第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)03—0192—02 文献标识码：A 中图分类号：TP39

基于前后文 n-gram 模型的古汉语句子切分

陈天莹[1]，陈 蓉[1]，潘璐璐[1]，李红军[1,2]，于中华[1]
(1. 四川大学计算机学院，成都 610064；2. 西南科技大学计算机学院，绵阳 621002)

摘 要：提出了基于前后文 n-gram 模型的古汉语句子切分算法，该算法能够在数据稀疏的情况下，通过收集上下文信息，对切分位置进行比较准确的预测，从而较好地处理小规模训练语料的情况，降低数据稀疏对切分准确率的影响。采用《论语》对所提出的算法进行了句子切分实验，达到了 81%的召回率和 52%的准确率。

关键词：n-gram 模型；数据稀疏；平滑技术；基于前后文的 n-gram 模型

Archaic Chinese Punctuating Sentences Based on Context N-gram Model

CHEN Tianying [1], CHEN Rong[1], PAN Lulu[1], LI Hongjun[1,2], YU Zhonghua[1]
(1. Dept. of Computer Science, Sichuan University, Chengdu 610064;
2. Dept. of Computer Science, Southwest University of Science and Technology, Mianyang 621002)

【Abstract】 An algorithm of punctuating the sentences in archaic Chinese language based on context n-gram model is proposed in the paper. The algorithm can make comparatively accurate prediction of the punctuating-positions of the text under data-sparse instances by collecting and calculating context information to better analyze small-scaled corpus and meanwhile, to bring down the effects of the data-sparse plight on the global accuracy. At last, the paper selects the analects of Confucius (Lunyu) to test the algorithm introduced, and the results show that the recall and the precision achieve 81% and 52% respectively.

【Key words】 N-gram model; Data sparse; Smoothing technology; N-gram model based on context

中国几千年的文明历史形成了浩如烟海的历史典籍。如何借助现代化的手段对这些历史典籍进行有效的挖掘，对于继承发展我国古代灿烂的历史文化具有重要意义。作为古代书面语的重要形式，古汉语的分析理解是历史典籍挖掘的关键和基础。本文针对古汉语句子中缺少句读的问题，研究并提出了句子自动切分(自动加句读)的算法，并在《论语》上验证了算法的有效性。

1 问题特点、难点及相关工作

本文要解决的问题是设计算法对无标点的古汉语文本进行句子切分，确定句读的位置，包括逗号、冒号、感叹号、问号、句号和顿号等。很显然，上述问题与句子边界识别完全不同，句子边界识别是识别作为句子边界的句号，其实质是对自然语言文本中出现的句号(如汉语的“。”和英语的“.”)根据前后文进行消歧[1,2]。

古汉语句子切分问题不是一个平凡问题，其难点主要体现在以下几个方面：

(1)古汉语具有句子简洁精练的特点，单字成词的现象比较普遍。

如：

1)子曰：“邦有道，谷；邦无道，谷，耻也。”

2)“克、伐、怨、欲不行焉，可以为仁矣？”

3)子曰：“先之，劳之。”请益。曰：“无倦。”

古汉语的简洁精炼使预测句读位置所依赖的局部前后文信息变少，这增加了句子切分的难度。

(2)古汉语文体繁杂，如《论语》、《道德经》、《韩非子》、《诗经》等，每一种文体都有自己独特的风格，且具有的文本数量少，很难获取训练统计模型所需要的足够样本，因此，对于古汉语句子切分来说，数据稀疏问题更加严重。

(3)古汉语中字和词的界限模糊，很难进行词的切分，无法利用单词一级的特征进行句读位置的预测，只能利用有关字或者字串方面的信息来进行决策。

目前，人们在英语和现代汉语句子边界识别方面进行了大量的研究工作，提出了一系列基于规则和基于统计的识别算法，达到了 99%左右的准确率。但是对于古汉语句子的自动切分，还未见相关的研究报告。本文设计并实现了一个基于前后文 n-gram 模型的古汉语句子切分算法，提出了有效的解决数据稀疏问题的平滑技术，在《论语》上的实验结果达到了 81%的召回率和 52%的准确率。

