

文章编号:1001-9081(2005)09-2025-03

一种基于提取上下文信息的分词算法

曾华琳,李堂秋,史晓东
(厦门大学 计算机科学系,福建 厦门 361005)
(hlzeng@xmu.edu.cn)

摘 要:汉语分词在汉语文本处理过程中是一个特殊而重要的组成部分。传统的基于词典的分词算法存在很大的缺陷,无法对未登录词进行很好的处理。基于概率的算法只考虑了训练集语料的概率模型,对于不同领域的文本的处理不尽如人意。文章提出一种基于上下文信息提取的概率分词算法,能够将切分文本的上下文信息加入到分词概率模型中,以指导文本的切分。这种切分算法结合经典 n 元模型以及 EM 算法,在封闭和开放测试环境中分别取得了比较好的效果。

关键词:中文分词; n 元模型;上下文信息

中图分类号: TP391.2 **文献标识码:** A

Segmentation algorithm for Chinese based on extraction of context information

ZENG Hua-lin, LI Tang-qiu, SHI Xiao-dong

(Department of Computer Science, Xiamen University, Xiamen Fujian 361005, China)

Abstract: Chinese segmentation is a special and important issue in Chinese texts processing. The traditional segmentation methods based on an existing dictionary have an obvious defect when they are used to segment texts which may contain words unknown to the dictionary. And the probabilistic methods those consider the probabilistic model of the training set only also do a bad job on the texts of a specific domain. In this paper, a probabilistic segmentation method based on extracting context information was proposed, which adds the context information of the segmenting text into the segmentation probabilistic model so as to guide the processing. The method combining n -gram model and EM algorithm achieves a good effect in the close and opening test.

Key words: Chinese segmentation; n -gram model; context information

0 引言

汉语与西方文字在书写形式上最大的不同在于,汉语词与词之间没有空格,在整个汉语信息处理过程中,首先必须解决汉语的自动分词问题。词是最小的能独立活动的有意义的语言成分。汉语处理应用系统只要涉及语法语义(如检索、翻译、文摘、校对等),就需要以词为基本单位。因此,汉语自动分词技术在现代汉语信息处理过程中的地位不言而喻,许多的语言研究人员在这个方面做了大量的工作。

在本文中,我们研究总结了前人的部分做法,分析在分词领域中流行的三类算法的不足之处,在结合经典算法(基于词典的算法和基于统计的算法)的基础上,提出一种基于二元 Bigram 模型,加入用 EM 算法训练的上下文信息的改进分词算法,并与几种经典分词算法进行了比较。

1 汉语分词算法

已有的分词算法可以归结为以下三种:

第一种是基于词典(Dictionary-Based)的机械匹配算法^[1],例如正向匹配,逆向匹配,最小匹配等等,这些算法的优点是易于实现,在对精确率要求不高的系统中得到了很好的应用。其缺点在于由于词典是在分词之前准备的,其规模和内容受到了限制,所以没有哪个词典是完备的;语言中常出现新的词语,所以没有一个词典能够囊括所有领域的词语;虽然可以通过加入一些构词规则的方法识别出一些可构造新词,但是基于词典的这一类算法无法解决文本中大量出现的

未登录词(OOV)的问题,致使分词的效果达到一定的瓶颈之后无法提升。这里的未登录词不仅包括了命名实体(人名,地名,组织名,时间词,数词等),也包括了新词。

第二种可以归纳为统计的分词方法^[2,3]。例如 N-Gram 算法,HMM 算法,最大熵算法,基于 EM 的算法等等。统计方法的优点在于它可以从已有的大量实例中进行归纳总结,分析语言内在的关联信息,将其加入到统计模型中去;简单的统计方法不需要词典,而是通过训练语料的迭代,建立统计模型。对统计的方法来说,训练语料库的规模严重影响着分词的效果,训练集规模小则模型的可信度低,分词效果差;而一旦训练集规模大了,则会引起数据稀疏的问题,使得分词的效率大大降低;另一方面,不同领域的语料对于统计模型起着决定性的作用,口语语料跟书面语语料,不同专业领域的语料都在内容上存在着很大的差异,拿书面语语料训练出来的统计模型去切分口语语料,势必不会得到很好的切分结果。

第三种较为成熟的分词算法将统计的方法与词典的方法进行结合,例如中国科学院计算技术研究所的汉语词法分析系统 ICTCLAS 采用的是多层隐马尔可夫模型^[4]。他们对原有隐马模型进行了扩展,将模型分别应用到原子切分、简单和复杂的未登录词识别及基于类的隐马分词等多个层面上。这种分词算法也存在着不足,其上下文信息都是从训练语料库中获取,忽略了切分文本的上下文反馈信息。

