词性标注规则的获取和优化*

陈文亮 朱靖波 吕学强 姚天顺

(中文信息处理实验室,东北大学计算机软件与理论研究所,辽宁 沈阳 110004)

Website:www.nlplab.com E_mail:chenwl_cn@263.net

摘要：本文提出一种词性标注规则自动学习算法。通过对规则进行评价、优化，有效提高标注正确率和标注效率。系统对 PFR 标注语料库（98 年 1 月）进行标注，相对于 NA 假设的词性兼类消歧模型标注结果，封闭测试正确率提高了 5.53%，开放测试提高了 4.57%。

关键词 词性标注；规则自动学习；中文信息处理；

1. 前言

汉语词性标注一直是中文信息处理领域的一个重要研究课题。它是中文信息处理其他领域应用的前提。目前，词性标注方法主要分为两种：基于规则的方法[1,2]、基于统计[3,4]的方法。基于规则的方法通常采用手工编制复杂的词性标注规则系统，可以充分利用人的语言知识，但是带有很强的主观性并且存在知识获取的瓶颈问题。基于统计的方法主要利用相邻词性标记之间的同现概率及隐 Markov 语言模型，来实现词性标注，获取的知识客观性好。

基于转换的错误驱动的方法[5,6,7]是 Eric Brill 提出，用于英文的词性标注。其基本思想是利用一个初始标注器来标注训练语料库，然后把标注结果和正确标注结果进行比较，遍历所有可能的变换模式，从中选出效果最好的一条变换式，作为系统的标注规则，再用该规则重新标注语料库，重复上述过程，每次循环都得到新的一条规则，直到没有新的变换式出现。这样就可以获得词性标注的规则集。在标注时，首先使用初始标注器进行标注，然后利用获得的规则集进行标注。

本文在 Brill 学习算法基础上做了一些改进。在自动学习过程中，获取所有可能的变换模式后，对这些变换模式进行评价，选出所有效果较好的变换模式，生成自动获取的规则集。初始标注器是采用基于 NA 假设的词性兼类消歧模型（Bi-Gram）词性标注系统[8]。该系统的二元语法模型参数统计是基于无任何标注的生语料，标注正确率达到 92%左右。本文利用自动学习的规则来改善该系统，实验结果表明封闭测试标注正确率提高了 5.53%，开放测试提高了 4.57%。

2. 基本定义

1、本文中 **Lex** 表示汉语词条，**Pos** 表示词条词性。

2、规则体系：本规则体系是根据各个词条来定义规则的，包含多个规则块，规则块是以词条加词性为入口，规则块内包含一条或多条规则。规则体系表示如图 1 所示。

图 1 中 Rs 表示整个规则集，Rb 表示规则块，Ri 表示规则条。

* 获得国家 863 计划项目资助（863-301-7-7-B）

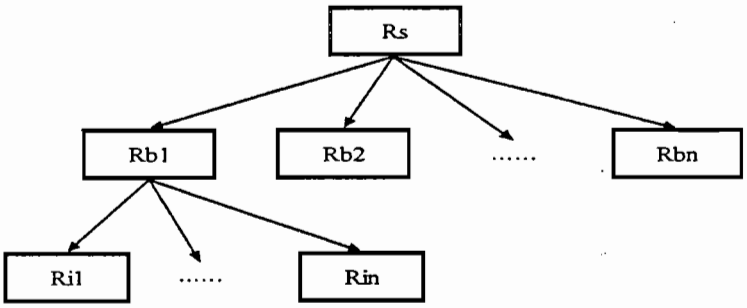


图 1 规则体系

规则体系文法 BNF 形式：

规则集 -> 规则块；{规则块}
规则块 -> 块名：规则
规则 -> 规则条；{规则条}
规则条 -> 测试条件->动作
动作 -> 设置 Pos 为某种词性
测试条件 -> 项测试条件；{项测试条件}

下面是一个规则块的例子：

lex=广播&pos=n
1^pos=n;^1pos=n;->pos=v
1^pos=p;^1pos=n;->pos=v
1^pos=p;^1pos=b;->pos=v
1^pos=c;^1pos=n;->pos=v

