实用型日语自动分词系统的算法及初步实现

陈群秀　孙勇　陈利人　莫林峰
（清华大学计算机科学与技术系）

摘　要：本文提出了一个实用日语自动分词系统的算法。在系统实现中我们使用了伪双向切分方法，并注重解决自动分词领域存在的三个问题：①分词系统中对分词词典中未登录词的处理；②分词词典规模与分词速度的协调；③分词过程中的歧义问题。

关键词：分词，伪双向切分，邻接表

Algorithm and Implementation of a Pragmatic and Automatic Japanese Word-Segmentation System

Chen Qunxiu　Sun Yong　Chen Liren　Mo Linfeng
(Department of Computer Science and Technology, TsingHua University)

ABSTRACT : This paper presents the algorithm and implementation of a pragmatic and automatic Japanese word-segmentation system. In this system we use the approach of pseudo-bidirection segmentation. We focus our attention on three key problems existing in automatic word segmentation system: treatment of unlisted words in the lexicons, trade-off between large lexicon and fast segmentation process, ambiguous word boundary determination.

Keywords : word-segmentation, psuedo-bidirection segmentation,adjacency-table

一、引言

与汉语相似，日语句子也是采用连续书写的形式，词和词之间没有间隔标志，用计算机对日语进行理解或翻译时，就必须先将日语字符串序列切分为词的序列。因此日语自动分词系统的研究是进行日语语料研究、日语自动翻译、日语理解的基础性和关键性的课题之一。而当前的研究状况为每一个有关日语处理的系统均是从日语自动分词开始研究，这样，每个项目均需投入大量的人力、物力、财力，同时也影响了系统的开发进程，基于以上原因，我们感到有必要研究一个能适用于不同领域，满足不同用户需要的可对未登录词进行某些处理的实用型日语自动分词系统。

目前，自动分词的算法普遍采用字符串匹配的原理，方法有最大匹配法（ MM ）、逆向最大匹配法（ RMM ）、有穷多层次列举法，联想回溯法（ AB ）、双向扫描法、邻接表约束法、最佳匹配法，正向带回溯法最大匹配法、基于词频统计的方法等等。

在上述切分算法中，正向最大匹配法（ MM 法）是最基本的分词方法，其思想为：假设自动分词中最长词是 i 个字，则取当前字符串列中的前 i 个字作为匹配字段去查找词典，若匹配不成功，则甩掉匹配字段最后一个字（变为 i -1 个字），剩余的字作为新匹配的字段再进行匹配，如此下去，直至匹配成功为止。此方法实现简单，但错分率较高，一

般辅以其他方法配合使用。

而逆向最大匹配法的分词过程与 MM 法相同，只是扫描方向相反。处理时从句子末尾开始，匹配不成功时甩掉句中位置靠前的那个字。根据统计， RMM 法的分词精度比 MM 法要高一些，但要求配置逆序分词词典。而且日语中很多单词具有形变，其词尾变化非常复杂，逆向匹配较为困难。

二、伪双向切分方法以及邻接表的构成

综合正向、逆向最大匹配法分词的优点，我们在系统中采用了伪双向切分方法，其基本思想如下：

切分开始时先采用正向最大匹配法；当待切分字串中包括有生词，正向切分不能继续下去时，则采用伪逆向的方法，从待切分字串的最后向前逆向切分。设“ $a_1\ a_2 \cdots\cdots a_i\ a_{i+1} \cdots\cdots a_j\ a_{j+1}\ a_{j+2} \cdots\cdots a_m$ ”为待切分字串，正向切分至 a_{i-1}，而从 a_i 开始是一个在词典中查不到的未登录词。此时开始逆向切分，对“$\cdots\cdots a_m$ ”的最长词字长的待匹配字段不断甩掉头一个字，直至剩余部分是一个单词，在此过程中仍然可以采用正序词典，所以称为伪逆向切分。我们期望正向切分和伪逆向切分“夹出”中间的生词；当逆向切分进行到 a_j 停止时，暂定 $a_i\ a_{i+1} \cdots\cdots a_{j-1}$ 为一新词，并给出标记，同时根据下述邻接表提供的信息推导出该词的品词信息。