关于切分文本的上下文信息对于分词的结果的影响,我们分析下面的两个例子。在这两个例子中,如果用抽取出的切分文本的上下文信息进行分词指导,分词的效果将得到改善。

收稿日期:2005-03-18;修订日期:2005-05-28 基金项目:国家 863 计划资助项目(2002AA117010)

作者简介:曾华琳(1980-),女,福建厦门人,硕士研究生,主要研究方向,机器翻译、计算语言学;李堂秋(1944-),男,福建长乐人,教授,主要研究方向:人工智能、自然语言处理、机器翻译;史晓东(1966-),男,江苏江阴人,教授,主要研究方向:自然语言理解、机器翻译。

第一个例子,对组合歧义问题的帮助。

“王洪超生前使用过的物品”

切分 a: 王洪超/nr 生前/v 使用/v 过/v 的/u 物品/n

切分 b: 王洪/nr 超生/v 前/f 使用/v 过/v 的/u 物品/n

这两种切分方法,从语意上来讲是完全不同的,但从语法上来讲,却都是正确的。如何判断到底哪个切分结果是正确的,单纯从规则或者概率的方法都是无法解决的,但是如果在上下文中我们可以找到“王洪超/nr”的切分词语,则从统计的模型上来衡量,切分 a 的概率值显然是正确的切分结果了。

第二个例子,对于新词发现的帮助。

“厦门大学 bbs 鼓浪听涛”

切分 a: 厦门大学/n bbs/n 鼓/n 浪/n 听/v 涛/n

切分 b: 厦门大学/n bbs/n 鼓浪听涛/n

对于新词的处理,基于文本局部出现的原理,可以在上下文中找到“鼓浪听涛”的多次出现,那么可以通过适当的加重该词的权重比例,使得其在统计的分词过程中被正确地切分出来。

本文正是基于上下文的语境,基于文本的局部出现的特性,将上下文中已经识别出的知识作为既定事实,加入分词系统的迭代过程中,以求取得在切分文本中,前文对后文的指导性效果。

2 引入上下文信息的二元概率模型分词算法

下面具体阐述本文实现的引入局部文本信息的二元概率模型的分词算法。首先以 EM 算法训练切分文本(即文本的局部上下文信息),建立切分文本概率模型,然后以最大概率法进行整体的分词切分。在分词切分过程中,分别采用了字和词的 Bigram,以切分文本作为基准值,大规模语料和分词词典作为调整值,进行加权平均,以确定最后的切分概率值。

2.1 准备语料

a) 大规模文本(未切分文本:读者文摘 200 期语料库,含 9929225 字,约 471760 句,共 23.6M),从中可以抽取字的频率 $\Pr(C_i)$,两个字之间的共现概率 $\Pr(C_i | C_{i-1})$ 。

b) 人工标注语料(包括北京大学计算语言研究所 1998 年 1 月人民日报标注语料,863 评测提供的标准标注数据),抽取字的频率 $\Pr(C_i)$,两个字之间的共现概率 $\Pr(C_i | C_{i-1})$,词的频率 $\Pr(W_i)$,两个词之间的共现概率 $\Pr(W_i | W_{i-1})$ 。进一步通过贝叶斯先验概率公式

$$\Pr(C_i | W_i) = \Pr(C_i, W_i) | \Pr(W_i) \quad (1)$$

得到字的成词概率。这些概率值构成基础语料资料。此处,都只考虑二元概率,暂不考虑三元概率。

c) 一个切分用电子词典,拥有词条 88727 条。一个常用词词典,拥有词条 5000 条。

2.2 基于 EM 算法的切分文本概率模型建立算法

下面提出一种改进的 EM (Expectation Maximization) 算法训练切分文本的概率模型^[5],这是一个基于对分词碎片中的字的出现概率构建汉语分词的模型,过程如下:(这里将要进行切分的文本称为“切分文本”)

1) 扫描文本,对于切分文本中的每个文字序列,以常用词词典给出粗切分结果。

2) 对粗切分后的分词碎片,给出所有可能切分的概率初值。

3) 进行如下的步骤,计算词的概率:

a) 使用当前词的概率值计算每个可能切分的可能性;

b) 对切分可能性进行“归一化”处理为尾数,使其和为

c) 对每种切分进行词计数,即将切分的“尾数”加到词数上。

4) 使用当前的词的概率值,对语料库中的句子进行分词处理。

5) 重复 2)、3) 过程,进行多次迭代,直到概率值收敛。

在以上的算法过程中,需要说明以下几点:

