文章编号：1007-791X (2009) 02-0124-06

基于优化最大匹配与统计结合的汉语分词方法

刘春辉，金顺福，刘国华*，李　颖

（燕山大学 信息科学与工程学院，河北 秦皇岛 066004）

摘　要：汉语自动分词是中文信息处理的前提，如何提高分词效率是中文信息处理技术面临的一个主要问题。基于词典和基于统计的分词方法是现有分词技术的主要方法，但是前者无法处理歧义字段，后者需要大量的词频计算耗费时间。本文提出优化最大匹配与统计结合的分词方法，首先提出优化最大匹配算法，在此基础上提出了规则判断与信息量统计两种消歧策略。然后，给出了优化最大匹配与统计结合的分词算法，提高了分词的效率。最后，基于分词算法实现中文分词系统，并通过实验对算法进行了分析和验证。

关键词：中文信息处理；词典；分词；优化最大匹配方法

中图分类号：TP391.1　　**文献标识码**：A

0　引言

词是汉语中最小的语义单位，是自然语言处理系统中重要的知识载体和基本操作单元[1]，而汉语文本中词与词之间却没有明显的分隔标记，是连续的汉字串，因而如何进行汉语分词（自动识别词边界，将汉字串切分为正确的词串）成为中文信息处理的首要问题。

目前，汉语分词[2]主要分为基于词典和基于统计两类方法。

基于词典[3]的分词方法主要依据构建一个分词词典并遵循一个切分评估原则来进行分词。基于词典的方法不需要大量的语言资源，程序实现简单，开发周期短，但由于该方法太过机械，得到的结果难以满足实际应用的要求。

基于统计[4-5]的分词方法主要是用于在分词过程中消除歧义现象，即消歧[6-8]。基于统计的方法主要靠一个或多个语料库获得相关信息统计的数据，该语料库一般都是训练语料库，规模虽然较小，但有一定的代表性。该方法根据从语料库中得到的数据（主要是词频和字间邻接关系[9-10]）来指导分词，如对可能的分词结果根据统计得到正确性最大的分词结果。这种方法完全抛弃了汉语的词法、语法、语义关系，而只根据统计算法的结果来进行分词，由于过分依赖统计算法，造成分词结果不能完全满足实际的需要。

鉴于上述问题，本文提出了一种基于优化最大匹配与统计结合的中文分词方法 OMSSA（Optimization Matching and Statistics Segmentation Algorithm），并实现了 CSS（Chinese Segmentation System）分词系统。

1　优化最大匹配和统计结合的方法

1.1　设计思想

输入待切分文本，首先通过系统的预处理模块，把句子切割出来，然后通过优化的正向最大匹配和逆向最大匹配方法对句子进行切分，如果两者一致就是正确的，否则通过比较词的个数，未登录词（词典中没有存储的词）来选择正确的切分，若仍然不能解决问题，则通过统计的方法确定歧义字段的切分。

1.2　基于优化最大匹配的方法

下面以优化正向最大匹配方法为例，介绍 OM-

收稿日期：2008-11-17　　基金项目：国家自然科学基金资助项目（60773100）

作者简介：刘春辉（1983- ），女，黑龙江集贤人，硕士研究生，主要研究方向为自然语言处理、数据库、信息安全；*通讯作者：刘国华（1966- ），男，黑龙江齐齐哈尔人，教授，博士生导师，主要研究方向为数据库安全、模式匹配、系统仿真，Email：ghliu@ysu.edu.cn。

SSA。

1.2.1 问题描述

传统的正向最大匹配方法主要存在以下问题：

1）正向最大匹配 [11] 方法通常采用整词二分[12]的匹配策略，该方法具有词典构造简单，易于实现等特点，但其 MAX（最大匹配方法中第一次截取汉字串的长度）需事先确定，若 MAX 设置过大，则易出现多次无意义匹配。