我们知道，日语词中存在大量的邻接歧义和结合歧义（包括组合歧义、包孕歧义和交集歧义），因此在使用伪双向切分方法进行切分的过程中同时采用邻接表对切分结果进行检验，以发现和纠正由此导致的错误切分。首先对多词性进行检验来解决词性错误问题；若不是词性错误则是单词切分错误，因为采用了最大匹配算法，造成错误切分的原因基本上是长词覆盖了短词，因此回溯时把前一个切分单词长词变短词，再进行邻接关系的检查。

在切分系统中使用了邻接表，其基本构造是一张矩阵，它含有左邻接属性和右邻接属性，表示的是词与词之间词性的邻接关系。邻接表的示例如下：

日语单词邻接关系表

		右邻接属性					
		1	2	3	4	…	
左邻接属性	1	1	0	0	0	…	
	2	0	0	1	0	…	
	3	0	0	1	0	…	
	4	1	0	0	1	…	
	…	…	…	…	…	…	
							…

其中：1，2，3，……为邻接属性的代号；矩阵中的 0，1 为是否可邻接的标志。

每个单词有左、右邻接属性各一个，当两个单词相邻时，以前一个词的右邻接属性和后一个词的左邻接属性为坐标，在邻接表矩阵中对应一个值，若为 1 表示这两个单词的当前词性可以邻接，若为 0 表示它们不能邻接。

三、一种能提高切分速度、压缩存储空间的词典设计

我们确定分词词典存储的信息有两项：词条和邻接属性。

为了方便用户管理，根据词的应用领域不同，系统将词典分为七类：包括基本词典、外来语词典、专用词典、姓名词典、地名词典、缩略语词典以及用户临时词典。它们采用定长格式，有利于用户对词典的删减，扩充，修改等操作。而在系统运行时，首先对七部词典进行归并，形成带索引词典的排序的压缩结构。

压缩词典包括词条与相应的信息两部分内容，它采用首字索引，同首字的词条按序排列，因此结构紧凑，便于查找。

例如：

方（法 X 面 X ……向 X）；其中，X 表示相应词条所带信息。

索引词典中相应存放每个首字在压缩词典中开始的位置。

经过如上处理，所有首字相同的词条首字无需保存，因此可以节约大量空间。

我们对所使用的 KJDD 日英汉操作系统中日文汉字的内码进行了研究，发现日文汉字的内码由两个字节构成，高字节的最高位为 1，这是为了与 ASCII 码区分，低字节的最高位为 0，这是为了与中文汉字的内码区别。因此我们把信息部分也用两个字节表示，高字节置为 00H，低字节用来存放邻接分类信息，这样，词条信息和邻接分类信息就可以区分开来。因此我们可以采用词条连续存放方式，并且具有不同的信息的词条就只存放一次，而信息可有多个。

另外，我们知道日文字内码的高位字节从 A4H 开始，而低位字节从 28H 到 7FH，因此，若某词条 W 的首字内码的两字节为 X1 和 X2，令：

$$i = (X1-A4H)*(7FH-28H)+(X2-28H) \quad \cdots\cdots\cdots(1)$$

则可由 i 直接得到词条在索引文件中的位置，而索引文件中存放的是每个首字对应在压缩词典中的开始位置，那么也就得到了匹配开始的位置。

由于采用连续存放，各词条实际长度不一，现有的二分法检索不能直接使用，我们对此进行了改进，在每次二分法求中点后，向前调整若干字节，使中点稍微偏移，取在某一词条的起始位置。这种改进后的二分法与原二分法相比，检索速度稍慢，但是由于能节省大量空间，把整个词典组织在内存，检索速度相对反而提高。

词典检索算法先确定给定词条在压缩词典中的位置，然后以同样首字开始的词全部读入内存，形成内部小词典，用改进的二分法查找。

四、带绝对切分标志的快速最大正向匹配算法及实现

在切分过程开始，先根据日语中的绝对切分标志将日语句子切分为词群，这样，就可以在较短的词群上进行切分。日语中的绝对切分标志有英文、数字、标点符号、日语格助词を以及连续的片假名串等。

基于词典的机械分词采用字符串匹配方式。由于最小匹配法分词精度低，故一般采用最大匹配法。最大匹配法可采用增字和减字两种算法，但每一增字（或减字）过程都需要重复类似的字符串匹配工作，效率较低。在匹配过程中，作为匹配成功的中间结果（词条）和最终结果不仅应该处于同一内部词典中，而且应该相距不远。直接匹配分词算法正是利用这一点，快速地实现分词过程。