表示当前词条是广播并且词性是 n 有 4 条规则，第一条规则的含义：左边第一个词条词性是 n; 右边第一个词条词性是 n;->设置当前词条词性为 v。

通过这一规则体系，可以有效对词汇的特殊情况进行描述。

3、规则有效值：在实际标注中每一条规则是否有效的度量指标。计算公式：$C=f_c/f_a$，其中 f_c 是指在标注时规则标对次数，f_a 表示规则发生作用次数。

4、规则跨度 **Len**：表示变换式的测试条件，是当前词条左边 Len 个词条加上右边的 Len 个词条，可以表示为：左边 Len 个词条+加上右边 Len 个词条。在上述规则例子中，Len=1。

5、规则出现频度 **f**：某一条规则可以从 f 个错误现象推出来。

3. 规则的自动学习

在本文中，规则的学习过程如下：首先利用初始标注器来标注训练语料库；然后把标注结果和正确标注结果进行比较，得到所有可能的变换模式；对这些变换模式进行评价，计算每一条规则的有效值，从中选择所有有效值大于最低有效阀值的变换式，作为系统的标注规则集，再利用这个规则集重新标注语料库，重复进行规则的学习。这样就可以获得词性标注的规则集。

学习算法如下：

1、 对语料进行初始标注；
2、 设 Len＝1；
3、 寻找所有的变换模式（规则）；
4、 进行规则评价，计算规则有效值；
5、 进行规则优化；

6、 利用获取的规则集进行标注；

7、 Len 增加 1，如果 Len 小于设定最大规则跨度 MaxLen，转到步骤 3；

规则评价过程：用规则 r_i 标注语料，可以得到规则标对的次数 f_c 和规则作用的次数 f_a，利用公式：$C=f_c/f_a$ 计算 r_i 的有效值。

4. 规则的优化

规则优化主要目的是对自动学习规则进行优化，来提高标注正确率。在给定的训练集 T，自动学习规则得到规则集 R_a，通过筛选和优化得到最终规则 R_r。

优化算法如下：

输入：基于转换错误驱动自动学习规则，即规则集 R_a

1、 进行规则的初选；

2、 解决规则冲突；

3、 分析规则有效性，使用规则有效值来选取规则；

4、 规则合并；

结果：得到最终规则集 R_r。

4.1 规则的初选

规则的初选是去掉出现频度太少的规则。设定最少应该出现频度 f_m，规则集 R 中某条规则 r_i 出现频度 f_i，如果 $f_i < f_m$，则 r_i 属于低频规则，r_i 将从规则集 R 中被删除；否则 r_i 属于高频规则，r_i 将保留在规则集 R 中。

4.2 规则的冲突解决

规则冲突指的是在同一个规则块中，相同的条件下，所标注的结果不同。在规则集 R 中某规则块 Rb 有这样的两条规则 r_i 和 r_j，它们可以表示为 r_i ：$Test_i$->$Setting_i$ 和 r_j：$Test_j$->$Setting_j$，如果 $Test_i = Test_j$，而 $Setting_i != Setting_j$，则存在冲突，r_i 和 r_j 这两条规则都从规则集 R 中删除。如果规则集中存在规则冲突，规则解释器就无法根据测试条件，来选择合适的动作。

4.3 规则有效性的分析

规则有效性分析：在规则集 R 中，某一条规则 r_i 的有效值是 C_i，和最佳有效阀值 C_m 进行比较，如果 $C_i < C_m$，则该规则将从规则集 R 中被删除。其中 C_m 是最佳有效阀值。下面是 C_m 的寻找算法：

输入：规则集 R

1、计算每一条规则的有效值 $C=f_c/f_a$ ；

2、设置 C_m 值为 1；

3、从 R 中选取有效值大于 C_m 的规则，组成一个临时规则 R_t；

4、使用 R_t 来标注语料；

5、计算标注正确率；

6、C_m 减去 0.05，如果 C_m 大于 0，转向步骤 3；

输出：不同 C_m 和标注正确率。

根据得到结果，选取标注正确率最大时的 C_m 值，就是最佳有效阀值。

4.4 规则的合并

选取规则后，所面临的问题就是重多的规则带来的标注效率下降，为了提高标注效率，所以要进行规则合并。规则的合并是在同一规则块内进行的。在规则块 Rb 中有规则 ri ： testi->settingi 和