例 1　“扒手是很令人痛恨的。”

假设 MAX 为 6，初次截取的字串为“扒手是很令人”，与词典中词匹配，如果匹配不成功，截取新字串为“扒手是很令”，依此类推，直到截取字串为“扒手”，在扒手后面划分切分标志，在这次匹配过程中，进行了 4 次匹配，严重影响了分词的速度。

若 MAX 设置过小，则易出现歧义字段 [13-14]。

例 2　“中华人民共和国是有着悠久历史的国家。”

假设 MAX 为 5，则经过两次匹配，切分为“中华人民”，继续切分其余字串，使词“中华人民共和国”被切分成“中华人民”和“共和国”，因此产生分词错误；

2）根据金山词霸 2007 和 2005 年由中国社会科学院语言研究所主编的《现代汉语词典》（第 5 版）建立的语料库共有 76021 条词条，其中双音节词 [15] 占总词条的 43.8%，而三音节词占总词条的 24.98%，四音节词占总词条的 19.4%，而词条平均长度为 2.97，这些数据表明语料库中的词主要是低音节词，可是在分词过程中，采用正向最大匹配方法，必须从字串长词开始匹配，因而在匹配过程中多数匹配操作都是无意义的，降低了分词的速度。

因此，本文提出优化最大匹配方法解决上述问题。

1.2.2　优化正向最大匹配方法的词典结构

最大匹配方法是基于词典的分词技术，提出的优化方法首先需要对词典结构进行优化。本文设计的词典结构主要由两部分组成，首字索引表和词表正文，如图 1 所示。

区位码	…	1624	…	1652	…
首字成词最大词长	…	5	…	3	…
首字词组指针	…	·	…	·	…

首字	词长	尾字 1	词组 1	…	尾字*N*	词组*N*
案	5	栏	犯照片	…	人	卷保管
案	…	录	件记	…	明	情不
案	…	…	…	…	…	…
案	2	秤	^	…	情	^
案	1	Y				

图 1　词典结构

Fig. 1　Structure of dictionary

1）首字索引表

词的首字索引可根据汉字的区位码采用公式

$$position=(c_1-176)\times 94+(c_2-161) \tag{1}$$

直接定位，*position*为词在首字索引结点中的位置，c_1为词首字第一个字节的无符号数，c_2为第二个字节的无符号数。

首字索引表还包括如下数据：

首字成词的最大词长：找出首字成词的最大词长，存入词典中，以此来解决问题 1 提出的如何设定正向最大匹配 MAX 的问题；

首字指针：索引字成词，则指向词表正文中该索引字所在位置，否则不分配指针。

2）词表正文

每个首字的词表正文以词长来划分成元组，每一元组包含如下数据：

词长：每个首字所成词的词长不同，一个首字对应一个词长作为一个元组。元组顺序按词长由大到小排列；

尾字：每个词的最后一个汉字；

词组：每个词的首尾字之外的字串。当词长小于等于 2 时，无词组项。

词长为 1 时，判断单字是否成词，Y 表示单字成词，N 表示单字不成词。

1.2.3　优化的正向最大匹配方法

基于优化正向最大匹配方法的词典结构，给定

汉字串$string=s_0s_1\cdots s_{(n-1)}$，$n$为汉字串的最长词长度，对$string$进行分词，分词方法如下：

1）取$string$的第一个字符s_0，根据式 (1) 求得s_0在首字索引表中的位置，读取s_0的信息，索引字成词最长词长为$length$和首字词组指针为p_{Str}；

2）$i=0$，如果$length>n$，转 9)；

3）读取s_0在词表正文不同元组下的词长字段$nlength_j$，$j\in(0, n_{Cou}-1)$，n_{Con}为s_0成词不同长度的个数，其中$nlength_0=length$；

