中 } p_{ij} = \frac{\text{标记 } i \text{ 与标记 } j \text{ 同现的次数}}{\text{标记 } i \text{ 出现的次数}} \times 100\%$$

三、CLAWS 算法

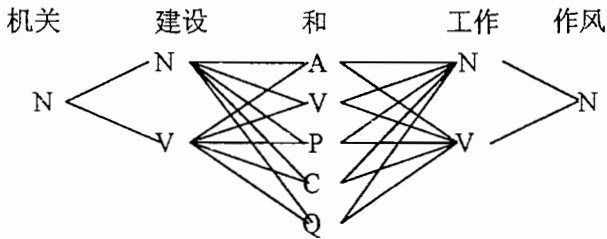
(1) 将测试语料首先分词。

(2) 查兼类词表、非兼类词表（假定词类共 26 种，在此仅考虑大类）。

(3) 选取 SPAN。(句中的 N 个相邻的兼类词极其前后的非兼类词构成一个 SPAN ,SPAN 中兼类词的个数叫做 SPAN 的长度，每一中词类标记叫做一条路径 SPAN)。

(4) 取路径上排列元素之间的同现概率的联合分布率最大的那条路径。

例如：从网上下载的北京青年报中的一个 SPAN 为

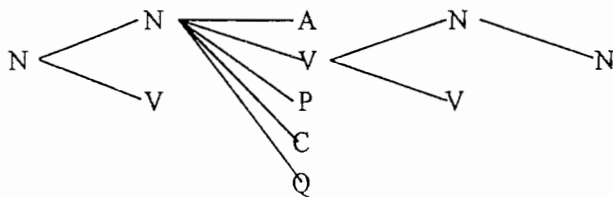


共有 $1 \times 2 \times 5 \times 2 \times 1 = 20$ 条路径，选取极大似然估计值最大 0.0028

四、VOLSUNGA 算法

对任一长度为 n 的 SPAN，从左到右， $W = w_0, w_1, w_2, \dots, w_n, w_{n+1}$ 对当前考虑的目标词设为 w_i ，只保留通往该词 w_i 的每个词的词类的最佳路径，然后，继续将这些路径与下一个词 w_{i+1} 的所有词类标记进行匹配，分别找出通往这个词 w_{i+1} 的每一个词的最佳路径，以下重复。

例如： 机关 建设 和 工作 作风



其中： $P(N,N)=0.2581$ $P(N,V)=0.1445$ $P(N,A)=0.0313$ $P(N,C)=0.0038$

$P(N,P)=0.0251$ $P(N,Q)=0.0006$ $P(V,N)=0.2916$ $P(V,V)=0.1639$

利用 VILSUNGA 算法进行标注，选择的最佳路径是：

机关/N/建设/N/和/V/工作/N/作风/N

从标注的结果可以看出：“和”的词性在此标为“动词 V”，而实际“和”的词性在此应为“连词 C”，出错的原因是由于 CLAWS 算法、VOLSUNGA 算法仅仅考虑词性与词性的同现概率，并没有考虑词本身的词义，以及上下文词与词之间的相关信息。本例中

$P(N, V) * P(V, N) > P(N, C) * P(C, N)$ ，因此，“和”的词性在此标为“动词 V”。为此本文将在一个 SPAN 上利用遗传算法，对兼类词进行词性标注。

五、遗传算法

词性自动标注也是一种优化问题，目的是寻找最佳路径，但以上提到的 CLAWS 算法、

VILSNGA 算法在开放集上的准确率不能令人满意，比如：利用 VOLSUNGA 算法标注下列 SPAN 为：

(1) 金融体制 N 改革 N 和 V 防范 V

正确的结果为：金融体制 N 改革 V 和 C 防范 V

(2) 继续 N 改善 N 和 V 发展 N 同 C 发达 V 国家 N 的 U 关系 N

正确的结果为：继续 D 改善 V 和 C 发展 V 同 P 发达 A 国家 N 的 U 关系 N

本文对文本中的含有兼类词的 SPAN，利用遗传算法进行词性的自动标注。

5.1 遗传算法的思想：

遗传算法是 70 年代美国的 Michigan 大学的 John.Holland 提出的，它的思想是从达尔文的自然选择，蒙德尔的遗传变异，根据生物的进化过程：繁殖、变异、竞争、选择。

遗传算法：是一类模拟生物进化过程和机制，来求问题自适应的人工智能技术。

适应度：在遗传算法中，衡量个体优劣的尺度是适应度的大小，决定某些个体是繁殖或消失。因此，适应度是驱动遗传算法的动力。从生物学角度讲，适应度相当于“物竞天择，适者生存”的生物能力。

5.2 算法的步骤：

(1) 编码：利用字符编码。通过一个随机函数 rand()，在每个词可能的词性中随机产生一组字符，即对于染色体 $chrom[i] \in \{A, B, \dots, Z\}$ ，且 $1 \leq i \leq SPAN$ 的长度，其中 26 个英文字母为词所有的词性，此处仅考虑词的大类。