1) 假设词的长度是可预见的,所以取长度为 1、2、3、4 的词作为我们的候选词序列。这样的考虑是基于对切分用词典的考察,不同长度的词的数量分布如表 1。

表 1 不同长度词语数量的统计

词的长度	1	2	3	4	5	6
所占比例	4.24%	38.43%	23.59%	25.87%	4.02%	1.90%

长度小于等于 4 的词条占词条总数的 92.13%,于是在第一步扫描文本过程中,分别建立一字、二字、三字、四字的共现概率词典。考虑到为了尽可能地降低算法复杂性,在切分过程中,只将句子中的连续汉字序列当作一个切分整体,即数字、标点等非汉字字符都作为切分文本的分隔符,同时一些高频率出现的但是成词率(构成词的频率)较低的助词也应当作为切分分隔,这些词可以从在对大规模语料的统计数据中得到,本文中,取以下词:“的、在、了、和、是、有、为、不、这、对”。

2) 切分可能性的计算。对于一个长度为 n 的句子,用全切分的方法,其可能的切分方法有 2^{n-1} 种。我们的切分则是针对经过切分分隔符预切分以后形成的连续汉字序列,所以这样大大降低了切分的可能性。对于一个具有 3 个字的连续汉字序列,其切分可能性为如下几种,以“|”代表切分位置: $C_1 | C_2 | C_3$, $C_1 C_2 | C_3$, $C_1 | C_2 C_3$, $C_1 C_2 C_3$ 。

根据极大似然原则 (MLP),假设词互相独立,即两个词共现概率只与各自的出现概率相关,于是第一种切分的切分可能性为

$$\Pr(C_1 | C_2 | C_3) = \Pr(C_1) * \Pr(C_2) * \Pr(C_3) \quad (2)$$

3) 归一化过程,尾数 α_i 的计算公式如下:

$$\alpha_i = \Pr(\text{切分方法 } i) | \sum \Pr(\text{切分方法 } j) \quad (3)$$

4) 词计数的过程,将尾数 α_i 加入词数上,即此词的出现概率上,则 $\Pr(C_1 | C_2 | C_3)$ 计算出的尾数将被加入到 $\Pr(C_1)$, $\Pr(C_2)$, $\Pr(C_3)$ 上。

5) 迭代的过程可以进行到概率值收敛,为了降低算法复杂性,实际的处理过程进行了 k 次 ($5 \leq k \leq 10$),取经验值 $k = 5$ 。

在 EM 算法结束后,我们得到切分文本的概率词典,其中所有单字的出现频率,双字的,三字的以及四字的,经过阈值的过滤,留下的信息构成了上下文的切分信息,以及对最后分词切分结果的统计得到的 Bigram 数据。

2.3 分词算法模型

1) 根据基于 EM 算法的无词典统计方法建立切分文本的概率模型,收集切分文本的上下文信息。利用建立好的切分文本的概率模型,成立切分文本词典,其中包含切分文本中的词条的概率,并且统计切分文本中词条的 Bigram 数据。

2) 利用 Bigram 模型对文本进行最大概率^[6]切分。评价公式如下:

$$\Pr(W | C) = \prod_i \Pr(w_i) * \Pr(w_i | w_{i-1}) \quad (4)$$

这里通过加权策略进行组合,对已有的概率数据(切分文本的数据,切分用词典的数据,人工标注文本的数据)进行结合:

$$\Pr(w_i) = \sum (wt_{c_j} * \Pr_i(w_i)) \quad (5)$$

$$\Pr(w_i | w_{i-1}) = \sum (wt_{b_j} * \Pr_i(w_i | w_{i-1})) \quad (6)$$

其中 wt_{c_j} 和 wt_{b_j} 为第 j 个概率数据的权重,且满足约束条件

$$\sum wt_j = 1 \quad (7)$$

3) 对切分完的文本进一步进行简单命名实体的识别:

a) 时间词的识别,以“xxxx年xx月xx日”的格式对文本进行扫描。

b) 简单的人名识别,以姓氏触发。

3 实验结果分析

实验评测分别在封闭集和开放集中进行。评测采用 863 评测所用的三个指标:正确率、召回率、F 值。各指标定义如下:

(1) P 正确率 = 识别出的词语出现在标准结果中的词语数 ÷ 识别出的词语总数 × 100%;

(2) R 召回率 = 识别出的词语总数出现在标准结果中的词语数 ÷ 标准结果中的词语总数 × 100%;

(3) F 值 = $2 \times \text{召回率} \times \text{正确率} \div (\text{召回率} + \text{正确率})$ 。

3.1 概率数据加权权值对分词结果的影响

$wt_{c_j}(j=1,2,3)$, $wt_{b_j}(j=1,2,3)$ 分别代表字的概率权重以及双字 Bigram 的概率权重。其中 j 的取值分别代表:切分文本数据,切分用词典的数据以及语料库数据(单字数据取大规模文本的数据,双字数据取人工标注文本的数据)。根据我们对不同概率数据的可信度进行了以下的排序:

