文章编号：1672-2892(2012)02-0184-04

基于条件随机场的中文分词算法改进

顾佼佼[1a]，杨志宏[2]，姜文志[1a]，胡文萱[1b]

(1.海军航空工程学院 a.兵器科学与技术系；b.外训系，山东 烟台 264001；
2.海军装备部驻武汉地区军事代表局，湖北 武汉 430064)

摘　要：在中文分词领域，基于字标注的方法得到广泛应用，通过字标注分词问题可转换为序列标注问题，现在分词效果最好的是基于条件随机场(CRFs)的标注模型。作战命令的分词是进行作战指令自动生成的基础，在将 CRFs 模型应用到作战命令分词时，时间和空间复杂度非常高。为提高效率，对模型进行分析，根据特征选择算法选取特征子集，有效降低分词的时间与空间开销。利用 CRFs 置信度对分词结果进行后处理，进一步提高分词精确度。实验结果表明，特征选择算法及分词后处理方法可提高中文分词识别性能。

关键词：中文分词；条件随机场；特征选择；置信度

中图分类号：TN911.72；TP391.1　　**文献标识码：**A

Improvement on CRFs-based Chinese word segmentation algorithm

GU Jiao-jiao[1a]，YANG Zhi-hong[2]，JIANG Wen-zhi[1a]，HU Wen-xuan[1b]

(1a.Department of Ordnance Science and Technology；1b.Department of Foreign Training，Naval Aeronautical and Astronautical University，Yantai Shandong 264001，China；2.Military Representatives Bureau of NED in Wuhan，Wuhan Hubei 430064，China)

Abstract：In Chinese word segmentation fields，the most widely used method is character-based tagging，which reformulates segmentation task to a sequence tagging task. The Conditional Random Fields (CRFs) tagger is the best tagger which can achieve state-of-the-art performance. The segmentation of the command orders is one of the basics of the auto-generation of command orders. Yet when using the model for command orders segmentation，problems of bad time and space efficiency are encountered. The model is analyzed and feature subsets are selected by using the feature selection algorithm，which cut the overhead of time and space effectively and improve the efficiency of the model. Then a novel post-process using CRFs confidence is presented to further improve performance. By combining the feature selection method and the confidence-based post-process，great improvement is achieved and the experimental results are satisfactory.

Key words：Chinese word segmentation；Conditional Random Fields；feature selection；confidence

如今随着信息化技术的迅猛发展，互联网上的信息量呈现指数爆炸的增长趋势，海量文本信息使得文本信息的挖掘成为迫切需求。与西方语言不同，中文文本中并不存在词的分隔符，故中文分词[1-2]是中文信息处理的基本步骤，是自然语言处理的经典问题。近些年来中文分词得到了长足的发展。主流方法有传统的基于规则的[3-4]方法和现在流行的基于统计的方法。传统方法如前向最大匹配和反向最大匹配等，基于统计的方法主要有支持向量机(Support Vector Machine，SVM)[5]、隐马尔科夫模型(Hidden Markov Model，HMM)[6]和条件随机场(Conditional Random Fields，CRFs)[7]等。基于统计的方法建立在统计推断基础上，可得到比传统方案更高的性能。随着分词算法的不断改进，各分词方法的性能已经相差无几。目前达到最好分词效果的是基于 CRFs 的分词模型，但 CRFs 的主要问题是其训练效率偏低，模型本身决定了其时间复杂度和空间复杂度非常高，尤其现在新的语料、词汇不断涌现，预先训练好的模型不能适应开放性语料，模型需要及时更新，高速实时处理的分词系统成为迫切要求。如何提高其训练效率，使之适应快速变化的环境是实现该模型的一大挑战。

收稿日期：2011-05-24；**修回日期：**2011-08-23

1　基于条件随机场的算法改进

1.1 条件随机场模型

CRFs 是一种判别式模型，采用的是无向图分布，没有严格的独立性假设，可以任意选取特征，而且因为采用全局归一化的方法，避免产生标记偏移问题，所以在中文分词上优于 HMM 和最大熵马尔科夫模型(Maximum Entropy Markov Model，MEMM)等模型，取得较好的效果[8]，其中链式 CRFs 在中文分词任务中最常用。在给定观察序列条件下，标记序列的条件概率为：

$$p(y|x) \propto \exp\left(\sum_{e\in E,k}\lambda_k f_k(e,y|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,y|_v,x)\right) \tag{1}$$

