针对SVM 中文分词特性的个性化后处理设计

王屹林，朱慕华，朱靖波

（东北大学自然语言处理实验室，辽宁 110004）

摘 要： 支持向量机是当前经常被使用的分类模型。本文使用支持向量机处理中文分词任务，并且在支持向量机的初步结果上，根据其分词特性设计实现了个性化后处理规则。实验结果表明，支持向量机与后处理规则结合而实现的分词系统，在Sighan 评测的数据上达到了良好的效果。

关键词： 中文分词；支持向量机；后处理规则

Specified Multiple PostProcess for Support Vector Machines Chinese Word Segmenter

Wang Yilin, Zhu Muhua, Zhu Jingbo

(Natural Language Processing Lab of Northeastern University, Shenyang 110004, China)

Abstract: Support vector machine is a state-of-art model in most classification task. We present a Chinese word segmentation system based on support vector machine that provides a frmawork to use a large number of linguistic features . Our system performs well especially followed with specified multiple postprocess rules.

key words: Chinese word segmentation; support vector machines; postprocess rule

1. 引言

中文自然语言处理是让机器理解中文的核心技术。其中，词是作为携带语义信息的最小单位。如何从连续的字符序列中确定词与词的边界，成为中文预处理的关键技术。

中文处理中确定词边界的任务称为中文分词。中文分词的难点在于非词典词的识别和词边界的消歧。同时，对于词的定义，虽然语言学家们提出了各种语言学的标准，但是这些标准并不通用于计算语言学的各种不同任务；目前计算语言学领域，普通接受以标注语料和详尽的分词规范作为词的定义。另一方面，语料的标注需要大量的人工参与。如何利用尽可能少的标注数据得到良好的分词性能已成为中文分词研究的热点。研究人员设计了多种不同中文分词技术，其中，常用技术包括基于词典的最大匹配，基于字的标记技术。

支持向量机作为模式识别领域应用广泛的分类模型，同样被用于处理各种自然语言处理任务。本文将中文分词转化为以字为单位的分类问题，提出了基于支持向量机的分词技术；同时，在支持向量机的初步分词结果上，针对其分词特性，设计了与支持向量机起到互补作用的个性化后处理规则。在Sighan2006 数据集进行的实验结果表明，利用支持向量机与个性化后处理规则相结合而实现的分词系统，同样可以取得不错的性能。

本文剩余内容包括：第二部分概要介绍了与本文工作相关的一些研究；第三部分是基于支持向量机的分词系

本文工作部分得到国家自然科学基金(NO.60473140)和国家教育部新世纪优秀人才计划项目资助

王屹林 男 1981 年生 辽宁 硕士研究生 E-mail: wangyl@ics.neu.edu.cn

统以及对后处理规则的详细描述；最后的第四部分是分词结果以及第五部分的结论和未来工作。

2 相关工作

[Xue] 利用最大熵马尔可夫模型（Maximum Entropy Markov Model ）实现的中文分词系统中，任何字都被赋予四个类别之一，包括LM（词的左端）、MM（词的中间）、MR（词的右端）和LR（单字成词）。这些类别用来标记字在词中的位置。其中使用的主要特征是基于字的n-gram 信息，即当前字的上下文窗口信息。为了解决最大熵马尔可夫模型产生的序列标记偏置问题，作者分别从正向和逆向进行分词，并使用基于转换的学习算法整合两种分词结果。该文首次提出以分类技术处理分词问题，但作者并没有深入研究决定分词性能的特征选择问题。

[Ng] 以最大熵模型实现了分词系统。作者针对分词任务，设计了简单但是实用的特征模板。在五个字的上下文窗口中，利用五个特征模板描述局部上下文。实际结果表明，采用该特征模板训练得到的分词系统，可以获得很好的分词结果。在Sighan2005，[Ng] 在开放性测试四个数据集上，获得了三个最佳性能。

3 支持向量机与后处理规则相结合的中文分词系统

本文使用支持向量机与后处理规则相结合的方法实现分词系统。主要包括两大部分：基于支持向量机的初步的分词系统；后处理系统，包含针对支持向量机分词结果的特性而设计的个性化后处理规则。

3.1 基于支持向量机的分词系统

3.1.1 支持向量机模型

支持向量机模型[Vapnik]以统计学习理论为基础，广泛应用于自然语言处理的各项任务中。在满足数据线性可分的约束条件下，支持向量机寻求最优的线性分类超平面，以最大边界正确分类训练数据。支持向量机的成功主要归因于错误泛化能力以及与核函数的结合。支持向量机的学习过程可以等价转换成如下的二次优化问题：

$$W(\partial)=\sum_{i=1}^{l}\partial_i-\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j \partial_i \partial_j \left\langle x_i \bullet x_j \right\rangle$$