2 基于前后文n-gram模型的古汉语句子切分算法

n-gram(n元语法)是自然语言统计建模的重要工具，在单词和字母预测等方面获得了广泛的应用。但是，传统的n-gram模型[5]在应用于古汉语句子切分时会面临数据稀疏的严重问题。针对古汉语句子切分所面临的数据严重稀疏的问题，本文设计并实现了一个基于前后文n-gram模型的切分算法，实验结果表明，新算法不但优于经典的n-gram模型，而且优于

基金项目：国家自然科学基金资助项目(60073046)；高等学校博士学科点专项科研基金“SRFDP”资助项目(20020610007)
作者简介：陈天莹(1982－)，女，硕士生，主研方向：自然语言处理，自动推理及智能软件设计；陈 蓉，工程师；潘璐璐，硕士生；李红军，在职硕士生、讲师；于中华，博士、副教授
收稿日期：2006-03-13 **E-mail：**yuzhonghua@cs.scu.edu.cn

现有的针对数据稀疏问题提出的各种平滑技术。

2.1 算法的基本思想

设待切分的古汉语文本串为 $w_1w_2...w_n$，其中 w_i 为任意的汉字。对于任意相邻的两个汉字 w_iw_{i+1}，算法通过计算它们之间有句读的可能性度量 $\sigma(w_i \bullet w_{i+1})$ 和没有句读的概率 $\sigma(w_iw_{i+1})$，然后比较 $\sigma(w_i \bullet w_{i+1})$ 和 $\sigma(w_iw_{i+1})$ 的大小关系来决定是否应该在 w_i 和 w_{i+1} 之间加上句读。当 $C(w_i \bullet w_{i+1})>0$ 且 $C(w_iw_{i+1})>0$ 时，$\sigma(w_i \bullet w_{i+1})$ 和 $\sigma(w_iw_{i+1})$ 分别简单地定义为 $C(w_i \bullet w_{i+1})$ 和 $C(w_iw_{i+1})$，否则定义为

$$\sigma(w_i \bullet w_{i+1}) = C(w_i\bullet)/C(w_i)+C(\bullet w_{i+1})/C(w_{i+1}) \quad (1)$$

$$\sigma(w_iw_{i+1}) = [C(w_i+)/C(w_i)+C(+w_{i+1})/C(w_{i+1})]\times d \quad (2)$$

其中 $C(w_i \bullet w_{i+1})$ 为训练语料中 w_i 和 w_{i+1} 之间存在句读的次数，$C(w_iw_{i+1})$ 为 w_i 和 w_{i+1} 之间不存在句读的次数，$C(w_i+)$ 为 w_i 后面不为句读的所有二元组出现的次数，$C(+w_{i+1})$ 为 w_{i+1} 前面不为句读的所有二元组出现的次数，$C(w_i\bullet)$ 为 w_i 后面接句读的二元组出现的次数，$C(\bullet w_{i+1})$ 是句读后接 w_{i+1} 的二元组出现的次数，d 为折扣因子。引入折扣因子的目的是为了使 $\sigma(w_i \bullet w_{i+1})$ 和 $\sigma(w_iw_{i+1})$ 具有可比性，由于在一般情况下，$C(w_i+)>C(w_i\bullet)$，$C(+w_{i+1})>C(\bullet w_{i+1})$，因此 $C(w_i\bullet)/C(w_i)+C(\bullet w_{i+1})/C(w_{i+1})<C(w_i+)/C(w_i)+C(+w_{i+1})/C(w_{i+1})$，d 的取值应该反映汉字和句读在语料库中分布的差异，即大约为 1/L，其中 L 为训练语料中由句读隔开的汉字串的平均长度(汉字数)。当 $\sigma(w_i \bullet w_{i+1})>\sigma(w_iw_{i+1})$ 时，确定 w_i 和 w_{i+1} 之间有句读，否则没有句读。在 $C(w_i)=0$ 或 $C(w_{i+1})=0$ 时，简单地认定 w_i 和 w_{i+1} 之间没有句读。

针对语料库规模小，数据严重稀疏的问题，式(1)和式(2)式在三元组频率为 0 时，将三元组频率的计算退化成二元组频率的计算，这样可以大大降低数据稀疏为决策带来的负面影响。从式(1)、式(2)可以看出，在识别句子的切分位置时，不仅利用了句读可能出现位置的前邻接汉字，而且还利用后面临接的汉字信息，综合这种前后文信息来决策句读的位置。