式中：x 表示需要标注的观察序列集；y 表示相应的标注序列集；在一阶链式结构的图 $G=(V,E)$ 中，V 代表图中的节点集，E 表示图中的边，最大团仅包含相邻的 2 个节点，即图 G 的边。对 1 个最大团中的无向边 $e=(v_{i-1},v_i)$，$f_k(e,y|_e,x)$ 为状态转移特征函数；$g_k(v,y|_v,x)$ 为状态特征函数；λ_k 和 μ_k 是由训练样本得到的特征权重；k 为特征函数编号；v 为 V 中的节点。计算特征权重函数采用极大似然估计方法。CRFs 指数模型为凸函数，可采用迭代方法找到全局最优解。目前常用的是有限记忆 BFGS(Limited memory Broyden,Fletcher,Goldfarb,Shanno，L-BFGS)迭代方法。

1.2 标注集

引入标注集可把分词问题转化成序列标注问题，对于 1 个句子中的每个字给出相应的标签，等效地就知道了分词结果。LRMS 体系是一种常见的标注方法，每个字依据其在词中出现的位置给予不同标签，句子中的每个位置被标注为 LRMS 4 个不同的标签之一。L(left)代表词的左边界，R(right)代表词的右边界，M(middle)代表词的中间部分，S(single)代表单字成词。经过标注，分词问题就转化为序列标注问题。例如：

位于/ XX/ 海域/ 活动/ 的/ 潜艇/。

L R　L R　L R　L R　S　L R　S

标注 LRMS 4 个标签之一也等价于分类问题，可引入特征选择方法来评价特征，从而选取最有效特征来提升效率。

1.3 中文分词中模型使用的特征

中文分词常用特征分为 unigram 特征及 bigram 特征 2 类，用特征模板表示，见表 1 第 1 列，其中 U00 和 U01 是特征序号，%x[0,0]指的是当前字(unigram)，%x[1,0]指下 1 个字(unigram)，%x[–1,0]/%x[0,0]指的是前 1 个字和当前字组成的二元组(bigram)。特征模板每 1 行代表一类特征。观察位置遍历整个句子时即可产生这句话的所有特征，这些特征中所包含的信息量是不同的，像(“艇。”，S)对分词的贡献就较小，因为“。”大多数情况下都是作为单字词 S，无用特征会导致特征空间膨胀。

表 1 PKU05 上不同种类特征的分词结果

Table1 Results for different features on PKU05

feature template	F1-value
U00:%x[–2,0]	0.662
U01:%x[–1,0]	0.743
U02:%x[0,0]	0.842
U03:%x[1,0]	0.760
U04:%x[2,0]	0.667
U05:%x[–2,0]/%x[–1,0]	0.816
U06:%x[–1,0]/%x[0,0]	0.933
U07:%x[0,0]/%x[1,0]	0.935
U08:%x[1,0]/%x[2,0]	0.831
U09:%x[–1,0]/%x[1,0]	0.851
U10:%x[0,1]	0.544
U02+U06	0.951
all features	0.963
U01+U02+U03+U06+U07+U09	0.965

本文分别对各个模板所产生的特征进行分词试验，结果见表 1。可以看出，unigram 特征中最有效的是 U02，也就是待分类的位置上的字本身，二元组特征中最有效的是 U06 和 U07，是该位置向前 1 个字形成的二元组和向后 1 个字形成的二元组。U02+U06 组成的特征模板所能达到的性能和全部特征模板所能达到的性能相差无几，但特征数量却差好几个数量级。可见不同的特征对分词的贡献是不同的。只有少部分特征起到了重要作用，一部分特征甚至有不利影响，而且特征多会导致训练时间及空间呈指数关系增长。