约束条件：

$$\sum_{i=1}^{l}y_i\partial_i=0 \quad \Big| \quad \partial_i \geq 0, i=1,\cdots,l. \quad \Big|$$

3.1.2. 支持向量机处理分类问题

中文分词问题可以看作是基于字的分类问题。本文以支持向量机实现分词系统，任何一个字可以属于词首，词尾，词中以及单字成词四个类别中的一类且只属于一类；同时，本文采用[引文,NG]所采用的基本特征模板中的四个，分别为：

(a) C_n (n=-2, -1, 0, 1, 2)

(b) C_nC_{n+1} (n=-2 -1, 0, 1)

(c) $P_n(C_0)$

(d) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

本文没有使用特征模板$C_{-1}C_1$，因为在实际使用中，我们发现这个特征起到了负面的作用。

3.2 针对SVM 分词特性的多重后处理规则

在确定采用支持向量机以及特征模板的前提条件下，其分词的结果具有一定的规律性。同时，我们发现，在引入其它特征以后，往往使词典词召回率和非词典词召回率发生振荡而整体的性能并没有发生改变。因此，本文在使用支持向量机得到初步分词结果以后，设计了与支持向量机分词特性相关的个性化后处理规则，改善系统的整体性能。

首先我们将训练语料中的出现的词称为词典词（IV，in-vocabulary）。相对应的，测试集中出现但未在训练语料中出现的词称为非词典词（OOV，out-of-vocabulary）。并且，部分词典词在不同的上下文环境下，有成词和切分两种情况。例如，“充满希望的新世纪”和“新世纪大酒店”两个片断。前者“新世纪”被切分成两个词；后一种情况，“新世纪”作为命名实体的一部分，组合成词。因此，我们把词典词还可以分为“切分一致词”和“切分不一致词”（如，“新世纪”）。

最终得到的后处理规则主要可以分为两部分：规则使切分结果与训练语料尽可能一致，我们称之为“词典词规则”；规则根据当前的切分结果，对非词典词进行猜测，我们称之为“非词典词规则”。

3.2.1. 切分一致的词典词被错误切分

训练数据中，各个类别的数据分布并不均衡，其中，以“词中”类别包含的数据最少，其它三类数据量相当。这种数据分布的性质，使支持向量机在分类置信度较低的情况下，倾向于选择两个字或者单字成词的情况。例如，训练语料中的词“复制品”属于一个整词，但在测试语料中，“复制品”被SVM 切分为“复 制品”；“统一”是训练语料中切分一致的词，支持向量机却将它切分成“统 一”。对于如上在训练数据中出现，且切分情况与上下文无关的词典词，我们设计规则，保证这类词在测试语料中的切分与训练语料中相一致。

3.2.2. 多个切分一致的词典词被错误合并

当切分片断包含有后缀词，例如“（淘金）者”，“（经理）人”，支持向量机在分类置信度较低的情况下倾向于将后缀与前一词进行合并。比如，在支持向量机的分词结果中，训练语料中的切分片断“符合 条件者”被SVM合并为：“符合 条件者”。

我们可以对SVM 切分结果中的非单字片断W 进行处理。如果W 不是个有不同切分的词，如“新世纪”， 则可以使用最大匹配法对词W 进行切分，得到切分片断S1，S2，..，Sn。如果该切分片断在训练语料中至少出现一次，后处理规则用该片断代替词W，作为最终的切分。

3.2.3. 切分不一致词的处理

某些字符片断，如“新世纪”，是否切分依赖于其所在的上下文环境。如果测试数据中包含该片断上下文未在训练数据中出现，支持向量机倾向于将字符片断合并成词。后处理规则利用训练数据中包含该片断的上下文，来选择是否切分。

我们从训练语料中，寻找切分不一致词的所有切分情况，并收集其上下文。出现在片断前与片断后的词分别形成前导词和后导词的集合。比如，词“新世纪”有切分“新世纪”和“新 世纪”，后处理规则取得切分“新世纪”的前导词和后导词的集合，以及“新 世纪”的相应集合。

在支持向量机的分词结果中，如果某个词W 是切分不一致的词（可能是个整词或被切开），规则就可以根据上下文信息来选择更适合的切分。在本文实验中，只考虑切分片断的前一个词和后一个词。

