基于错误驱动的现代汉语方位词用法规则的自动更新*

吴云鹏，咎红英

郑州大学信息工程学院 郑州 450001

E-mail: ztyzoiwyp@163.com ;iehyzan@zzu.edu.cn

摘 要：基于规则的现代汉语方位词的用法标注有助于文本内容的自动理解，由于人工制定的规则具有不完备性，为了提高准确率，往往需要人为考察错误标注语料来完善方位词的用法规则，这是费时费力的事情。本文尝试采用基于错误驱动的原理设计一种可行的算法让机器实现用法规则的自动更新，实验结果表明本算法能实现大部分方位词的用法规则的自动更新，进而提高了方位词用法标注准确率。

关键词：自然语言处理，方位词，用法标注，错误驱动

Automatic Updates of Position Words Usage Rules in Modern Chinese Based on Error-driven

Wu Yunpeng, Zan Hongying

College of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China

E-mail: ztyzoiwyp@163.com ;iehyzan@zzu.edu.cn

Abstract: Rule-based usages annotation of position words in modern Chinese is helped to automatic understanding of text content, but the rules made by manual is not complete, We have to inspect error tagging corpus to perfect position usage rules if we want to improve the accuracy of rule-based usages annotation, which is laborious job, This paper attempt to use the theory based on error-driven to design a possible solution to update the rules automatically, the results show that the method can achieve most update of the rules of position words automatically, and then improve the accuracy of annotation of position words usages rules.

Keywords: natural language processing, position words, usages annotation, error-driven

1. 引言

汉语虚词用法的自动标注是自然语言处理领域的一个崭新的研究课题。目前，虚词用法自动标注方法目前有两种：基于统计的标注方法和基于规则的标注方法。本文采用基于规则的方法，做了以下的研究：

第一，方位词词典[1]，根据《现代汉语虚词辞典》[2]、《现代汉语八百词》[3]、《现代汉语词典》[4]和《人民日报》语料整理出所有方位词的用法词典。

第二，方位词规则库，根据方位词用法词典，整理出每个方位词的用法规则。

第三，根据规则实现计算机自动标注。

然而目前所得到的方位词用法规则库由于语料的局限性，还不能做到“完备精确”，需要在考察大规模真实语料的基础上人工去补充、完善用法规则，而人工在不断重复

*基金项目：本文承国家自然科学基金项目（项目号 60970083）、北京大学计算语言学教育部重点实验室开放课题基金（KLCL-1004）和河南省科技创新人才杰出青年基金项目（项目号 104100510026）的资助。

判别、更新规则需要耗费大量时间、精力，刘锐在《基于错误驱动的现代汉语副词用法的自动识别研究》[5]一文对具有单条规则的副词用法规则的自动更新完善进行过初步研究，买志玉在《现代汉语方位词用法知识库的研究》[6]得出以词典例句作为实验语料时自动标注的准确率能够达到 81.32%，本文在其基础上，以更大规模更真实的《人民日报》作为实验语料，尝试采用错误驱动的方式设计一个切实可行的算法让计算机更加高效的进行自学习、更新具有多条规则的方位词规则库，得到更加完备的用法规则库，使方位词用法的标注准确率得到进一步提高。

2. “三位一体”方位词用法知识库

根据俞士汶等提出的构建“三位一体”的现代汉语虚词知识库的思想[7]，本文做了以下准备工作，且算法设计在其基础上展开。

2.1. 方位词词典和用法规则库

方位词词典、方位词规则库的完成为实现计算机自动标注奠定了基础。词典方面，目前现代汉语方位词用法信息词典中共收录方位词词条 230 个，涉及 567 个不同的用法（或义项）属性的描述；规则库方面，在现代汉语方位词用法信息词典的基础上，根据方位词用法特征的不同表现，抽取其中可操作的判断条件特征，以有序的 BNF 形式进行方位词用法的规则描述，为现代汉语方位词用法的自动识别提供形式化依据。[8]

