提高汉语自动分词精度的多步处理策略①

赵铁军 吕雅娟 于浩 杨沐昀 刘芳

（哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001）

摘要：汉语自动分词在面向大规模真实文本进行分词时仍然存在很多困难。其中两个关键问题是未登录词的识别和切分歧义的消除。本文描述了一种旨在降低分词难度和提高分词精度的多步处理策略，整个处理步骤包括7个部分，即消除伪歧义、句子的全切分、部分确定性切分、数词串处理、重叠词处理、基于统计的未登录词识别以及使用词性信息消除切分歧义的一体化处理。开放测试结果表明分词精确率可达98%以上。

关键词：汉语自动分词；歧义；多步处理

中图分类号：TP391.12

Increasing Accuracy of Chinese Segmentation with Strategy of Multi-step Processing

ZHAO Tie-jun LV Ya-juan YU Hao YANG Mu-yun LIU Fang

(School of Computer Science and Technology, Harbin Institute of Technology Harbin 15001)

E-mail: {tjzhao, lyj, yu, ymy, liufang}@mtlab.hit.edu.cn

Abstract: The automatic word segmentation of Chinese sentences is difficult when the processing mechanism faces large-scale real texts. The crucial two issues in Chinese segmentation are the identification of unknown words and the disambiguation of segmentation strings. This paper describes a strategy based on multi-steps processing for decreasing the difficulties and improving the accuracy of the segmentation. The processing steps include seven parts, i.e., disambiguation of pseudo-ambiguities, full segmentation of a sentence, determinate segmentation for some words, processing of numeral string, processing for reduplication of words, statistical identification for unknown words and final correction for segmentation ambiguities with part-of-speech which is integrated in the tagger. The output of this procedure is promising with above 98% accuracy in open-test.

Keywords: Chinese segmentation; ambiguity; multi-step strategy

---

① 收稿日期：2000－05－23

基金项目：国家863计划(863－306－ZT03－06－3/863－306－ZD13－04－4)；国家自然科学基金(69775017)

作者赵铁军，男，1962年生，博士，教授，主要研究方向为机器翻译、计算语言学、人工智能。

一、引言

汉语自动分词在面向大规模真实文本进行分词时仍然存在很多困难。其中两个关键问题是未登录词的识别和切分歧义的消除。近年来国内外对此进行了大量研究。除了传统的最大匹配方法(Maximum Matching,简称 MM 方法),许多基于统计的方法也引入到分词过程中。例如分词与词性标注一体化方法[1],随机有限状态算法用于分词[2],模拟物理研究中结晶过程的统计方法也被尝试于分词过程[3]。此外,还有大量的基于统计或规则的汉语未登录词识别的研究,这里不能一一列举。本文提出了一种在整个分词过程中进行多步处理的策略,目的是提高汉语自动分词的精度。

本文第 2 节描述了整个分词中的多步处理策略,随后的两节分别介绍基本切分算法、具有特殊词性的某些词确定性分词、数字串切分、重叠词切分等。第 5 节描述了使用统计方法和与词性标注集成进行未登录词识别的实现机制。最后本文给出了实验结果,表明本文提出的策略取得了较好的效果。

二、基于多步处理策略的自动分词过程

为了避免汉语自动分词过程中出现切分盲点[4],切分的基础应该是全切分算法。通过把整个分词过程分解为若干子过程,实现多步处理策略,从而解决了某些特殊的分词问题。图 1 表示多步处理策略的自动分词实现流程。

切分歧义中最常见的一类歧义是交集歧义,但有许多交集歧义属于伪歧义。所谓伪歧义是指包含交集歧义但实际文本中只能有或几乎只有一种切分可能的字段。我们采用查表的方式直接确定分词形式,而不再参与后续的分词处理过程。例如:“办法规定”,实际文本中的切分形式一般就是“办法/规定”,其中的“法规”很难成为一个词。这样,通过查伪歧义表[注],某些切分形式一开始就决定了。大规模的统计表明 4000 多个列表项可以覆盖全部交叉歧义的 53.35%[5]。

在流程中,首先使用全切分算法给出一个汉语句子的所有可能的切分方式,即任何匹配词典中词的字串都被加入到全切分结果集合当中。这样,就保证了句子中任何可能的切分结果都不会被后续处理所遗漏。