规则 rj：testj->settingj 中，如果 settingi = settingj 并且 testi 和 testj 有部分项测试条件相同，这样就可以进行规则合并。例如：

合并前：

lex=广播&pos=n

1^pos=n;^1pos=n;->pos=v

1^pos=p;^1pos=n;->pos=v

1^pos=p;^1pos=b;->pos=v

1^pos=c;^1pos=n;->pos=v

合并后：

lex=广播&pos=n

1^pos=n|pos=p|pos=c;^1pos=n;->pos=v

1^pos=p;^1pos=b;->pos=v

从例子可以看出，在一个规则块内，规则数目从 4 条减少到 2 条。规则合并可以大大减少规则的规模，从而提高标注效率。

5. 实验结果

5.1 实验用的语料库

实验使用了北京大学计算语言学研究所的 PFR 标注语料库（98 年 1 月），它包括 1,062,174 个词条数。本实验只做词性大类标注规则的学习。下面是语料库（词性为大类）的片断：

中共中央/n 总书记/n 、/w 国家/n 主席/n 江/n 泽民/n

本报/r 驻/v 法国/n 记者/n 果/n 永毅/n

词性列表：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、习惯用语 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。

选取其中的 80%作为训练语料，20%作为测试语料。

5.2 自动学习规则

利用初始标注器标注训练语料，标注结果中标错的词条个数是 68973。使用自动学习规则算法进行训练，得到所有跨度 Len=1 的初始规则集。该规则集包含规则块 4919 个，所有规则条数总和是 20180，平均每个规则块包含规则条数是 4.102。

5.3 规则的初选

对初始规则集进行分析如表 1：

表 1 规则分析

出现次数	规则条数	占总规则比例	涉及错误数	占总错误比例
1	13186	65.3%	13186	19.1%
2	2837	14.1%	5676	8.2%
3	1250	6.2%	3750	5.4%
>=4	2907	14.4%	46361	67.3%
合计	20180	100%	68973	100%

表 1 中“涉及错误数”表示是从这么多的错误推出对应的规则。

从表 1 可以看出出现次数是一次的规则占总规则数的 65.3%，但是涉及错误数占总错误数比例仅为 19.1%，也就是出现次数是一次的规则数目庞大，但是标注作用不是很大。

去掉出现一次的规则，得到初选规则集，它包含规则块 1561 个，规则总数是 6994。该规则集占自动学习规则集的 34.7%，而涉及错误数占总错误数比例则达到 80.9%。这样可以既有效的保证了标注正确率，又大大降低了规则的规模。

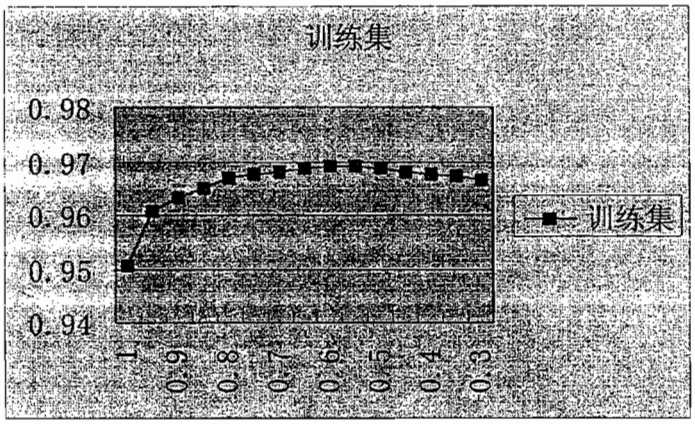
5.4 规则冲突的解决

在初选规则集中，涉及冲突的规则总共有 195 条，冲突对有 97 对，占初选规则集规则的 2.8%。去掉冲突规则，得到新的规则集，它包含规则块 1559，规则总数 6799。

