文章编号：1003-0077(2009)04-0016-06

基于 SVMTool 的中文词性标注

王丽杰，车万翔，刘 挺

（哈尔滨工业大学 计算机学院 信息检索研究室，黑龙江 哈尔滨 150001）

摘 要：SVMTool 是建立在支持向量机(SVM)原理上的序列标注工具，具有简单、灵活、高效的特点，可以融入大量的语言特征。该文将 SVMTool 应用于中文词性标注任务，将基于隐马尔科夫模型的基线系统准确率提升了 2.07%。针对未登录词准确率不高的问题，该文加入了中文字、词的特征，包括构成汉字的部首特征和词重叠特征，并从理论上分析了这两个特征的可行性，实验显示加入这些特征后，未登录词标注的准确率提升了 1.16%，平均错误率下降了 7.40%。

关键词：计算机应用；中文信息处理；词性标注；SVMTool；未登录词；偏旁部首

中图分类号：TP391 **文献标识码**：A

An SVMTool-Based Chinese POS Tagger

WANG Lijie, CHE Wanxiang, LIU Ting

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: The SVMTool is a simple, flexible and effective generator of sequential tagger based on Support Vector Machines, capable of dealing with a large number of linguistic features. In this paper, SVMTool is applied in Chinese POS tagging task and improves the accuracy by 2.07% compared with the baseline system on the Hidden Markov Model. To further improve the accuracy of unknown words, we introduce some features of Chinese characters and words, such as radicals of Chinese characters and reduplicate words, and probe into a theoretical analysis for their feasibility. Experiments indicate that these features can improve the accuracy of unknown words by 1.16% as well as reduce the error rate by 7.40%.

Key words: computer application; Chinese information processing; part of speech tagging; SVMTool; unknown word; radicals of Chinese

1 引言

所谓词性标注(Part of Speech tagging)就是根据句子上下文中的信息给句中的每个词一个正确的词性标记，即确定每个词的词性是名词、动词、形容词或者是其他词性。词性标注主要是针对多标记词(即有多种词性的词)和未登录词(即在训练语料中未出现的词)进行的。据统计，Brown 语料库中英语词汇多标记词的出现次数占 55.2%[1]；一个 13 万词的汉语语料库中多标记词的出现次数占 23.6%[1]。然而，各种多标记词在特定的上下文中总是具有确定的词性。同样，未登录词在特定的上下文中也是具有词性的。

词性标注是自然语言处理领域的基础，可以提高信息检索的效果和效率，它在信息检索领域有着非常重要的作用[2]。国内外该方面研究人员很重视它，成功设计出很多词性标注模型。归纳起来，比较典型的标注算法有：

(1) 基于规则的方法

基于规则的标注系统与系统设计者的语言能力有关，规则集直接体现了设计者的语言能力。最初

收稿日期：2008-06-23 **定稿日期**：2008-12-29

基金项目：国家自然科学基金资助(60803093,60675034)；国家 863 计划资助项目(2008AA01Z144)

作者简介：王丽杰(1986—)，女，硕士生，主要研究方向为自然语言处理；车万翔(1980—)，男，讲师，主要研究方向为自然语言处理和信息检索；刘挺(1972—)，男，教授，博士，主要研究方向为自然语言处理和信息检索。

的词性标注系统就是采用了这种方法，著名的TAGGIT系统，利用3 300条上下文规则，对100万词的Brown语料库进行标注，准确率达到77%[3]。不幸的是，要对某一种语言的各种语言现象都构造规则的话，将是一项很艰难也很耗时的任务。基于规则的标注系统另一个常见问题是：当根据规则判断一个词的词性时可能面临多种选择，如果不根据上下文则很难做出正确的选择。

(2) 基于统计的方法

20世纪80年代初，随着经验主义方法在计算语言学中的重新崛起，统计方法在词性标注中占据了领导地位，是目前最常使用的一种方法。对于给定的输入词串，基于统计的方法先确定其所有可能的词性串，然后对它们分别打分，并选出得分最高的词性串作为最佳的输出。常见的方法有基于N元模型的方法和基于隐马尔科夫模型的方法。其中，隐马尔科夫模型结合Viterbi算法的词性标注方法[4]最为常见与成熟。近年来，决策树[1]、最大熵模型[1]和条件随机域[5]等方法也被用在词性标注上，并取得了不错的效果。