全切分完成后,某些切分结果(即出现在词典中的词)具有特殊的词性,不会与其他词产生交叉切分歧义,因而很容易被首先确定下来。这些词性包括成语、惯用语、叹词、语气词等 4 类。这就是确定性分词处理步骤。实用的汉语分词系统除了要对上述情况予以处理以外,还可以专门建立用户自己定义的辅助词典,把某些满足汉语惯用表达方式的短语(如四字词语)存入其中,一开始就将其切分为一个词。因为在这样的情况下,分开来不易分析,而且常常会出现错误。

汉语数词(即一、二、三、百、千、万 … …)是典型的单字词,数词串就是单字词串。如果句子中出现数词串,有必要首先将其合并,从而减少句子中切分结果的数量,以利于分词过程后续阶段的处理。

汉语的重叠现象需要单独处理。因为许多重叠词的存在,会使句子中单个汉字节点增多。如果不合并它们,就会把属于同一词或同一节点的汉字分割开来,可能会使句子的句法结构分析产生错误。所以,需要将重叠部分合并以后再设置相应的标志,有利于后续的句法分析。

整个分词处理过程中的重要部分是未登录词的识别。识别过程采用了统计方法，对句子中出现的单字串进行中国人名、中国地名、外国人名译名或地名译名的识别。统计单个汉字在上述名字中出现的频率，作为识别的依据。最后，少数尚未确定的组合或交叉歧义以及未登录词将在下一步词性标注过程中借助词性信息和上下文信息予以确定。这一过程也称为分词词性标注一体化过程。

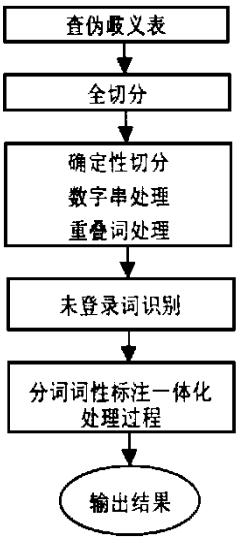

图 1　基于多步处理策略的汉语自动分词过程

三、全切分与确定性分词处理

整个分词过程的基础是全切分算法。其输出结果是一个句子的所有可能切分的形式，即任何出现在词典中的词都放入全切分列表。例如，对于汉语句子"全切分的结果是一个有向图。"，其全切分结果如图 2 所示。这是一个有向图，图中的每条弧表示一个切分结果，即一个词典中存在的词。如："全、切、分、切分"等。全切分算法保证了任何出现在句子中字串如果它属于词典中的一个词，则其必然为有向图中的一个弧。

全部分词形式通过词典查询来确定，所有汉语单字(6763 个)都包含在词典中看作一个单字词。与词典查询相关的数据结构主要是词典的索引结构，本系统采用两级索引结构。第 1 级索引是 6763 个汉字，组织成哈希表形式。第 2 级索引包括词典中的全部词，其第 1 个字对应于上述 6763 个汉字。第 2 级索引使用二分查找算法。词典的外部存储形式使用 Huffman 编码。

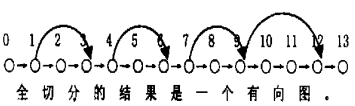

图 2　一个汉语句子的全切分形式

确定性分词处理的目标是把句子中那些不包含歧义的部分分离出来，并确定其切分结果。汉语成语(词性标记为 i)、惯用语(词性标记为 l)、叹词(词性标记为 e)、语气词(词性标记为 y)以及拟声词(词性标记为 o)等都几乎不可能与其他词产生交叉歧义。所以这些词从分词角度看是孤立的，可以首先被确定为一个确定的切分形式，同时也就确定了词性。例如，在句子"这件事使人浮想联翩"中，成语"浮想联翩"决不可能有组合歧义或与其他词产生交叉歧义，其边界可以确定无误。再如，句子中一个标点前面的叹词或语气词"啊呀/e"、"呀/y"等都是确定的。

四、数字串和重叠词的处理

由纯单字词串组成的字段包括以下一些情况：

- 数词或数词量词串；
- 重叠词形式；
- 未登录词；
- 前缀或后缀。

其中第三种情况在下节叙述，第四种情况即前后缀与相应词干结合的处理将在分词之后

基本短语处理阶段实现，本文不再介绍。数词串的处理只涉及长度大于等于 2 的情况，在处理过程中必须区分一般数词（如“一”到“九”）和那些表示幂数的数词（如“十”、“百”），以便转换为正确的数值，用相应的阿拉伯数字表示。