5.5 规则有效性的分析

下图是寻找最佳有效阀值 C_m 结果：

图 2 寻找最佳有效阀值 C_m 结果



从图 2 中可以分析得出当有效值 C_m=0.55 的时候，训练标注正确率达到最好，达到 96.90%。删除 $C < C_m$ 的规则，得到新的规则集，包含规则块是 1445，规则条总数是 5823 条。

5.6 规则的合并

规则合并前，规则集所包含的规则总数是 5823 条，通过合并，规则集的规则总数下降成 2570 条，减少了 55.86%。通过规则合并，可以大大降低规则集内规则的规模。有效提高标注效率。

5.7 测试

首先利用初始标注器进行标注，标注结果：封闭测试集正确率是 91.87%，开发测试集是 91.90%。然后利用获取的规则标注结果如表 2 所示：

表 2 标注结果

规则类型	封闭测试	开放测试
1	96.90%	96.24%
2	97.37%	96.47%
3	97.41%	96.47%

表 2 中规则类型 1 表示 Len=1 情况下学习到的规则标注结果。

表 2 中规则类型 2 表示在类型 1 规则标注下结果学习 Len=2 情况下的规则标注结果。

表2 中规则类型3 表示在类型2 规则标注下结果学习 Len=3 情况下的规则标注结果。

从表2 可以看出，在规则类型1，封闭测试标注正确率提高了 5.03%，类型2、3 对类型有很好的补充效果，标注正确率又提高了 0.51%。

6. 结束语

词性标注规则自动学习一直是一个重要课题。本文提出一种词性标注规则自动学习和规则优化的算法。通过对规则进行优化，有效提高标注正确率和标注效率。利用这种算法，对 PFR 标注语料库（98 年 1 月）进行学习，获取 1445 个规则块，总共有 2570 条规则，从而实现了一个词性标注系统。该系统进行测试，相对于初始标注器的标注结果，封闭测试正确率提高了 5. 54%，开放测试提高了 4. 57%。

参考文献

[1] 孙杰,林鸿飞,姚天顺. 一种获取机器翻译系统词类搭配规则的机器学习方法,模式识别与人工智能,1999.6

[2] 李晓黎,史忠植. 用数据采掘方法获取汉语词性标注规则,计算机研究与发展,2000.12

[3] 魏欧,孙玉芳. 基于非监督训练的汉语词性标注的实验与分析,计算机研究与发展,2000.4

[4] 魏欧,吴健,孙玉芳. 基于统计的汉语词性标注方法的分析与改进,软件学报,2000.11

[5] Eric Brill. A Simple Rule-based part of speech tagger. In: Proc 3rd Conference on Applied Natural Language Processing, ACL,Trento,1992

[6] Eric Brill. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. Computational Linguistic, 1995,21(4):543-565

[7] 周明,吴进,黄昌宁. 用于词性标注的一种快速学习算法—对 Brill 的基于变换算法的一项改进,计算机学报,1998.4

[8] 朱靖波. 面向机器翻译的统计消歧技术研究,东北大学博士论文,1999.7

作者简介：陈文亮(1977-)，男，福建龙岩人，博士生，主要研究方向：中文信息处理和信息安全；朱靖波(1973-)，男，浙江人，副教授，博士，主要研究方向：中文信息处理；吕学强(1970-)，男，辽宁抚顺人，博士生，主要研究方向：机器翻译；姚天顺(1934-)，上海人，男，教授，博士生导师，主要研究方向：计算语言学理论。

Acquisition and Optimization of Rules for Part of Speech Tagging

Chen WenLiang　Zhu JingBo　Lv Xueqiang　Yao TianShun

Institute of Computer Software & Theory, Northeastern University, Shenyang 110004

Abstract: A learning algorithm is presented to acquire and optimize rules which are applied on part of speech tagging. It can increase the tagging precision and efficiency. Applying the algorithm, an experiment is conducted with PTR corpus. Comparing with tagged result of grammatical tagging based on NA assumption, the tagging precision increases 5.53% in close text and 4.57% in open text.

Keyword: Part of speech tagging; Automatic learning of rules; Chinese information processing