(3) 规则与统计相结合的方法

这种方法结合规则和统计两种方法的优势，弥补对方的缺点，能够有效地进行词性标注。国内北京大学计算语言学研究所提出了一种先规则、后统计的规则和统计相结合的标注算法，其准确率达到了96.6%[6]。

本文中训练的中文词性标注模型使用了一种基于统计方法的词性标注工具——SVMTool[7]。SVMTool是建立在支持向量机原理上的序列标注工具，它使用待标注语言的特征。本文加入了中文词语的特征，即构成汉字的部首特征和词重叠特征，应用它们提高未登录词的词性标注准确率。在理论上分析了这两个特征的可行性，并在实验中证明了这两个特征可以提高未登录词的标注准确率。通过实验证明，SVMTool同样适用于中文词性标注任务。加入这些新的中文词语特征后，当未登录词的比率为1.88%时，总的词性标注准确率可达97.84%，未登录词的标注准确率达85.48%。

2　运用SVMTool进行中文词性标注

SVMTool是基于SVM[8]原理的简单的、灵活的、高效的序列标注工具，每秒标注上千词[7]，非常适合于自然语言处理中的序列标注问题。根据文献[7]介绍，SVMTool的标注准确率已经超过基于隐马尔科夫模型的Trigrams'n'Tags(TnT)[9]词性标注工具，在《华尔街日报》语料上英文标注准确率为97.2%[7]，是目前世界上最好的词性标注工具之一。SVMTool已经成功应用于西班牙语和加泰罗尼亚语，准确率分别为96.89%[7](LEXESP语料)和97.77%[7](3LB语料)。

通常，词性标注既要求准确率又要求高效率，但是二者往往不可兼得，SVMTool在二者之间达到了一个很好的平衡[7]。对于标注的其他问题，诸如：(1)不同语言的词法规则是不同的，这就需要考虑语言的特征；(2)不同语言的标注集和多标记词的概率是不同的；(3)当训练数据较小时，则会出现大量的未登录词等；SVMTool都能很好的解决它们：(1) SVMTool提取的就是语言的特征，即使不同的语言有不同的词法规则，它能够提取这种词法规则并利用其进行词性标注；(2)SVMTool只需要一个标注好的训练语料，标注集是从训练语料中提取的，当然你也可以给定标注集；(3)SVMTool在训练的时候会选出部分词作为未登录词，训练时会生成未登录词的标注模型，这对未登录词的标注准确率的提高起到了很大的作用。

SVMTool在训练模型时根据丰富的特征集提取训练语料的信息，应用SVM原理来训练分类器，使用了SVM-light①。SVMTool在训练模型时不仅生成了已登录词的标注模型，还会根据给定的比率选出部分词作为未登录词，应用这些词生成未登录词的词性标注模型，用于标注未登录词，由此可以提高未登录词的标注准确率。

SVMTool标注时对文件进行按句处理，根据每一个待标注词的分类(是否为已登录词)，选择训练好的模型(针对已登录词和未登录词的)。然后根据选择的模式(贪心模式或句子模式)进行按词标注。贪心模式是按词进行标记的，每个词选取一个概率最大的标记，标记好的词可以为其他词提供上下文环境。句子模式对词进行标记，并进行概率和相加，然后针对整个句子选择一个概率最大的标注序列。

SVMTool应用SVM原理主要是利用待标注语言的特征，特征提取的好坏直接影响词性标注的准确率，特征收集的越好越贴切则准确率会越高。本文主要考虑的是提高未登录词的标注准确率，所有加入的新特征都是针对未登录词的。本文的主要

① http://svmlight.joachims.org

研究内容是提取中文词语的特征并将其付诸于实验中，观察其对未登录词标注准确率的影响，对可以提高未登录词标注准确率的特征进行模型训练及优化组合，使未登录词标注准确率及总标注准确率进一步提高。

SVMTool的未登录词的原特征[7]如表1。

表1 未登录词的原特征

Word	W(-2),W(-1),W(0),W(1),W(2),W(-2,-1),W(-1,0),W(0,1),W(-1,1),W(1,2),W(-2,-1,0),W(-2,-1,1),W(-1,0,1),W(-1,1,2),W(0,1,2)
PoS	P(-2),P(-1),P(-2,-1),P(-1,1),P(1,2),P(-2,-1,1),P(-1,1,2)
Ambiguous	k(0) k(1) k(2)
Maybe's	m(0) m(1) m(2)
前/后缀化特征	a(1) a(2) a(3) a(4) z(1) z(2) z(3)z(4)
其他特征	L SA AA SN CA CAA CP CC CN MW