(2) 初始化：当通过随机函数产生几组编码即初始的几个个体，这样 SPAN 中的每个词都标了可能的词性，即对给定的 SPAN $W = w_1, w_2, \dots, w_{n-1}, w_n$ 有

$$\begin{array}{cccc} w_1 & w_2 & \dots & w_{n-1} & w_n \\ t_{1,1} & t_{1,2} & \dots & t_{1,n-1} & t_{1,n} \\ \dots & \dots & \dots & \dots & \dots \\ t_{m,1} & t_{m,2} & \dots & t_{m,n-1} & t_{m,n} \end{array}$$

其中： $t_{i,j} \in \{t_{1,j}, t_{2,j}, \dots, t_{m,j}\}$ ，据兼类词表第 j 个词的有 m_j 种词性， $t_{i,j}$ 为第 j 个词的第 i 种可能词性，m 为群体的规模，n 为 SPAN 的长度。

(3) 适应度函数：在词性标注中，我们定义适应度函数为：

$$fitness(T_S) = P(t_{s,1})P(w_1/t_{s,1}) \prod_{i=2}^n P(t_{s,i}/t_{s,i-1})P(w_i/t_{s,i})$$

(4) 复制：先利用轮盘选择法从旧群体中选择进行繁殖的个体，通常是适应度最大的个体，用它代替适应度最小的个体，以便产生一个新的群体。

(5) 杂交：通过复制产生的新群体，其性能得到改善，但不能产生新的个体。为了产生新的个体，利用生物学中杂交的原理达到这一目的。据词性标注的编码为字符序列，因此，采用离散杂交中的生成模板形式。

(6) 突变: 由于突变概率很小, 约为 0.008。我们还是继续随机地决定在染色体某个基因上突变, 随机产生某个词的可能词性。

反复执行(3)~(6), 利用复制、杂交、突变三种操作不断更换标注策略, 迭代的次数越多准确率越大。这样对每个 SPAN 进行标注, 直到把所要标注文本的所有词的词性都标出来。

六、算法分析与讨论

若 SPAN 的长度为 n , 词性的种类为 m , CLAWS 运行的时间效率极低, 它的运行时间与 SPAN 的歧义级别的乘积成正比, 最坏的情况的时间复杂度为 $O(m^n)$, 因此, 对 SPAN 的长度来说, 时间复杂度是以指数增长。而 VOLSUNGA 算法, 它的时间复杂度为 $O(n \cdot m^2)$ 。在遗传算法中, 利用启发式对每个词进行搜索最佳的词性, 通过一代一代地进化, 最终达到最优解, 它的时间复杂度将 CLAWS 算法的指数阶 $O(m^n)$ 降到 $O(f(n) \cdot l \cdot g)$, 其中 l 为群体的规模, g 为遗传的代次。

词性标注是自然语言处理的基础课题。目前, 主要方法利用统计、规则方法, 以及统计与规则相结合的方法。本文对含有兼类词 SPAN 进行词性标注, 通过实验, 对 70, 904 个字的语料库中, 有 5, 271 个句子, 20, 708 个兼类词进行测试, 利用 VOLSUNGA 进行测试, 正确率为 82.375%, 若用遗传算法, 正确率能达 90.5%, 原因是在适应度函数中, 除使用了同现概率外还使用了相对概率, 但用遗传算法比用 VOLSUNGA 的时间效率低。另外句中的词较少时, 遗传的代次少时, 准确率高; 句中的词较多时, 遗传的代次多时, 准确率高。但 SPAN 的长度小于 7 的占到 SPAN 总数的 99% 以上, 因此利用遗传算法对含有兼类词 SPAN 进行词性标注是可行的。我们也将利用规则与遗传算法相结合的方法来实现词性标注, 有关规则正在进行从大量的语料库中自动获取, 获取的方法见[4], 下一步的工作将对适应度函数加入规则, 以减少遗传的代次。

参考文献

- [1]. 刘开瑛、郑家恒、赵军. 语料库词类自动标注算法研究, 《机器翻译研究进展》, 电子工业出版社, 1992. 8. 378—385.
- [2]. 云庆夏、黄光球、王占权, 《遗传算法和遗传规划—一种搜索寻优技术》, 冶金工业出版社, 1997 年.
- [3]. 潘正君、唐立山、陈毓屏, 《演化计算》 清华大学出版社、广西科学技术出版社, 1998 年.
- [4]. 王素格、苗夺谦、刘开瑛, 基于 Rough Set 自动获取词性标注规则初探, 机器翻译与计算机语言信息处理国际学术研讨会 (ISMT-CLIP) 1999 年, 北京.
- [5]. Qing Ma, Hitoshi Isahara, Sun Maoson A Multi-Neuro Tagger Applied In Chinese Texts, Proceeding 1998 International Conference on Chinese Information Processing, November 18-20, 1998, Beijing, China. 200-207.