下面是方位词“上边”用法规则的样例：

$上边

@<f_shang4bian5_1e>→ML ^M→在|往|朝|向|自|从|由|p ^L→n|r

@<f_shang4bian5_1c>→L ^L→(j|r|s|n)[的]

@<f_shang4bian5_1d>→L ^L→n

@<f_shang4bian5_1b>→M ^M→在|往|朝|向|从|由|自|p

@<f_shang4bian5_1a>→F ^F→~

规则中的符号系统采用的是昝红英等的虚词用法规则的用法属性及规则描述规范[9]及赵丹在《面向机器识别的现代汉语方位词用法形式化描述研究》[8]中所述，如下所示

F→词 1 | 词 2 | … | a | v | n | …

M→词 1 | 词 2 | … | a | v | n | …

L→词 1 | 词 2 | … | a | v | n | …

R→词 1 | 词 2 | … | a | v | n | …

N→词 1 | 词 2 | … | a | v | n | …

E→词 1 | 词 2 | … | a | v | n | …

其中，(1)句首 F (2)前合用 M (3)前连用 L(4)后连用 R(5)后合用 N(6)句末 E

2.2. 实验语料

本文选取 1998 年 1 月份和 2000 年 1 月份《人民日报》分词与词性标注的语料作

为自学习算法的实验语料。为了说明方便，我们抽取了标注语料中部分例句如下：

(1)在/p 十五大/j 精神/n 指引/v 下/f<f_xia4_1k> 胜利/vd 前进/vi ——/wp 元旦/t 献辞/n

(2)科学/a 理论/n 指导/v 下/f<FAIL> 的/ud 实践/vn 解决/v 了/ul 无数/m 难/a 解{jie3}/v 的/ud 课题/n ，/wd 给/p 我们/rr 带来/v 许多/m 便利/vn ，/wd 但/c 同时/d 也/d 一定/d 会/vu 展示/v 更/dc 大/a 更/dc 多/a 的/ud 未知/vn 领域/n ，/wd 吸引/v 我们/rr 去/vi 探索/v 。/wj

如上所示(1)为标注成功的样例 (2)为标注失败的样例。其中标注规范采用的是俞士汶等在《北京大学现代汉语语料库基本加工规范》[10]一文中所描述的内容。

买志玉在《现代汉语方位词用法知识库的研究》[6]中采用的主要是词典中的例句语料库，与其不同的是本文描述的算法都是在《人民日报》这样大规模真实语料环境下进行的实验，可能对实验结果的准确率造成的一定的影响，但在可接受范围之内。

3. 基于错误驱动的方位词规则自学习算法及实现

3.1. 错误驱动原理

基于转换的错误驱动的学习方法是EricBrill[11]提出的，最初是用于英文的词性标注。其基本思想是：用标注过的文本作为训练语料库。首先采用一种初始标注方法对语料库进行标注，然后将标注结果与正确的文本进行比较，通过预先设计好的转换模板和目标函数，找出应用一条转换模板后可产生标注错误次数最少的转换式，作为一条系统的新的标注规则，再用该规则重新标注语料库。重复上述过程，每次循环都会得到一条新的标注规则，直到找不出这样的规则为止。

因此基于转换的错误驱动的学习方法首先应有以下三部分资源：

(1)带标注的训练语料库，对于语料标注任务来说，训练语料要标注出其中所有正确的标记信息。

(2)规则模板集合，用于确定可能的转换规则空间。

(3)一个初始标注程序。

具体的基于转换的错误驱动的学习算法是:①初始标注利用一个初始标注器来标注训练语料库。②生成候选规则集在每个初始标注错误的地方，规则模板便用来生成候选规则，规则的条件就是词的上下文环境，动作就是改正错误标记所要做的动作。③获取规则把候选规则集中的每条规则分别运用于初始标注的结果，选出得分最高的规则(采用目标函数进行评分)。把这条规则运用于初始标注的结果作为下一轮循环的基础，并把这条规则作为规则序列中的第一条规则输出。重复以上过程直到得分最高的规则的得分为0或低于某个阈值为止。