其中各特征的含义为：(1)Word、PoS、Ambiguous、Maybe's分别表示词n元组、词性n元组、该词所有可能标记、每一个可能的标记，下标是相对于待标记词的位置；(2)a(i)、z(i)表示前后缀特征，i为相对于词开始或结尾的位置；(3)SA、CA、CAA、AA、SN、CP、CN、CC、MW、L分别表示以大写字母开头、包含一个大写字母、包含若干大写字母、所有字母均为大写、以数字开头、包含"."、包含数字、包含","、包含连字符、词长度。

英语的词干加上不同的后缀可以向名词、动词、形容词以及副词转化，在英语中后缀对于识别未登录词的词性是很有帮助的。同样，中文词语也有后缀特征，但不像英语词语的后缀特征那样暗示词性的转化，而是主要被用于识别地名、机构名或其他专有名词。例如：

后缀为"市"，如北京市(地名，ns)，多为名词；

后缀为"化"，如市场化(动词，v)，多为动词；

后缀的组合为"部门"，如政府部门(名词，n)，多为名词；

后缀的组合为"委会"，如奥委会(简称，j)、特委会(简称，j)、村委会(简称，j)，多为简称。

由此可见，中文词语也可以提取后缀作为特征，本文保留了SVMTool的后缀特征。由于汉字占两个字节，所以本文对后缀特征进行了优化(使用z(2)、z(4))。同时本文还针对中文文本的特点，另外还优化了SVMTool的源程序(主要是标点符号的处理)，使其更适合于中文词性标注。

2.1 中文词语特征

根据中文词语以及汉字的构成方式，本文共提取了两个中文词语的特征并将其加入训练模型的特征集来训练模型。实验结果表明这两个特征是有效的。加入的两个特征如下：

1) 部首特征

汉字形成了一套运用非常灵活的基础结构——偏旁部首，汉字虽有数万之巨，偏旁部首却只有几百个。偏旁部首是汉字表意的基础，在汉字形义关系中有着重要的作用。它具有高度抽象的符号功能，这种抽象性更表现为意义上的抽象。偏旁部首实际上是对表现万物的汉字进行逻辑分类。在形式上表现为以形为纲，但实质上，偏旁部首的类聚作用表现为意义上的统领，某一部首下的所列的一系列具体汉字几乎都与该部首有着意义上的联系。例如，列在"木"部的字如杨、柳、森、林等，都与"木"相关；列入"车"部的字如轮、轻、辑、轩等，都与"车"有关；列入"示"部的字如神、祖、禅、祀等，都与祭祀有关。

正是因为这种特殊的形体结构，汉字字符本身就具有了直观性、象征性的特色。而且部首与汉字这种意义上的联系，对猜测某字的词性是有帮助的。所以，可以考虑把汉字的部首也作为中文词语的特征。如：

言字旁：说、记、论等，一般为动词；

立刀旁：剁、刮、判等，一般为动词；

提土旁：地、场、城等，一般为名词。

所以知道了一个汉字的偏旁部首就可以初步猜测它的词性，当然也不能够借此就完全准确地猜到它的词性，还要根据它所在的词以及该词所在的上下文环境进一步判断。

2) 重叠特征

由于中文词语中常会用重叠来强调语气，这种重叠词的出现频率不高，容易以未登录词的形式出现，而且这种词语可以通过去重后判断词性，原词语与去重后的词在词性上一般是相同的。例如：

高高兴兴：可以通过"高兴"来判断词性；

轻轻的：可以根据"轻"来判断词性等。

可见，这些词去重后一般不影响它的词性，可以根据去重后的词来判断它的词性。

2.2 特征的提取及组合

如果要通过部首特征来猜一个字的词性，则我

们需要考虑如何提取偏旁部首。通过调查得知，Unicode 的汉字编码就是根据偏旁部首进行的，例如部首为“宀”的汉字的 Unicode 编码是从 0X5186 到 0X519A，部首为“亻”的汉字的 Unicode 编码则是从 0X4EAA 到 0X512E，所以通过汉字 Unicode 编码就可以很容易找到该字的部首。