1.4 特征选择算法

由此本文引入特征选择算法来优化特征区间，此处特征选择[9]是指选出分词能力强的特征组合。试验证明冗余的特征会对效果产生干扰，此处根据标准选出缩减的特征子集使得分词任务达到和特征选择前近似甚至更好的效果。通过选择算法删除无关或冗余的干扰特征，降低训练复杂性，简化的数据集会得到更精确的模型。常用的特征选择算法包括分支定界法、聚类算法等。在文本分类领域[10]，$Chi^2_{\max}$ 是比较常用的特征选择算法之一，在多个数据集上有较好的分类效果。此处引用此算法，衡量特征单元 t 和类别 c 的独立程度，如果 1 个特征和某个类

别非常相关，那么该特征在分词过程中的贡献比较大。

算法流程为：首先使用之前选定的特征模板产生特征全集，接着对全集中的每个特征，针对每个类别计算特征选择算子，这样对于每个特征 t 可计算出其与任意一个类别的关系权重，根据 $Chi^2_{\max}$ 算法要求，计算方式如下：

$$Chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{2}$$

式中：A 为该特征指向特定类别的次数；B 为该特征指向其他类别的次数；C 为特定类别中该特征没有出现的次数；D 为其他类别中不出现该特征的次数；N 为以上 4 个变量的总和。针对每个特征单元可计算出其与所有类别的关系度，选取其最大值代表该特征单元的重要程度：

$$Chi^2_{\max}(t)=\max_{1\le i\le m} Chi^2(t,c_i) \tag{3}$$

式中：m 为类别数。

在用 PKU05 数据集训练出 1 个模型后，本文计算了所有特征的 $Chi^2_{\max}$，并按照这个值进行排序，统计发现只有很少一部分特征具有较大的 $Chi^2_{\max}$，大部分的都很小。为了验证特征的有效性，选取具有较大 $Chi^2_{\max}$ 的特征进行实验，全部特征都生成共有 5 504 732 维，见表 2。由此可知特征选择算法是有效的，只需使用 3 000 000 维即可达到最高分词性能，只用了原有特征集合的一半，很大程度上降低了算法的时间复杂度和空间复杂度。当维数再取高时性能反而下降了，这说明增加的这些特征起到了干扰作用。使用特征选择算法可以在不损失分词性能的前提下提高分词的效率。

表 2 PKU05 不同特征数量下的性能

Table2 Performances with different features on PKU05

No. of features	F-1
500 000	0.933
1 000 000	0.951
2 000 000	0.964
3 000 000	0.965
4 000 000	0.963
5 504 732	0.963

1.5 基于置信度的分词后处理

为进一步提高准确率[11]，本文引入 CRFs 置信度(CRFs confidence)来对模型的分词结果进行后处理，对于 CRFs 给出的每个切分，可以计算置信度[12]。置信度本质是 1 个边缘概率，指对 1 个特定位置的切分的可能性。首先通过标准的前向—后向算法计算得到整个序列的似然度，再通过受限的前向—后向算法算出指定切分时整个序列的似然度，二者比值即为该切分的置信度。

通过在 PKU05 上进行统计发现，CRFs 置信度和切分错误率呈反相关，即低置信度区间中出现切分错误的可能性更大。这样只需对置信度低于设定阈值的片段进行处理，即可对分词错误进行有效修正。

后处理流程为：首先用 CRFs 切分得到置信度，然后按照设定的阈值，筛选出低置信候选片段，根据片段长度的不同将其分类处理：

a) 长度为 2 个字的低置信片段：这种类型的错误，一是可能将二字词错切为 2 个单字，二是可能将 2 个单字合并成 1 个词。在训练集词表中搜索，若训练集中有该二字词，则将其合并为二字词；若没有该二字词则将其标注为错划的二元组；

b) 长度为 3 个字的低置信片段：直接用词表重新切分处理，若结果与原切分结果不同，则进行修正，否则保留原分词结果；

c) 长度为 4 个及 4 个字以上的低置信片段：只有在词表中有整个片段的成词时，才对之前的结果进行修复。

2 实验

实验采用 Bakeoff2005 的 PKU 语料库，抽取 70%作为训练集，30%作为测试集，后处理词表为 Bakeoff 2005 中 PKU 训练集的词表。实验采用 CRF++工具包，标注集使用 LMRS，特征选用 $Chi^2_{\max}$ 排序后的前一半，在取低置信区间时设定 CRFs 置信度阈值为 70%，这个阈值是多次实验得出的可获得最大分词性能的阈值。