4）通过p_{Str}读取s_0在词表正文中的信息，按照词长$length$读取相对应的尾字字段 last W_k，$k\in(0, n_{Count}-1)$，n_{Count}为s_0成词词长为$length$时组成词的个数，如果$length>2$，还要读取词组字段 full W_k；

5）取$string$的第$length$个字符s_{length}，与 lastW_k进行匹配，如果匹配不成功，$i=i+1$，转 8)；

6）读取字符串$s_1\cdots s_{(length-1)}$与 lastW_k对应的词组字段 fullW_k进行匹配，如果匹配不成功，$i=i+1$，转 8)；

7）$s_{(length+1)}$作为待分字符串的第一个字符转回 1)，如果$s_{(length+1)}$为空，转 9)；

8）读取$nlength_i$值，$length=nlength_i$，转回 4)；

9）$length=n$，转回 3)；

10）返回。

1.3 基于统计的方法

待分文本进行正向和逆向最大匹配后，如果结果一致则得出正确的分词结果，否则找出不一致的最小字段，对其进行如下处理。

1.3.1 规则判断

规则 1 专有词组 如果用户标注待分文本的专业代码，首先按专业词库中词作为划分标准。

规则 2 成语、熟语优先 如果歧义字段中含有成语或熟语，则尽可能保证该部分成词。

规则 3 比较词数 如果词的个数不相等，则选择词数较少的一个作为切分结果选择切分段数少的切分方案作为切分结果。如正向结果“证实/行之有效/的”与逆向结果“证/实行/之/有效的”，那肯定选择正向结果作为结果。如果词的个数相等，则转至下一步。

规则 4 未登录词数 若未登录词的个数不相等，则未登录词的个数少的一个就作为结果。如正向结果“发电/子 [未登录词]”与逆向结果“发/电子”，则选择逆向结果作为结果；若未登录词的个数相等，则通过统计信息量统计方式处理歧义字段。

1.3.2 信息量统计

本文利用已有的训练语料库计算词频度，通过互信息和t-信息这两个信息量统计 [9] 的方法，判断相邻汉字是否成词来对歧义字段进行处理。

1.3.2.1 信息量的计算公式

1）互信息

定义 1 设有序汉字串为xy，汉字x, y之间的互信息定义为

$$I(x: y)=\frac{p(x, y)}{p(x)p(y)} \tag{2}$$

其中，$p(x, y)$是x, y为汉字串xy作为二字词出现的概率，$p(x)$，$p(y)$分别代表x和y作为单字词独立出现的概率。

设xy作为二字词出现的次数为$f(x, y)$，而x和y可作为单字词独立出现的次数分别为$f(x)$，$f(y)$，N为所有的汉字，于是

$$\begin{cases} p(x, y)=\dfrac{f(x, y)}{N} \\ p(x)=\dfrac{f(x)}{N} \\ p(y)=\dfrac{f(y)}{N} \end{cases} \tag{3}$$

根据上述定义的互信息能够反映出汉字间结合关系的紧密程度。

如果$p(x, y)=0$，表明x和y不成词。

如果$p(x, y)\neq 0$，当$p(x)=0$或者$p(y)=0$时，表示x和y成词的机率比较大，否则，当$p(x)\neq 0$而且$p(y)\neq 0$时，表示x和y成词的机率不确定，进行判断则需要参照上下文，进一步寻找依据。

2）t-信息

定义 2 设有序汉字串为*xyz*，汉字*y*相对于*x*及*z*的*t*-信息定义为

$$tx,\ z(y)=\frac{p(z|y)-p(y|x)}{\sqrt{p^2(z|y)-p^2(y|x)}} \tag{4}$$

其中，$p(z|y)$表示*y*作为单字词出现的条件下，*yz*作为二字词出现的条件概率，$p(y|x)$表示*x*作为单字词出现的条件下，*xy*作为二字词出现的条件概率。于是

$$\begin{cases} p(z|y)=\dfrac{p(y,\ z)}{p(y)} \\ p(y|x)=\dfrac{p(x,\ y)}{p(x)} \end{cases} \tag{5}$$