当知道一个汉字的部首时，可以初步猜测该字的词性，但是当由该字组成一个词时，则不能简单的通过一个字来猜测该词的词性，有时含有同一个字的不同词具有不同的词性，所以这就需要一个词的部首的组合。可以考虑类似于词的前后缀形式，提取一个词的前几个或后几个字的部首组合起来作为部首前后缀特征。

对于重叠词这一特征，通过对语料的调查，当二字重叠为“某某”，如果它与“地/的”分为一个词时（如“轻轻地/的”）多为形容词，如果它与“地/的”单独分开（如“轻轻/地”）多为副词。所以，提取重叠这一特征时分为三种情况：(1)当重叠词是类似于“高高兴兴”这种形式时，提取“高兴”；(2)当它是“轻轻地/的”这种形式时，则提取“轻”；(3)当它是“湛蓝湛蓝”这种形式时，则提取“湛蓝”。

本文使用了 1998 年 1 月～6 月的《人民日报》语料。在对新加入的特征进行组合及优化时，为提高系统开发的速度，在 1998 年 2 月～6 月《人民日报》(共 6 092 949 词)中随机抽取 30 万词作为训练集，并抽取 5 万词作为开发集。最终，本文加入的新特征组合结果见表 2。

表 2　中文词性标注的新特征组合

新特征组合	实现时表示
部首前/后缀化特征	bsa(1) bsa(2) bsa(3) bsz(1) bsz(2)
重叠特征	DOU

其中各特征含义为：(1)bsa 表示部首的前缀组合，bsz 表示部首的后缀组合，bsa(i)/bsz(i)中的 i 表示距离词开始(bsa)或词结束(bsz)的偏移量；(2) DOU 表示词重叠特征。

3　实验结果

本文优化了 SVMTool 的后缀特征，使其适合于中文词性标注。以下实验都是在优化后的基础上进行的，标注过程使用了 SVMTool 贪心模式，从左到右依次标注每个词的词性，同时使用了 SVMTool 的默认参数。

3.1　SVMTool 与隐马尔科夫模型(HMM)的比较

本文首先进行了与隐马尔科夫模型的比较。由于传统的隐马尔科夫模型中，未登录词的词汇发射概率是未知的，而且在开放测试的时候，该值的确定对准确率影响较大。所以，在训练隐马尔科夫模型的时候，我们应用了文献[10]中提到的计算未登录词的词汇发射概率的算法。根据词在训练语料中出现的次数，我们设置了一个阈值确定了部分词作为未登录词(出现次数小于该值的作为未登录词，在本文中我们把阈值定为 5)，然后由文献[10]给出的公式来计算未登录词的词汇发射概率。

本文采用 1998 年 2 月～6 月的《人民日报》作为训练语料，1998 年 1 月的《人民日报》作为测试语料，进行的是开放测试，未登录词的比率为 1.88%。在标注准确率上，SVMTool 比隐马尔科夫模型高出两个百分点(结果见表 3)。

表 3　SVMTool 与隐马尔科夫模型比较

	SVMTool	隐马尔科夫模型
准确率	97.80%	95.77%

同时，为了进一步比较 SVMTool 和 HMM 模型，本文还针对两个模型的各词性标注精度差异情况进行了分析(这里仅列出了准确率相差在 2.5% 以上的，其余的均为 SVMTool 的标注准确率比 HMM 模型的偏高)，如图 1。

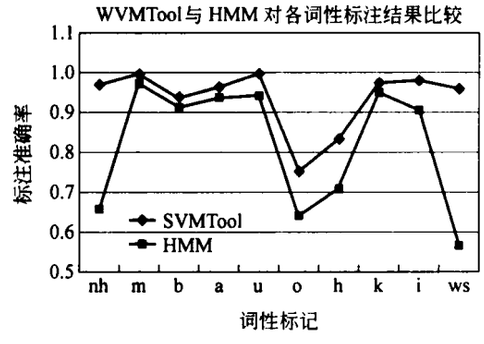

图 1　SVMTool 与 HMM 模型对部分词性标记结果比较

SVMTool 不仅使用了词的 n 元组和词性 n 元组等特征，还使用了待标注语言本身特有的特征，这些特征非常精确地刻画了训练语料的信息，训练出了一个准确率很高的模型。图 1 的结果充分显示了

