一种结合规则和统计的中文分词系统

摘要：中文分词是中文信息处理的基础。是搜索引擎、机器翻译、文本分类、文本朗读、语音合成、信息抽取、情绪分析、中文问答系统等中文信息处理各项任务中的基础课题。目前常用的中文分词方法大致分两类：基于规则和基于统计的分词方法。近年来，中文分词研究取得了一定的进展，本文通过对现有分词算法的分析，提出了一种新的分词方法，这种方法基于 MMSEG 算法，并结合统计分词方法实现句子的分词，能准确高效地进行中文的分词。
关键词：分词；MMSEG；统计与规则

A Chinese word segmentation system which combines rules and statistic method

Abstract: Chinese word segmentation is the basic of processing Chinese information. It is the basal task in search engine, machine translation, text classification, text reading, speech synthesis, information extraction, sentiment analysis and Chinese question-answering system, etc. Those common methods to segment Chinese text are divided into two categories roughly: based on rules and based on statistics. In recent years, researches in this field have made great development. This paper presents a new and effective method which combines MMSEG and statistic method to segment Chinese text after studying nowadays Chinese word segmentation method.
Keywords: word segmentation; mmseg; rules and statistics

1 引言

随着社会信息化智能化程度的不断提高，诸如语音助手、搜索引擎、推荐系统等智能服务越来越深入社会各个领域，而在计算机处理中文方面，和处理英语、西班牙语等印欧语系的语言极为不同的一个基础课题是分词问题。英语、西班牙语等语言都具有自然分隔符，而中文则没有这样的自然分隔符，所以分词是计算机处理中文信息任务中必须进行的工作，而且是大多数这样的场景下的首要工作，经过多年的探索和研究，从早先的完全依赖规则到后来引入统计模型，中文分词课题的研究过程中涌现出了各种分词模型，这里面比较具有代表性的诸如规则处理上的最大匹配法（MM）及其各种变体，统计模型上的 N-gram，隐马尔可夫模型（HMM），最大熵马尔可夫模型（MEMM）和条件随机场（CRF）等。中文分词的效果也越来越好了。基于规则的分词方法有其自身局限性例如规则覆盖率不高和规则的主观性较强，统计分词则过于依赖训练语料库以及分词过程忽视了汉语的句法语法等信息[1]。而结合规则与统计的方法是中文分词领域研究的热点[4]。本文结合规则和统计分词算法提出一种改进的分词方法，这种方法基于最大匹配法的一个变体——MMSEG 算法[2-3]，在 MMSEG 分词算法的基础上引入统计算法，改善分词的精确率（precision）和召回率（recall）。

2 分词算法及其实现

2.1MMSEG 分词算法

MMSEG 包含简单最大匹配法和复杂最大匹配法，本文选取其中的复杂最大匹配法与统计方法进行结合。MMSEG 的复杂最大匹配法首先识别出从当前读取位置开始所有的块（chunk），块是包含三个词的组合，这三个词中如果是存在单字成词则不用必须存在字典中也被看做一个词对待，如果是多字词则必须存在字典当中，对于每一个块，它的第二个词和第三词可以为空，且块中的词是保持句子的前后序列的。对于句子 S 由多个字符 C 构成，则

$$S = C_1C_2C_3...C_{n-1}C_n \tag{1}$$

当前分词位置为 $start$ 时则其所有块组成的集合表达式为

$$chunks = \{[C_{start}...C_i, C_{i+1}...C_j, C_{j+1}...C_k] \mid start \le i < j < k \le n\} \tag{2}$$

例如有字典 dict=[研究，生命，研究生]，对句子“研究生命”进行分词则能分出所有的块为

[研究，生命]
[研究，生，命]
[研究生，命]
[研，究，生命]
[研，究，生]

如果 $chunks$ 元素个数等于 1 则选择唯一的块中的第一个词作为分词结果返回，如果元素多于 1 个则依次应用规则 1 到 4 进行过滤。

规则 1 最大总词长

取总词长度最长的块中的第一个词作为分词结果返回。

假设 $chunks$ 的元素个数为 N， $chunks$ 中的元素 i 的有效词数（即不为空的词的个数）为 M_i， $l(w)$ 为词 w 的长度， $w^{(i)}$ 为 i 中的词，则 i 的总词长 L_i 计算公式为

$$L_i = \sum_{j=1}^{M_i} l(w_j^{(i)}), i \in chunks \tag{3}$$

取得最大词长的块的公式为

$$\arg\max_{i \in chunks}(L_i) \tag{4}$$

例如对于以上例子有

表 1 各块的总词长

块(chunk)	长度
[研究，生命]	4
[研究，生，命]	4

[研究生，命]	4
[研，究，生命]	4
[研，究，生]	3

如果过滤后得到的候选块个数等于 1 则选择该块中的第一个词作为分词结果返回，如果多于 1 则应用规则 2 进行过滤。例子中根据规则 1 排除第 5 种方案，接着对剩下的 4 种应用规则 2。

规则 2 最大平均词长

取平均词长度最大的块中的第一个词作为分词结果返回。

计算平均词长公式为

$$avg_i = \frac{L_i}{M_i}, i \in chunks \tag{5}$$

选择最大平均词长的块的公式为

$$\arg\max_{i \in chunks}(avg_i) \tag{6}$$

上文例子的平均词长为

表 2 各块的平均词长

块(chunk)	平均词长
[研究，生命]	2.00
[研究，生，命]	1.33
[研究生，命]	2.00
[研，究，生命]	1.33

如果过滤后得到的候选块个数等于 1 则选择该块中的第一个词作为分词结果返回，如果多于 1 则应用规则 3 进行过滤。例子中排除两种方案，剩下两种继续应用规则 3。