3.2.4. 非词典词被错误切分的情况

支持向量机倾向于将多于等于三个字的非词典词切分为连续的两个字的片断或连续两个字的片断后跟着一个单字。如，分词结果中会有切分片断“巴拿 马籍”，就是词典词“巴拿马”被错误切分后与周围窗口的词错误绑定后的结果。这样会造成连续的非词典词片断的存在。我们设计两条后处理规则，用于部分解决非词典词被错误切分的问题：对分词结果文件中连续两个非词典词字符片断，将其合并，如果合并后的片段中含有词典词，则词典词与其余部分切分；连续两个非词典词字符片断，如果其中至少有一个是单字的情况，则进行合并（考虑到大部分汉字都具有独立成词的能力，且训练数据的规模足以包含大部分单字成词的情况）。

3.3 语料库构建过程中的错误切分

手工标注的语料库不可避免，存在一部分切分错误。本文考虑到语料库中存在标注错误，对标注错误与切分不一致的情况进行了区分。如果词典词存在不一致的切分情况，则不同的切分至少出现多次；如果词典词的不同切分的频次比值大于某个阈值，且其中一种切分只在训练语料中出现一次则认为这种切分属于标注错误。在本文实验，阈值选择为7，后处理规则将排除这种切分情况。

4 实验

SigHan Word Segmentation BakeOff 是针对中文分词而提供的一个公共、开放的评测平台。它提供了统一、公开的数据集。本文的系统是在Sighan 中文分词评测所提供的数据上衡量系统的性能的。

4.1 性能指标

本文实验的评测与Sighan 分词评测的指标一致，使用传统的召回率、正确率、宏F1 来评价分类结果。计算公式如下：

正确率P = Count_ws_correct / Count_ws_sum

召汇率R = Count_ws_correct / Count_test

宏F1 = P*R*2 / P+R

其中， Count_ws_correct 代表分词系统正确识别的词数；

Count_ws_sum 代表分词系统识别的词的总数；

Count_test 代表测试语料中词的总数

4.2 实验语料及结果

本中文分词系统参加了SigHan2006 Word Segmentation BakeOff 中四个数据集的closed track 。实验中的训练语料和测试语料皆为SigHan 所提供。在四个数据集上的实验结果如图：

表4.1 分词系统评测结果

Tab4.1 Result in SigHan 2006 Word Segmentation Bakeoff

	准确率	召回率	OOV 召回率	IV 召回率	F-measure
UPenn	94%	91.4%	0.634	0.969	92.7%
Msra	95.5%	95.6%	0.650	0.966	95.6%
Cityu	97.1%	96.5%	0.719	0.981	96.8%
Ckip	94.9%	94%	0.694	0.960	94.4%

从评测结果可以看到，我们的分词系统在词典词的识别方面表现具有可比性，具有切分歧义的词几乎都能够被正确切分。错分的词典词大部分是那些切分不一致的词。所以针对切分不一致词的后处理规则仍需要更细化的设计；在未登陆词的识别方面，系统的性能一般。虽然对于人名、地名等专名，系统都能够很好的识别。但系统所加入的特征，对于更普通的未登陆词的识别的帮助还是比较有限。

5 结论和未来的工作

本文主要介绍了以支持向量机基础而实现的中文分词系统，并且针对支持向量机分词结果中的一些性质设计的多重个性化后处理规则。在Sighan 分词评测数据集上的实验表明，基于统计的分词系统与基于规则的后处理系统具有互补性，能够使词典词召回率和非词典词召回率同时得到提高，从而使分词系统整体性能得到改善。另一方面，规则往往只能处理复杂问题中的部分子问题，而且不同规则之间容易产生冲突，从而对系统性能造成损害。更恰当的方法，可以将已经得到验证的有效的后处理规则，转为学习算法的特征，自动地从数据中进行学习。

这将是我们下一步的工作。

参考文献

[1] Xue Nianwen. Chinese Word Segmentation as LMR Tagging[A], Computational Linguistics 2003

[2] Ng Hwee Tou. A Maximum Entropy Approach to Chinese Word Segmentation[A] , SigHan2005

[3] Tseng Huihsin. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005[A] , SigHan2005

[4] Emerson Thomas. The Second International Chinese Word Segmentation Bakeoff [A], SigHan2005

[5] Brill Eric. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging[A], Computational Linguistics 21:543-565

[6] Gao Jianfeng. Chinese Word Segmentation: A Pragmatic Approach[A] , MSR-TR-2004

[7] Vapnik N. 统计学习理论本质[M]. 清华大学出版社，2000

[8] 姚天顺等, 自然语言理解——让机器懂得人类语言的研究[M]. 第二版，清华大学出版社，2002