丰富的特征集和良好的上下文环境在标注中所起的重要作用，譬如 m(数词)、h(前接成分)、k(后接成分)、ws(字符串)等词性的标注精度比较结果充分显示了特征集的作用，nh(人名)、u(助词)、a(形容词)等词性的标注精度比较结果则表明了其所利用的上下文环境较好。

为了与其他词性标注工作进行比较，本文还在2007年 SIGHAN 的 pku 词性标注语料进行了测试。在此，列出了在 pku 语料上标注结果最好的参赛单位的实验结果(文献[11])，以便对该任务的整体性能有所认识，如表4。

表4 与其他中文词性标注实验比较

实验出处	未登录词①准确率	总准确率
文献[11]	58.18%	94.50%
本文未加特征	57.09%	94.30%
本文加新特征	58.20%	94.45%

通过以上的实验结果比较，SVMTool 的标注准确率与文献[11]相当(总准确率相差 0.05%)。但是在进行词性标注时，本文的标注策略为贪心模式、一遍标注，标注速率约为 2 160 词/秒②；然而文献[11]采用的是 beam search 模式、两遍标注，因此本文的标注效率会更高。以上两点说明了 SVMTool 适用于中文词性标注任务。

3.2 新特征对未登录词的影响

本文针对未登录词，考察了每一个新加入特征对其总准确率的影响(结果见表5)，并进一步分析了在不同特征条件下各词性标记的准确率变化情况：部首特征使未登录词的标注准确率提升了0.95%(85.27%～84.32%)，图2显示了其对部分词性标注准确率的影响(这里列出了准确率提高在5%以上的)；重叠特征使未登录词的标注准确率提升了0.27%(84.59%～84.32%)，图3显示了其对部分词性标注准确率的影响(这里列出了准确率提高在1%以上的)；两个特征合起来使未登录词的标注准确率提升了1.16%(85.48%～84.32%)，图4显示了它们加入后对部分词性标注准确率的影响(这里列出了准确率提高在5%以上的)。

表5 加入新特征后与原特征的比较

	未登录词③准确率	未登录词的标注错误率下降比④
优化后缀特征	84.32%	—
部首特征	85.27%	6.06%
重叠特征	84.59%	1.72%
全加	85.48%	7.40%

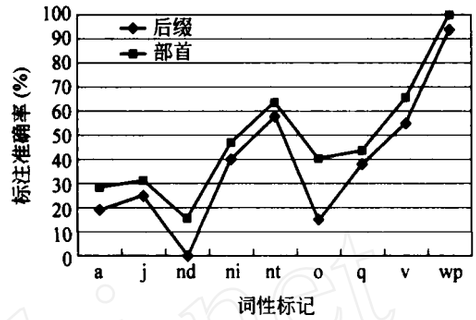

图2 部首特征对部分词性标注准确率的影响

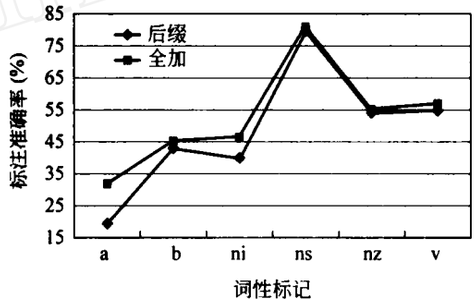

图3 重叠特征对部分词性标注准确率的影响

由表5可见，两个特征的加入使未登录词的标注错误率下降了7.40%。其中部首特征使未登录词的标注错误率下降达6.06%，主要是由于部首特征体现了汉字本身的特点，且其前后缀提取方式更好的描述了未登陆词词本身的特点。通过图2的结果可知，部首这一特征对中文未登录词的词性标注是有效的，譬如 a(形容词)、j(简称)、v(动词)、nd(方位名词)等词性标注准确率的提升充分说明了部首特征的重要性。重叠特征对性能的提升不如部首特征明显，这主要是由于重叠特征主要是针对形容词