重叠是汉语当中存在的普遍现象，这些重叠类型可以列表如下。

表 1　汉语词的重叠类型

类型	标识	合并后的形式	合并后的词性	例子
A/ AB	1	AB	动词	握/ 握手
A/ 了/ AB	2	AB	动词	握/ 了/ 握手
A/ 一/ AB	3	AB	动词	握/ 一/ 握手
A/ AB/ B	4	AB	与 AB 相同	漂/ 漂亮/ 亮
A/ 不/ AB	5	AB	与 AB 相同 是动词或形容词	喜/ 不/ 喜欢 高/ 不/ 高兴
A/ 里/ AB	6	AB	状态词	晃/ 里/ 晃荡
AB/ B	7	AB	状态词	热呼/ 呼
A/ A	8	A	动词	停/ 停 瞅/ 瞅
A/ 了/ A	9	A	动词	瞅/ 了/ 瞅
A/ 一/ A	10	A	动词	瞅/ 一/ 瞅
A/ 了一/ A	11	A	动词	瞅/ 了/ 一/ 瞅

重叠词的识别过程相对较为简单，只要一个词序列的词性满足上述组成，就视为一个重叠词，并查词典获取其合并后的词形在词典当中的信息。将节点合并以后，把标识存放到合并后节点的特殊域内。有关例子参见图 3。

五、未登录词的统计识别及与词性标注的集成

这里的未登录词是指词典中未收录的中国人名、中国地名、外国人名或地名的译名，因为汉语中 90％左右的未登录词是由单字串构成[6]，所以这里提出的算法只识别那些包含在一系列单个汉字构成的串中的未登录词。如果某个人名、地名或译名含有一个普通词（如“马胜利”、“红领巾路”），我们则暂时忽略其识别而推迟到短语分析时再做处理。

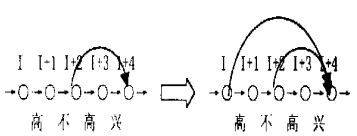

图 3　重叠词处理的一个例子

要同时实现未登录词识别的高精确率和高召回率是很困难的，理想情况是在保证普通词分词的高精确率的条件下识别更多的人名、地名或译名。

未登录词的识别需要一系列资源[7]，表 2 列出了有关资源。其中包括各个汉字作为人名、地名、译名的频率，同时这些汉字作为普通单字词出现的频率也要在识别过程中加以考虑，以避免普通词和未登录词之间的冲突。

对于汉语人名（包括姓氏），其长度一般为 2 或 3 个汉字，同时亦有可能包含中国地名或译名。因此，串长为 2 个或 3 个汉字的字段将分别进行人名、地名、译名的混合识别。对于串长大于 3 个汉字的单字串，主要进行地名和译名的识别。在此处理过程中，如果一个串能够分成 2 个或更多的长度为 2 或 3 的子串，则可以递归地调用串长为 2 或 3 的处理过程。这样，地名

表 2　未登录词识别所需资源

识别类型	所需资源	资源容量或使用情况
中国人名识别	* 姓氏用字频率	729 字
	* 人名用字频率[8]	3345 字
	* 称呼	使用
	* 上下文	未使用
中国地名识别	* 地名用字频率	32200 地名
	* 行政区划、地理单位	使用
外国人名译名识别	* 译名用字频率	30000 译名
外国地名译名识别	* 译名用字频率	30000 译名

(或译名)的处理过程可以分为以下 2 个步骤: 寻找可能的候选地名串, 排除串头或串尾的普通词。当大于 3 的字段能够减缩为 2 个或 3 个汉字组成的串时, 调用串长为 2 或 3 的识别过程。下面给出地名识别的例子。

第 1 步: 寻找可能的候选地名串。

假设一个长度大于 3 的单字串表示为 $S = d_1 \cdots d_i \cdots d_n$, 其中包含候选地名为 $S = d_p \cdots d_i \cdots d_q$, 其中每个单字出现的频率应该满足下述不等式:

$$\begin{cases} Fs(d_p) > 0 \\ Fm(d_i) > 0 \qquad (p < i < q) \\ Fe(d_q) > 0 \end{cases}$$