本文的规则更新算法是在方位词用法初标语料中，把当前规则识别失败的关键字（标注为<FAIL>）作为更新目标，主要基于规则，同时借助统计的方法，实现

(1)对待更新规则的选择

(2)以选出的规则为基础，进行规则库的更新

这样循环往复，实现了规则库自学习，完善了方位词用法规则库，进而提高了方位词用法自动标注的正确率。

3.2. 待更新规则选择算法

对于语料中标注失败的关键字（语料中标注为<FAIL>），将其前后一定范围（即小句，碰到任何除了顿号、以外的其它标点，如逗号、句号、叹号、问号等就停止搜索）的词语提取出来，对L、R、M、N、F、E六类用法特征统一用评估函数进行估值，量化了当前规则接近程度，选出评估值最高的规则，在此基础上，结合当前语料，提出相应的更新建议，实现对规则的自学习。

我们将整个规则更新算法的实现分为四个阶段：

1） L，R用法特征的评估

由于 L，R 位置清晰明确，即前紧跟或后紧跟一个词性或词语，所以第一阶段集中实现满足这两个规则的储存及评估。

当前 L，R 位置的评估值：

Value_LR[rule_num] = (key_L in Set_L[rule_num]) + (key_R in Set_R[rule_num])

其中 rule_num 为当前关键字的第 rule_num 条规则（为说明方便，本文 rule_num 指代同一含义）。key 代表当前关键字的词性，key_L 和 key_R 代表其前后紧邻词语的词性，Set_L、Set_R 来分别保存当前关键字的 L 和 R 规则。

2） M，N用法特征的评估

相对于 L，R 位置的固定，M，N 要灵活的多，所以我们必须要采取其他方法来处理。

在 L，R 的处理过程中，我们关注的是某一个关键字前后词语或词性，从而很容易利用评估函数找到最适合的规则进行更新。而 M，N 要求的是在小句中搜索，这样前后都增加了不少词语或词性，面临的新问题是到底选取哪些进入新规则。仅仅看单一的关键字及其前后语境显然不能满足要求，对此，我们采用下面的方法：

扫描实验语料，找出某一关键字 k_word 所有标注失败的记录，并按照查找顺序将其标号，设其为 i(i = 1..N, N 为关键字是 k_word 并且用现有规则标注失败的总个数)，对于第 i 个标注失败记录，定义 S[i]保存从当前关键字位置起，在小句范围内向前向后扩展，保存出现的词语和词性。最后把这样的 S[i]聚合在一起进行统计，把出现位置类似且频率较高的词语或词性抽取出来，作为更新规则的候选项。

这种方法的前提是语料要足够大足够丰富，这样某种标注失败会出现不止一次，最后在 S[i]合并筛选时，能把共同的缺陷提取出来，这样提出了 M，N 用法标注的解决方案

当前 M，N 位置的评估值：定义统计结果中词性出现的频率 f = Frequency($\bigcup$ S[i])，在前合用及后合用中选取其中相对出现频率较高的词语或词性，作为进行评估的参数，分别将他们保存在集合 T_front 和 T_after 中。因此我们定义其评估值：

Value_MN[rule_num]=(T_front[j] in Set_M[rule_num])+(T_after[j] in Set_N[rule_num])

其中 Set_M、Set_N 来分别保存当前关键字的 M 和 N 规则。

3） F，E用法特征的评估

由于 F， E 的位置也很固定，所以可以在 L，R 的基础上，完成对 F, E 的存储及评估函数的进一步丰富。当前 F，E 位置的评估值：

Value_FE[rule_num] = (key_F in Set_F[rule_num]) + (key_E in Set_E[rule_num])