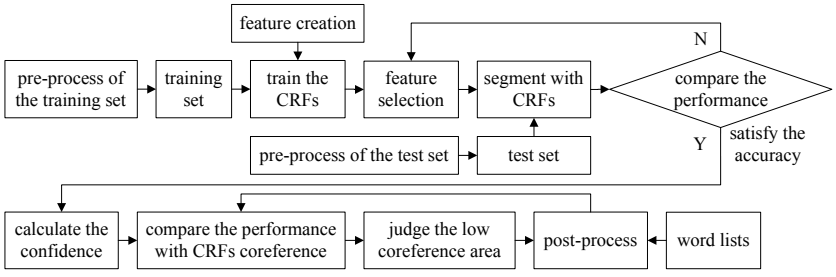

Fig.1 Flow chart of the experiment

图 1 实验流程图

实验流程见图 1。表 3 比较了经特征优化并后处理前后的分词效率。实验证明后处理算法可以有效提高分词的正确率，分词的召回率和准确率同时得到提升。

表 3 后处理算法与原算法性能比较

Table3 Performances without and with the post-process progress

PKU	precision	recall	F-1
CRFs only	0.965	0.961	0.963
CRFs + post-process	0.969	0.967	0.968

3　结论

本文针对当前分词算法中存在的问题，分析各种特征，通过特征选择算法选取分词能力强的特征，构造有效的特征子空间，有效降低分词的时间与空间开销。然后利用 CRFs 模型的置信度，对分词结果进行后处理，进一步提高了分词精确度。实验证明该改进可以有效提高分词正确率，同时节省时间，分词性能得到综合提升。后期将在作战命令分词中进行该算法的实际应用。

参考文献:

[1] 姜文志,蒋伟俊,范洪达. 汉语分词技术在信息工程中的应用[J]. 信息与电子工程, 2007,5(5):385-387.

[2] 周文帅,冯速. 汉语分词技术研究现状与应用展望[J]. 山西师范大学学报(自然科学版), 2006,20(1):25-29.

[3] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报, 2007,21(3):8-19.

[4] 岳中原. 词典与统计相结合的中文分词的研究[D]. 武汉:武汉理工大学, 2010:23-36.

[5] 崔和,龙玉峰. 支持向量机学习算法的研究现状与展望[J]. 信息与电子工程, 2008,6(5):328-332.

[6] 刘群. 汉英机器翻译若干关键技术研究[M]. 北京:清华大学出版社, 2008.

[7] Charles Sutton,Andrew McCallum. An Introduction to Conditional Random Fields[R]. Foundations and Trends in Machine Learning, 2010:18-26.

[8] 韩雪冬. 基于 CRFs 的中文分词算法研究与实现[D]. 北京:北京邮电大学, 2010:14-24.

[9] 吕峰,高春林. 支持向量机在皮肤症状图像识别中的应用研究[J]. 计算机仿真, 2010,27(11):267-269,362.

[10] 巩知乐,张德贤,胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真, 2009,26(7):164-167.

[11] 李丽双,黄德根,陈春荣,等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报, 2006,20(5):51-57.

[12] John D Lafferty,Andrew McCallum,Fernando C N Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of ICML-2001. San Francisco,CA,USA:Morgan Kaufmann Publishers Inc. 2001:282-289.

作者简介:

顾佼佼(1986-)，男，山东省青岛市人，在读硕士研究生，主要研究领域为武器装备信息化.email:vxgu86@hotmail.com.

杨志宏(1980-)，男，湖南省益阳市人，本科，工程师，主要研究领域为导弹发射.

姜文志(1964-)，男，山东省莱州市人，博士生导师，主要研究领域为计算机科学与技术及其现代化.

胡文萱(1986-)，女，山东省烟台市人，本科，主要研究领域为汉英翻译.