根据式 (3)，可计算出字*y*与*z*，字*y*与*x*结合的紧密程度，以此判断应该是*yz*成词还是*xy*成词。

①$t_{x,\ z}(y)<0$时，字*y*与*x*具备结合趋势，值越小，结合趋势越强；

②$t_{x,\ z}(y)=0$时，$t_{x,\ z}(y)$不表示任何结合趋势；

③$t_{x,\ z}(y)>0$时，字*y*与*z*具备结合趋势，值越大，结合趋势越强。

1.3.2.2　信息量统计方法的分词结果示例

本文的统计基础是 76021 条词条，词容量*N*为 1025836，其中总词频度为 12638035。其中本文测试所用实例的词语频度如表 1 所示。

表 1　实例测试所用的词频度

Tab. 1　Word frequency of Example Test

实例	词频度	实例	词频度	实例	词频度
的	18639	结	3261	成分	1580
确	2619	合	3809	分子	4275
实	2784	成	21402	子时	15
在	20216	分	16806	时有	8
理	6207	子	5731	应	2689
的确	2163	时	46249	用	10652
实在	2581	有	84067	于	6831
确实	2059	结合	2089	应用	3958
在理	228	合成	896	用于	3560

例 3　（他说）的　确　实　在　理

$I(x:\ y)$　45.45　289.69　47.04　1.89

$t_{x,\ z}(y)$　0　0.86　0.29　−0.98　0

正向：（他说）　的确　实在　理

逆向：（他说）　的　确实　在理

词频：（他说）　的　确实　在理

本文：（他说）的确实在理

例 4　结　合　成　分　子　时　有（…）

$I(x:\ y)$　172.53　11.27　4.51　45.53　0.058　0.002

$t_{x,\ z}(y)$　0　−4.53　−0.725　0.741　−0.98　−0.93　0

正向：结合　成分　子时　有（…）

逆向：结　合成　分子　时有（…）

词频：结合　成　分子　时有（…）

本文：结合成分子时有（…）

例 5　（这个项目）应　用　于（…）

$I(x:\ y)$　141.75　50.19

$t_{x,\ z}(y)$　0　−0.794　0

正向：（这个项目）应用　于（…）

逆向：（这个项目）应　用于（…）

词频：（这个项目）应用　于（…）

本文：（这个项目）应用于（…）

2　优化最大匹配与统计结合的分词算法 OMSSA

下面给出优化最大匹配与统计结合的分词算法，其思想主要是对待切分文本依次进行预处理，优化最大匹配分词，规则判断和统计方法处理 4 个步骤。

算法 1　OMSSA（Inputfile，Outputfile）

输入：经过预处理的待切分文本

输出：词之间加入分割符“/”的分词结果

1）首先根据标点符号将 Inputfile 文本断句，句子作为下一步处理的单位；

2）对待切分文本通过使用有限自动机（FSA）进行预处理，识别其中有明显特征的中英文数字（包括分数、小数、百分数、基数词、序数词等）、日期、人名、域名、图表等，之后是对包括标点符号、特殊字符、段落等的去除；