设当前待切分的日语字符串序列为 C = C1C2C3 ……,对于分词词典中某词条 W = W1W2W3 …… Wk,若 Ci = Wi, i =1,2,…… r ，其中 r ≤ k，则称词条 W 与日语字符串 C 的匹配数为 r。若 r = k，则称 W 与 C 完全匹配。

快速最大正向匹配算法的关键是先找到与待切分序列的匹配计数大于或等于 2 的词条，然后在其附近寻找匹配计数最大且与待切分序列完全匹配的词条，显然，它就是按最大匹配原理切分出来的词。

以下是上述切分过程的伪码算法：

```
PROCEDURE 快速最大正向匹配分词（D， C）
  /*对于待切分日语字符串序列 C = C1C2C3 ……，在分词词典 D 中查找与其完全匹
  配且有最大匹配计数的词条，变量 MAXCOUNT 中保存当前最大匹配计数值， P 和 S
  指向词典中词条。*/
BEGIN
  按式(1)计算以 C1 为首字的内部词典 DC1；
  在词典 DC1 中，用词典检索算法检索次首字为 C2 的词条 W；
  IF (DC1 中不存在次首字为 C2 的词条 W) THEN
      IF (C1 在词典中是单字词) THEN 返回 C1 是单字词;
      ELSE C1 是未登录词或未登录词的一部分;
  }
  P =词条 W 的位置;
  S =NULL；
  MAXCOUNT =2;
  DO {
      IF ( P 所指词条与 C 完全匹配 ) THEN { S =P； }
      COUNT = P 所指词条的匹配计数;
      IF (MAXCOUNT ≤ COUNT) THEN {
          MAXCOUNT =COUNT;
          调整 P 至下一词条的起始位置;
      }
      ELSE {
          IF (S !=NULL) THEN { S 所指词条为切分结果，返回; }
          ELSE BREAK；    //退出 DO 循环
      }
  } WHILE (1);
  P =词条 W 的位置;
  DO {
      调整 P 至前一词条起始位置;
      IF ( P 所指词条与 C 完全匹配) THEN { P 所指词条即为切分结果，返回; }
      COUNT = P 所指词条的匹配计数;
  } UNTIL ( COUNT =1 )
  IF (C1 在词典中是单字词) THEN 返回 C1 是单字词;
  ELSE C1 是未登录词或未登录词的一部分;
END/* END OF PROCEDURE */
```

这一算法在具体实现时，还要考虑边界情况。在一个正排序的词典中，对待切分的日语字符串 C = C1C2C3 ……，满足：

1. 内部词典 DC1 中所有词条的匹配计数值大于等于 1；

2. 对于内部词典 DC1 中每一个词条进行匹配计数时，应有且有一个递增，然后递减的过程。

3. 满足完全匹配的词条一定在匹配计数值递增的那一段中。

五、词尾表的构成与作用

在日语中，动词、形容词等均有一个形变的问题，我们没有在词典中存放词的各种形变，因为这样将占用大于的存储空间，而且影响查找速度，索引只在词典中存放动词、形容词、形容动词等的各种词干，然后总结出一个词尾表，这样，得到一个词的词性后，通过查词尾表，便可得到这个词的各种变化形式，以进行切分和词形还原。

例如：

く尾动词的词尾表为：{か，き，く，け，こ，い，つ}
す尾动词的词尾表为：{さ，し，す，せ，そ}
つ尾动词的词尾表为：{た，ち，つ，て，と}
る尾动词的词尾表为：{ら，り，る，れ，ろ}
一段动词的词尾表为：{る，れ，ろ，よ}
形容词的词尾表为：{かろ，く，かつ，い，けれ}

等共 14 个词尾表。

六、未登录词的处理策略

未登录词就是指在系统的分词词典中不曾记录的词。语言是一个开放性大系统，不仅原有词汇量浩瀚，而且语言不段发展，随着科学的进步、社会环境的转变、生活习惯的变化，新词将会层出不穷，而由于现有计算机的容量、速度的限制，词典只能包含尽可能多的词，而不可能穷尽所有的词，特别是人名、地名、机构名、新的科技术语等。因此，未登录词的出现是必然的。