其中 key 代表当前关键字的词性，key_F 和 key_E 代表其句首和句尾单词的词语或词性，Set_F、Set_E 来分别保存当前关键字的 F 和 E 规则。

4） 待更新规则的选择

六类用法特征的识别、储存、评估工作已经分别完成，但我们更新规则时要综合考虑上述因素，不妨设 W_LR, W_MN, W_FE 为 Value_LR ，Value_MN，Value_FE 对最终评估值的影响程度，即权值。我们定义最终的评估值：

Value = Max{Value_LR[rule_num] * W_LR + Value_MN[rule_num] * W_MN + Value_FE[rule_num] * W_FE }

这样完成了对待更新规则的选择。有种特殊情况是全部规则的评估值都为 0，即当前语料中语境与规则中有很大出入，根据人工校验时的经验，我们采用添加一个空规则项，这时没有前后语境的规则限制，用以表示当前词语语义为“单用”。

3.3. 规则自学习

在完成了待更新规则的选择后，将刚刚评估过程中得到的词语或词性更新候选项代入规则，分别对规则中的六类用法特征进行添加更新工作，这样，最终完成了对于识别 FAIL 的方位词关键字的规则更新工作。

4. 实验与结果分析

将 1998 年 1 月份《人民日报》经过分词与词性标注的语料作为自学习算法的实验语料，标注失败<FAIL>的方位词共有 241 处，其中 205 处可以被正确更新，自动更新的比率(简称更新率)达 85.0%。剩下不能更新的规则有相当一部分是兼类词在语料中词性初始标注错误造成的，数量为 34 个，除去这个因素，更新率高达 95.6%，达到了预期目标，使人工校对工作大大减轻。

又进一步用 2000 年 1 月份《人民日报》经过分词与词性标注的语料进行实验，标注失败<FAIL>的方位词共有 152 处，其中 143 处能够被正确更新，更新率达 94.0%，更加验证了实验效果。

但其中可能出现对规则修改后，原来标注正确的词语，在使用新规则的情况下变为错误标注问题，这将在下一步工作中进行优化。

下面是一些程序结果演示样例：

语料例句 1：

(1) 仅/d 1 9 9 6 年/t 冬/Tg 以来/f<FAIL> 一/m 年/qt 时间/n ，/wd 以/p 户/Ng 办/v 、/wu 联/Vg 户/Ng 办/v 的/ud 形式

/n ，/wd 投资/v 2500/m 多/m 万/m 元/qd ，/wd

“以来”的原始规则：

$以来

@<f_yi3lai2>→[M]L ^M→在|自|自从|从|p ^L→t|v|n|m|j|f|<rz>

更新后的规则

@<f_yi3lai2>→[M]L ^M→在|自|自从|从|p ^L→t|v|n|m|j|f|tg|<rz>

语料例句 2：

(1)陈/nrf 金水/nrg 出名/a 后/f<FAIL> ，/wd 几乎/dc 挤/v 不/df 出/vq 时间/n 照看/v 自家/rr 的/ud 菜园子/n 了/y 。/wj

“后”的原始规则

$后

@<f_hou4_1ac>→L ^L→最

@<f_hou4_1da>→R| L ^R→m|q ^L→先#[，]

@<f_hou4_2>→F ^F→~

@<f_hou4_1bb>→L ^L→q|t|饭|月|小时|赛|会|式

@<f_hou4_1db>→R ^R→m(秒|分钟|刻|小时|时辰|天|周|月|季度|年)

@<f_hou4_1ba>→L ^L→j|r|n|s

@<f_hou4_1aa>→M ^M→前#[，]