3）对句子通过优化的正逆向最大匹配方法进

行分词，得到初切分结果；

4）对出现歧义的字段根据 1.3.1 节的规则进行处理；

5）如果歧义现象仍然存在，通过信息量统计的方法进行歧义消解；

6）重复 3）→4）→5），直到处理完 1）分出的所有句子单元；

7）输出待切分文本的分词结果存入到 Output-file。

3 CSS 分词系统设计及结果分析

3.1 CSS 分词系统原型

本文基于 OMSSA 实现了分词系统 CSS，图 2 给出了分词系统的基本流程图。

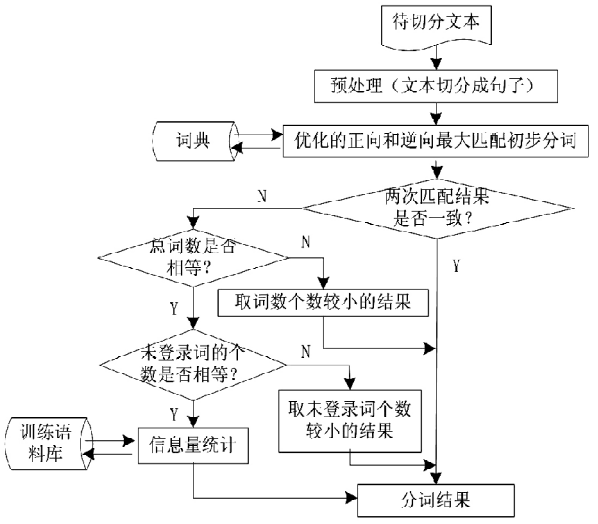

图 2 CSS 分词系统原型

Fig. 2 CSS model of segmentation system

3.2 实验结果及分析

为了证明本文提出方法的正确性，通过 CSS 对 OMSSA 展开验证。首先选取了医药、法律、交通、计算机、艺术 5 个类别文章作为测试文本 [16]，从中提取出主要含有词长为 2 的低音节词 50 篇文本，分别对其采用 OMSSA 和最大匹配方法、文献 [7] 提出的 DSfenci、文献 [12] 提出的组合 Hash 索引分词算法四种方法分词，同理对含有词长为 3~7 的低音节词文本各 50 篇进行对比，如图 3 所示。

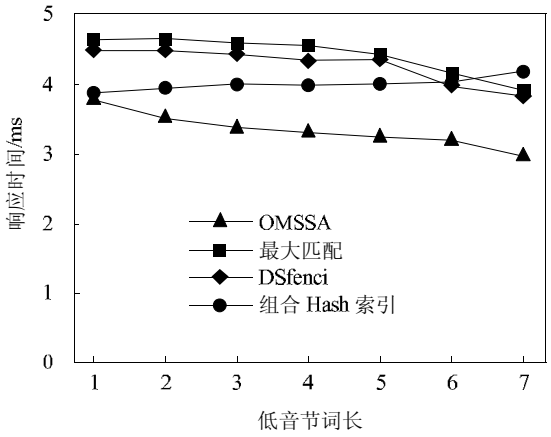

图 3 低音节词识别效率示意图

Fig.3 Word recognition efficiency of the bass section diagram

从低音节词识别效率示意图可见，用 OMSSA 对文本预切分，提高了区分低音节词的速度。而且在优化最大匹配分词过程中又能将歧义字段标注出来，本文为了验证 OMSSA 消歧的效率，从约 40 万字的语料库中提取出含有歧义字段的汉语句子 320 条，与一些其它有歧义的句子组成一篇约 3500 字的文章进行测试，其中歧义字段数为 476 条，测试结果如表 2。

表 2 不同实验结果的比较

Tab. 2 Comparison of different experiment results

使用方法	正确切分数	错误切分数	正确率
正向最大匹配	174	302	36.6%
逆向最大匹配	286	190	60.1%
词频信息统计	392	84	82.4%
有向图和统计加规则 [6]	402	74	84.5%
OMSSA	431	45	90.1%

从表 2 的统计结果表明，文章提出的方法将歧义字段切分正确率分别提高了 53.5%、30%、7.7%。可见本文首先采用优化正向和逆向最大匹配方法得到预切分结果，之后利用规则判断和信息量统计方法消歧，使分词正确率有了明显的提高。

4 结束语

汉语自动分词是中文信息处理的“瓶颈”问题，提高分词系统的性能也是改善信息检索、搜索引擎等领域功能的前提，而分词系统的性能与分词的速度和歧义的消解程度直接相关。本文以提高分词的效率为目标，提出了基于优化最大匹配与统计结合的分词方法。通过动态设定 MAX 和词末尾字

的查询，有效地提高了分词的速度。给出了消解歧义的方法，通过分词规则来对歧义字段进行第一步处理，以此为基础运用信息量统计策略进一步处理歧义字段。实验证明，通过本文提出的方法实现的分词系统，有效地提高了分词的效率。

参考文献

[1] 杨宪泽, 谈文蓉, 刘莉, 等. 自然语言处理的原理及其应用 [M]. 成都: 西南交通大学出版社, 2007.