$Fs()$, $Fm()$, $Fe()$分别表示串中第 1 个汉字、中间各汉字以及最后一个汉字作为地名的频率。如果串中任何一个汉字对于上述频率值为 0, 则该汉字就会成为该串的一个分点。例如, 在字段"新康村的王芳"中, 汉字"的"对于 $Fm()$和 $Fe()$的值为 0, 所以"的"就把这个串分为两个部分—"新康村"和"王芳"。

第 2 步: 排除一个字段中串头或串尾的普通词。

为了排除字段当中不属于未登录词的普通单字词, 每个单字词可能作为未登录词的频率以不同频级而出现在不同集合中。这些频级目前分为 K 级, 这样, 对于任何给定的汉字, 就存在一个它作为某类未登录词的概率级别。一个基本约束为: 一个汉字作为未登录词候选字出现在一个未登录词词串的某个位置, 其出现在相应位置的频级必须大于它作为普通词出现在该位置的频级。设函数 $Fw(d)$表示一个汉字 d 作为普通词的频率, $K()$表示频级, 则作为地名的新候选 $S = d_p \cdots d_i \cdots d_q$ 应该满足下述公式:

$$\begin{cases} K(Fs(d_p)) >= K(Fw(d_p)) \\ K(Fe(d_q)) >= K(Fw(d_q)) \end{cases}$$

这样, 寻找一个地名候选字段的过程就是按照上述约束从串两端逐渐向中间靠拢的过程。其他类型的未登录词识别也类似于此。

经过上述步骤处理以后, 还剩下一些未确定的分词形式包括某些未登录词的识别可以在词性标注过程中一起来确定, 这就是分词词性标注一体化的处理方法。许多交集歧义可以通过词性标注过程加以解决。上述分词过程的结果是一个可能包含歧义的有向图, 对有向图的遍历将得到不同分词结果的多条路径。在这些路径中, 词性标注将选择一条其分词结果的词性组合具有最大概率的路径。其词性标注算法采用动态规划算法, 其中对未登录词的规划过程将最佳切分路径分为若干较短的路径, 并得到每一个最佳结果。整个处理过程就是计算从

起始到结束的局部代价函数，记录走过的路径并在路径的最后回溯最佳路径。更详细的描述参见文献[6]。

六、实验结果与结论

本文提出的基于多步处理策略的汉语自动分词方法，其目的是提高自动分词的精度，实验表明取得了较好的结果。封闭和开放测试结果参见表3。

表3　封闭和开放测试结果

测试类型	测试集句子数	测试集词数	错误个数（包括分词和词性标注错误）	精度
封闭	2000	19872	122	9.36%
开放	400	3170	33	98.96%
开放	600	11184	174	98.44%

为了进一步改善分词结果，我们将加入规则来解决部分交叉歧义和组合歧义。并且还将尝试解决汉语离合词问题。

汉语自动分词对于任何中文信息处理尤其是大规模真实文本的处理都是一个基本但又困难的任务。分词、词性标注结果的改善需要使用多种大规模的资源，这些资源通常来自于大规模的具有正确分词和词性标注结果的语料库。而这些资源的建立必须花费大量的人力，因此资源的积累和建设对于分词系统是十分重要的。这也是我们今后的努力方向。

[注] 本文研究的实现过程中使用了清华大学计算机系孙茂松老师研制的伪歧义表，特此致谢。

参 考 文 献

[1] 刘继武，赵铁军，刘挺. 词性信息在汉语自动分词中的应用. 见：‘99智能计算机接口与应用进展. 北京：电子工业出版社，1999，147－150

[2] Richard Sproat *et al*. A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics，1996，22(3)：377－404

[3] Kok-Wee Gan *et al*. A statistically emergent approach for language processing：application to modeling context effects in ambiguous Chinese word boundary perception. Computational Linguistics，1996，22(4)：531－553

[4] 沈达阳. 基于统计和规则的汉语真实文本自动分词和词性标注系统的研究与实现[硕士学位论文]. 北京：清华大学，1996

[5] 孙茂松，左正平，邹嘉彦. 高频最大交集型歧义切分字段在汉语自动分词中的作用. 中文信息学报，1999：13(1)：27－34

[6] 吕雅娟等. 基于分解与动态规划策略的汉语未登录词识别. 中文信息学报，2001，15(1)

[7] 孙茂松，黄昌宁等. 中文姓名的自动识别. 中文信息学报，1995，9(2)

[8] 中国社会科学院语言文字应用研究所. 姓氏人名用字分析统计. 北京：语文出版社，1991