规则 3 最小词长方差

取词长方差最小的块中的第一个词作为分词结果返回。

词长方差计算公式为

$$\sigma_i^2 = \frac{1}{M_i}\sum_{j=1}^{M_i}(l(w_j^{(i)}) - avg_i)^2, i \in chunks \tag{7}$$

得到最小方差的块的公式为

$$\arg\min_{i \in chunks}(\sigma_i^2) \tag{8}$$

例子中计算得到

表 3 各块的词长方差

块(chunk)	方差
[研究，生命]	0.00
[研究生，命]	1.00

如果过滤后得到的候选块个数等于 1 则选择该块中的第一个词作为分词结果返回，如果多于 1 则应用规则 4 进行过滤。例子中选出了正确的分词方案，即返回[研究，生命]中的“研究”一词作为此次分词的结果。

规则 4 最大单字语素自由度之和

取单字语素自由度之和最大的块中的第一个词作为分词结果返回。

假设有多重集（multiset）SC_i为i中单字词的集合，则有

$$SC_i = \{w^{(i)} \mid l(w^{(i)}) = 1\}, i \in chunks \tag{9}$$

以各个单字词w的出现频数f_w作为$\log$函数的真数并相加得到最大语素自由度之和d，其计算公式为

$$d_i = \sum_{w \in SC_i} \log(f_w), i \in chunks \tag{10}$$

取得最大语素自由度之和的块的公式为

$$\arg\max_{i \in chunks}(d_i) \tag{11}$$

依靠以上规则可以解决大部分的分词歧义问题，但也有相当的一部分句子无法进行正确分词，对于应用规则 4 后候选块个数仍然多于 1 个的情况，MMSEG 则无法处理，所以本文引入了下文的统计分词算法。

2.2 统计分词算法

本文使用的统计分词算法类似中文分词中的最大概率法，最大概率分词是认为一个句子可能有多种分词结果，而把概率最大的那种分词结果作为正确的结果[6]，不完全和最大概率分词方法相同的是本文的统计算法是计算得到概率最大的块选择其第一个词作为正确分词结果返回，同时因为基于这样的假设：每个词的出现概率独立。虽然这个假设并不实际，所以相对于 N-gram 和 HMM 等模型却大大减少了概率的计算量，提高了分词的速度，在 MMSEG 处理后由统计分词处理 MMSEG 不能处理的情况，在保持较快的分词速度的同时保持了比较高的精确率，取得不错的效果。根据大数定律，可以知道只要语料库足够大，一个词的出现的频率就等于其概率[5]，在这里直接使用词w在训练语料库中出现的频率代表其概率，如果发现未登陆词，为了避免导致分词概率为 0，则将该词的出现次数设置为 1。设语料库词汇量为$\#$，w出现次数为$\#(w)$，频率为f_w，概率为$P(w)$，有

$$f_w = \frac{\#(w)}{\#} \tag{12}$$

由大数定律得

$$P(w) = f_w \tag{13}$$

求出最大概率的块，选取其中的第一个词作为分词结果返回

$$P_i = \prod_{j=1}^{M_i} P(w_j^{(i)}), i \in chunks \tag{14}$$

$$\arg\max_{i \in chunks}(P_i) \tag{15}$$

3 实验结果

为了测验本文提出的算法的效果并且保证结果对比的客观，本实验采用封闭测试，使用的训练语料，测试语料以及评测答案均来自山西大学整理的 Bakeoff 第四届比赛封闭测试中所使用的语料，其中训练语料词数为 528238，测试语料词数为 109873，对比测试算法为 FMM 算法，评测指标采用精确率和召回率，实验结果如下表所示

表 4 实验效果对比

	准确切分数	总切分数	切分正确的句子数	总句子数	句子正确率	精确率	召回率
FMM	108296	117200	2212	3654	60.54%	92.40%	98.56%
MMSEG	108347	116625	2490	3654	68.14%	92.90%	98.61%
本文算法	108820	117143	2498	3654	68.36%	92.90%	99.04%

通过实验结果分析，本文提出的规则和统计结合的算法比 FMM 算法以及 MMSEG 分词算法的精确率和召回率都有所提高，同时在切分完全正确的句子方面相比 FMM,MMSEG 也有所改善。

4 结束语

本文提出的规则和统计相结合的分词算法，能有效改善 MMSEG 的分词效果，但也还存在不少尚待改进的地方，这部分将简要地归纳一下本文算法的尚待改进之处：

1 词出现概率独立假设这个和现实情况不符合，虽然简化了计算过程降低了计算量但也同时丢失了词间的上下文信息，将统计算法部分参考 N-gram 模型改进应该会使效果大大提高。

2 未登录词的频率替代问题，本文使用的替代档案是用 1 代替使得该词对于所求概率影响无效化，这将在处理大量未登陆词的情况下导致分词效果严重降低，所以仍然需要更为有效的平滑方法处理这种情况。

参考文献

[1] 赵伟，戴新宇，尹存燕，陈家骏.一种规则与统计相结合的汉语分词方法 [J] .计算机工应用研究，2004，(3):23-25

[2] 张中耀，葛万成，汪亮友，林佳燕.基于 MMSEG 算法的中文分词技术的研究与设计 [J] .信息技术，2006，(6):17-20

[3] Tsai C H.MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm [EB/OL] .http://technology.chtsai.org/mmseg/，2000

[4] 奉国和，郑伟.国内中文自动分词技术研究综述 [J] .图书情报工作，2011，55(2):41-45

[5] 吴军.数学之美 [M] .北京:人民邮电出版社，2014

[6] 丁浩.基于最大概率分词算法的中文分词方法研究 [J] .科技信息，2010，(21):587