[2] 曾华琳, 李堂秋, 史晓东. 一种基于提取上下文信息的分词算法 [J]. 计算机应用, 2005,25 (9): 2025-2027.

[3] 黄昌宁. 中文信息处理中的分词问题 [J]. 语言文字应用, 1997,6 (1): 72-78.

[4] 朱鉴, 张建, 李淼. 一种有效解决汉语歧义切分的方法 [J]. 计算机工程与应用, 2007,43 (11): 175-177.

[5] 王显芳, 杜利民. 利用覆盖歧义检测法和统计语言模型进行汉语自动分词 [J]. 电子与信息学报, 2003,25 (9): 1168-1173.

[6] 苗夺谦, 卫志华. 中文文本信息处理的原理与应用 [M]. 北京: 清华大学出版社, 2007.

[7] 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法 [J]. 小型微型计算机系统, 2006,27 (9): 1766-1771.

[8] 马玉春, 宋海涛. Web 中文文本分词技术研究 [J]. 计算机应用, 2004,24 (4): 134-136.

[9] 郑德权, 于凤, 王开涛. 基于汉语二字应成词的歧义字段切分方法 [J]. 计算机工程与应用, 2003,39 (1): 17-19.

[10] 朱巧明, 温滔, 李培峰, 等. 一种基于多元信息库的自适应汉语歧义切分方法[J].小型微型计算机系统,2006,27(8):1597-1600.

[11] 金瑜, 陆启明, 高峰. 基于上下文相关的最大概率汉语自动分词算法 [J]. 计算机工程, 2004,30 (16): 146-148.

[12] 王东, 陈笑蓉. 一种改进的高效分词词典机制 [J]. 贵州大学学报, 2007,24 (4): 380-384.

[13] 闫引堂, 周晓强. 交集型歧义字段切分方法研究 [J]. 情报学报, 2000,19 (6): 637-643.

[14] 李凯, 左万利, 吕巍. 汉语文本中交集型切分歧义的分词处理 [J]. 小型微型计算机系统, 2004,25 (8): 1486-1490.

[15] 蒋斌, 杨超, 赵欢. 基于二字词位图表的汉语自动分词词典机制 [J]. 湖南大学学报 (自然科学版), 2006,33 (1): 121-123.

[16] http: //www. nlp. org. cn/docs/doclist. php? cat_id=16&type=15 [EB/OL].

A Chinese segmentation method based on optimization maximum matching and statistics

LIU Chun-hui, JIN Shun-fu, LIU Guo-hua, LI Ying

(College of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China)

Abstract: Chinese automatic segmentation is the precondition of Chinese information processing. A primary problem of Chinese information processing is how to improve segmentation efficiency. The segmentation method based on dictionary and statistics is main method of present segmentation technology; the former can not deal with ambiguity and the latter need a large amount of time to calculate word frequency. A method based on optimization maximum matching integrated with statistics is proposed. The method uses the segmentation method based on optimization maximum matching in the first step, then propose ruler judgment and information quantity statistics. During the second step, an algorithm of based on optimization maximum matching and statistics is presented focus on improving segmentation efficiency. Finally, Chinese automatic segmentation system is implemented by using the algorithm. Finally, the algorithm is analyzed and validated by experiments.

Key words: Chinese information processing; dictionary; segmentation; optimization maximum matching