切分文本数据 > 切分用词典 > 大规模文本数据(人工标注文本的数据)。

这样排序的原因在于,充分利用了切分文本的局部信息,将其作为指导我们分词过程中的最大可信度标准。由于人工标注文本数据中可能存在着一些与切分用词典重复的词或者是一些与切分词典不兼容的未登录词条,于是,我们将切分用词典的可信度设置得比人工标注文本的数据要高。根据这样的排序,对每一组数据进行了全排列的测试,测试数据以 0.1 进行递增。图 1 是一组测试数据 F 值的结果图。

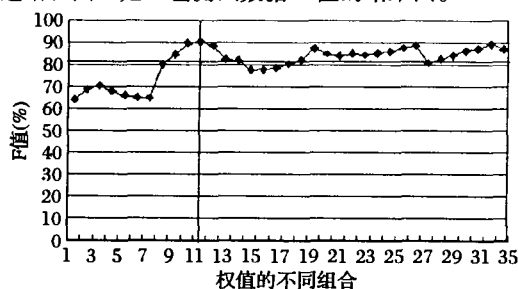


图1 概率数据加权权值对分词结果的影响

上面的数据分别是对测试文本进行全部切分得到的切分数据平均值。从以上的图表中,我们可以看到:平均切分 F 值是 80.97%,最高的切分 F 值为 90.20%,出现在以下概率权值中:

表2 最高 F 值概率的数据加权权值

wt_{c_1}	wt_{c_2}	wt_{c_3}	wt_{b_1}	wt_{b_2}	wt_{b_3}
0.6	0.2	0.2	0.7	0.2	0.1

这样的—个数据结果也从另一个方面验证了我们用“局部数据”指导分词的思路的正确性。

由于测试数据的复杂,只显示了 35 组的数据,在进一步的工作中,我们将采用统计学中的抽样方法对权值进行更精确全面的估值。

3.2 与经典分词算法的比较结果

与本文算法进行评测的对象是正向最大匹配算法,逆向最大匹配算法以及最大概率分词算法。选择的算法是分词领域中最常用的几种分词算法。我们的评测分别在封闭集和开放集进行。封闭集测试语料选择网上公布的北京大学计算语言研究所 1998 年 1 月人民日报标注语料。开放集测试语料选择 863 评测提供的带有标准答案的原始文本 200 个。评测结果取的是所有测试数据的平均值。

表 3 是封闭集测试的数据评测比较结果。可以看到,本文提出的将上下文信息融入分词算法的过程在一定程度上改善了这些经典算法的分词效果。

表3 封闭集常用分词算法结果比较

分词算法	P 正确率 (%)	R 召回率 (%)	$F1$ 值 (%)
正向最大匹配	93.51	89.97	91.71
逆向最大匹配	94.06	90.46	92.22
最大概率分词	90.34	89.98	90.16
本文算法	96.32	91.51	93.85

表 4 是开放集测试的数据比较。与在封闭集测试的结果一样,经典算法的分词效果得到了提高。但总体效果降低,其原因是由于训练语料和测试语料的范围不同造成的。

表4 开放集常用分词算法结果比较

分词算法	P 正确率 (%)	R 召回率 (%)	$F1$ 值 (%)
正向最大匹配	86.76	86.58	86.67
逆向最大匹配	89.05	88.73	88.89
最大概率分词	85.04	87.65	86.33
本文算法	90.39	90.02	90.20

4 结语

本文阐述的算法在分词精确率上对于经典算法来讲有一定的提高,证明了此种算法还是实际可行的,但是在算法复杂性上本算法还有待进一步改进。第一,由于算法中建立了对待切分文本的概率模型,在切分之前就对文本进行了多次的扫描,这样在一定程度上增加了算法复杂度;第二,对于不同领域文本的权值的选择也是需要合理考虑的。所以在进一步的工作中,我们将从提高分词速度的角度出发,寻找在效率和准确率之间的一个平衡点。另一个改进的方向是,结合其他概率模型实现算法,考虑最大熵模型等经典概率模型。

参考文献:

- [1] 刘开瑛. 中文文本自动分词和标注[M]. 北京: 商务印书馆, 2000.
- [2] 黄昌宁. 统计语言模型能做什么?[J]. 语言文字应用, 2002, (1): 77-84.
- [3] MANNING C, SCHUTZ H. Foundations of Statistical Natural Language Processing[M]. MIT Press. Cambridge, MA: 1999.
- [4] ZHANG HP. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model[A]. Second SIGHAN workshop affiliated with 41th ACL[C], 2003. 63-70.
- [5] 李家福, 张亚非. 基于 EM 算法的汉语自动分词方法[J]. 情报学报, 2002, (6): 269-272.
- [6] 刘挺, 吴岩, 王开铸. 最大概率分词问题及其解法[J]. 哈尔滨工业大学学报, 1998, (12): 37-41.