@<f_hou4_1c>→M ^M→v

@<f_hou4_1ab>→M ^M→在|往|朝|向|自|从|由|p

待更新选择算法选取的规则：

@<f_hou4_1ba>→L ^L→j|r|n|s

更新后的规则：

@<f_hou4_1ba>→L ^L→j|r|n|s|a

以上两例是基于规则，对单个关键字的 L 规则更新。

语料例句 3：

(1)我/rr 愿/v 借此机会/lv ，/wd 向/p 长期/d 来/f<FAIL> 为{wei4}/p 促进/v 中国/ns 和/c 南非/ns 人民/n 之间/f<?f_zhiljianl_1c> 的/ud 了解/vn 和/c 友谊/n ，/wd

(2) 百年/mq 来/f<FAIL> 中华民族/n 灾难深重/l ，/wd 到/v 了/ul 危亡/n 的/ud 最后/f<?f_zui4hou4_1> 关头/n ，/wd

(3) 更/dc 是/vl 对/p 百年/mq 来/f<FAIL> 我们/rr 民族/n 的/ud 深重/a 苦难/n 和/c 中国/ns 革命战争/nz 的/ud 艰难/a 历程/n 所/us 知/v 甚/Dg 少/a 。/wj

(4) 多年/mq 来/f<FAIL> ，/wd 宜阳县/ns 工农业/jn 用电/v 紧张/a ，/wd

“来”的原始规则：

$来

@<f_lai2>→M ^M→t|q|n

更新后的规则：

@<f_lai2>→M ^M→t|q|n|mq

本例基于统计方法，对关键字多个<FAIL>样例统计分析，完成 M 规则更新。

5. 下一步工作

由于上述方法主要采用了基于规则的方法，而人工总结出来的规则是从词典例句语料中提取出来的，不能涵盖大规模真实语料中方位词的全部用法信息，这种情况不可避免地造成了算法识别准确率提高的瓶颈问题，所以在下一步工作中，我们将重点考虑基于统计的方法来改进规则更新算法，使其更新准确率进一步提高。另外在规则自学习的过程中，可能会出现原本正确的标注被更新后的规则标注错误，这个问题需在下一步的算法优化工作中考虑。

参 考 文 献

[1] 刘云. 汉语虚词知识库的建设. 北京大学博士后出站报告, 2004

[2] 张斌 主编. 《现代汉语虚词词典》. 商务印书馆看, 2005

[3] 吕叔湘. 《现代汉语八百词》. 商务印书馆, 1980

[4] 中国社会科学院语言研究所词典编辑室. 《现代汉语词典》. 商务印刷馆, 2005

[5] 刘锐、昝红英、张坤丽. 现代汉语副词用法的自动识别研究. 计算机科学, 2008 年 8 月

[6] 买志玉. 现代汉语方位词用法知识库的研究. 郑州大学自然语言处理实验室, 2009 年

[7] 俞士汶、朱学锋、刘云. 现代汉语广义虚词知识库的建设. 《汉语语言与计算机学报》, 2003 年第 1 期, 89-98

[8] 赵丹、张坤丽、昝红英、买志玉. 面向机器识别的现代汉语方位词用法形式化描述研究. 第十一届汉语词汇语义学研讨会（CLSW2010）论文集, 苏州大学, 2010 年 5 月

[9] 昝红英、张坤丽、柴玉梅、俞士汶. 现代汉语虚词知识库的研究. 《中文信息学报》, 2007 年第 5 期, 107-111

[10] 俞士汶、段慧明、朱学峰、孙斌. 北京大学现代汉语语料库基本加工规范. 《中文信息学报》, 2002 年第 6 期, 58-64

[11] EricBrill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics, Volume 21, Number 4, 1995

[12] 俞士汶、段慧明、朱学峰、张化瑞. 综合型语言知识库的建设与利用. 《中文信息学报》, 2004 年第 5 期, 1-10

[13] 俞士汶、朱学锋、刘云. 面向自然语言理解的汉语虚词研究. 《民族语言文字信息技术研究》, 2007 年 2 月, 270-277

[14] 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 年 4 月第 21 卷第 4 期