① 这里的未登录词定义采用07年 SIGHAN 的未登录词定义，见 http://www.china-language.gov.cn/bakeoff08/。

② 实验机器配置：Pentium-IV，2GHz，1 GB RAM，C++ 版本的 SVMTool。

③ 这里的未登录词是指词本身在训练语料中未出现的，不包括那些该词在训练语料中出现，但是在测试语料中的某个词性却没有在训练语料中出现。

④ 标注错误率下降比 $p=\frac{Acc1-Acc2}{1-Acc2}$，其中 $Acc1$ 表示提升后标注准确率，$Acc2$ 表示提升前标注准确率。

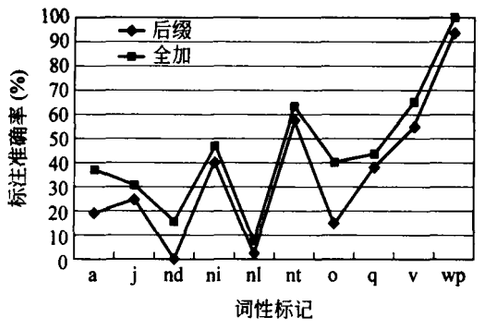

图 4　两个特征对部分词性标注结果的影响

的，由图 3 可知，重叠特征使未登录词中形容词的标注准确率提升了 17.34%；其次是由于重叠词出现的次数较少造成的，通过对测试语料（共 100 万左右词次）的统计，重叠词共出现 596 次，占总词次的 0.05%，而重叠特征使总标注准确率提升了0.02%；由此可见重叠特征对未登录重叠词的词性标注也是较为有效的。两个特征全加后，进一步提高了未登陆词的各词性标注准确率，图 4 的结果充分显示了两个特征在未登录词标注中所起的作用。

最终，两个新特征的加入使得词性标注的整体准确率由原来的 97.80%提升为 97.84%，性能提升不显著的主要原因是由于训练语料和测试语料来源相同，未登录词所占比重不大造成的。若将此词性标注工具应用于其他领域，如生物文本等，整体性能提升会更加明显①。

4　结语

通过与其他中文词性标注方法的实验结果比较，可以看出 SVMTool 的准确率是相对最高的。本文在针对中文优化后的 SVMTool 的原特征基础上加入了中文词语的特征（部首和重叠特征），有效地提高了未登录词的标注准确率，进一步提高了 SVMTool 的总标注准确率。尤其是部首特征，根据我们的调研，本文首次应用了偏旁部首进行中文词性标注。将部首作为特征加入特征集进行模型训练，有效地应用了汉语的这种特殊的形体结构及部首在汉字构成上的作用。实验结果证明了部首这一特征的有效性，它将未登录词的标注准确率提高了近 1 个百分点。下一步可以考虑加入更多中文词语特征来提高中文词性标注的准确率，还要进一步对 SVMTool 参数进行调整，使准确率进一步提升。

参考文献：

[1] 郭永辉，吴保民，王炳锡. 一种用于词性标注的相关投票融合策略[J]. 中文信息学报，2007，21(2)：9-13.

[2] 苏祺，胡景贺，等. 词性标注对信息检索系统性能的影响[J]. 中文信息学报，2004，19(2)：58-65.

[3] 张民，李生，等. 统计与规则并举的汉语词性自动标注算法[J]. 软件学报，1998，9(2)：134-138.

[4] 梁以敏，黄德根. 基于完全二阶隐马尔科夫模型的汉语词性标注[J]. 计算机工程，2005，31(10)：177-179.

[5] 洪铭材，张阔，等. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学，2006，33(10)：148-155.

[6] Zhou Qiang. An algorithm of tagging Chinese POS based on statistics and rule[J]. Chinese Information Journal, 1996, 9(3):1-9.

[7] Jesus Gimenez and Luis Marquez. SVMTool: A general pos tagger generator based on support vector machines[C]//Proceedings of the 4th LREC Conference. Lisbon, Portugal, 2004: 43-46.

[8] T. Joachims. Making large-Scale SVM Learning Practical[M]. Cambridge, MA, USA: MIT-Press, 1999: 41-56.

[9] T. Brants. TnT A Statistical Part-of-Speech Tagger [C]//Proceedings of the Sixth ANLP Conference. Seattle, WA, 2000: 224-231.

[10] 张孝飞，陈肇雄，等. 词性标注中生词处理算法研究[J]. 中文信息学报，2003，17(5)：1-5.

[11] Aitao Chen, Yang Zhang and Gordan Sun. A Two-Stage Approach to Chinese Part-of-Speech Tagging [C]//Sixth SIGHAN Workshop on Chinese Language processing. Indian, 2007: 82-85.

① 在本实验中当未登录词的比率为 1.88%时，两个特征的加入使总准确率提高 1.16%，其中部首特征使准确率提高 0.95%；当未登录词的比率为 6.74%时，它们使总准确率提高 1.57%，其中部首特征使准确率提高 1.37%。