2.2 算法描述

图 1 给出了基于上述思想设计的古汉语句子切分算法。

```
IF      C(w_i) = 0 OR C(w_{i+1}) = 0
THEN    设定 w_i 和 w_{i+1} 之间没有句读；
ELSE
BEGIN
    IF  C(w_i • w_{i+1}) ≠ 0  AND  C(w_i w_{i+1}) ≠ 0  THEN
    BEGIN
        σ(w_i • w_{i+1}) := C(w_i • w_{i+1});
        σ(w_i w_{i+1}) := C(w_i w_{i+1});
    END
    ELSE
    BEGIN
        σ(w_i • w_{i+1}) := C(w_i•)/C(w_i) + C(•w_{i+1})/C(w_{i+1});
    σ(w_i w_{i+1}) := [C(w_i+)/C(w_i) + C(+w_{i+1})/C(w_{i+1})] × d;
    END
    IF  σ(w_i w_{i+1}) > σ(w_i • w_{i+1})  THEN
          设定 w_i 和 w_{i+1} 之间没有句读；
    ELSE 设定 w_i 和 w_{i+1} 之间有句读；
END
继续处理下面两个汉字;
```

图 1 基于前后文 n-gram 模型的古汉语句子切分

很明显，算法采用的前后文 n-gram 模型是传统 n-gram 模型的变形和改进，以便充分利用前后文信息对句读的位置进行预测，并克服数据稀疏带来的问题。

3 实验结果及分析

为了验证算法的有效性，选择《论语》进行了实验。《论语》按照 80%对 20%被随机划分成两个集合，分别用作训练语料库和测试语料库。表 1 最后一行给出了本文所提算法在测试语料库上进行句子切分的召回率、准确率和F度量值(F-measure)，其中折扣率选定d=0.25，其它行是传统数据平滑技术[3~6]利用相同训练语料和测试语料得到的结果。底线模型是认为每个汉字后面都有句读。

表 1 本文算法与各种平滑技术的实验对比

算法	平均召回率(%)	平均准确率(%)	F 度量
底线(baseline)模型	100	23.598 7	38.186
MLE	97.266 67	32.834 85	49.096
Laplace 平滑技术	32.124 705	91.259 325	47.521 2
Lidstone 平滑技术	43.236 102 5	75.007 977 5	54.853 5
Jelinek-Mercer 平滑技术(λ=0.5)	83.491 822 5	41.785 8	55.696 66
Jelinek-Mercer 平滑技术(λ=0.16)	76.768 05	49.399 782 5	60.115 56
Witten-Bell 平滑技术	59.654 7	51.817 045	55.460 34
Absolute-Discounting 平滑技术	85.044 807 5	40.933 125	55.266
基于前后文 n-gram 模型	81.024 15	52.595 42	63.785 55

从表 1 可以看出，基于前后文 n-gram 模型的古汉语句子切分算法在 F 度量方面明显优于其它数据平滑技术，能够更好地解决古汉语句子切分所面临的数据严重稀疏的问题。图 2 给出了所实现的系统原型的运行实例，其中左边文本为系统自动切分的结果，右边为人工断句的结果。

图 2 原型系统运行实例

4 总结

本文针对传统 n-gram 模型及各种平滑技术在识别古汉语句读位置方面存在的缺点和不足，提出了基于前后文 n-gram 模型的古汉语句子切分算法，该算法能够在数据稀疏的情况下，通过收集上下文信息，对切分位置进行比较准确的预测，从而较好地处理小规模训练语料的情况，降低数据稀疏对切分准确率的影响。采用《论语》对所提出的算法进行了验证，达到了 81%的召回率和 52%的准确率。

参考文献

1 Palmer, David D, Hearst, at al. Adaptive Multilingual Sentence Boundary Disambiguation[J]. Computational Linguistics, 1997,23(2).

2 Charoenpornsawat P, Sornlertlamvanich V. Automatic Sentence Break Disambiguation for Thai[C]//Proceedings of ICCPOL'01, Korea, 2001: 231-235.

（下转第 196 页）