我们系统中对未登录词的处理策略有：

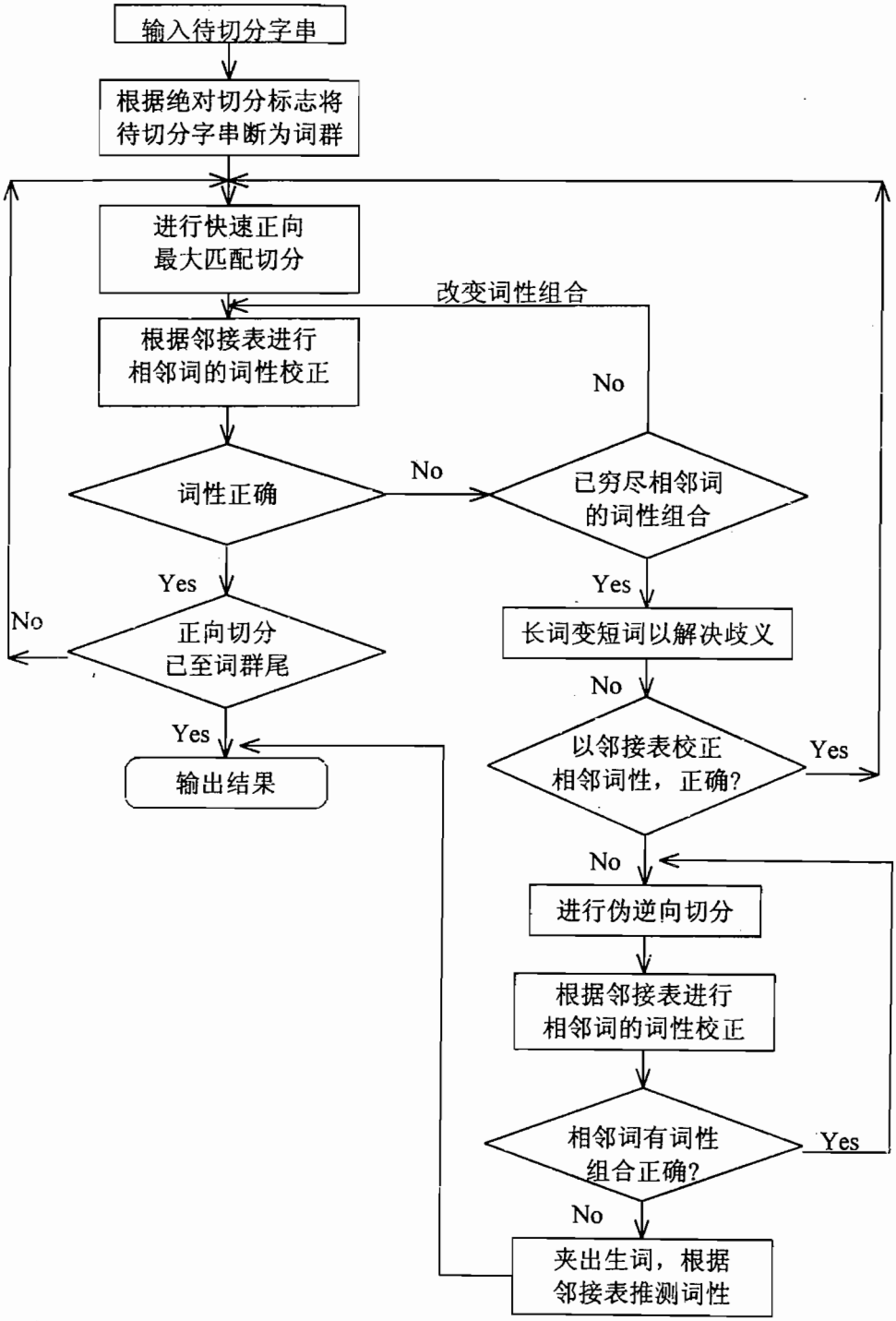
1. 建立人名词典、地名词典、缩略语词典和专业词典；记录一部分常在各个领域出现的词汇。

2. 建立用户临时词典；不同的用户可根据自己研究的领域，将该领域出现的新词存入用户词典中，这样，经过不断的积累，用户临时词典可以做到比较完备，从而大大提高切分的正确率。

3. 全自动切分时，采用伪双向切分的方法；如“二、”中所述，伪双向切分可以解决一定的未登录词问题，它的基本出发点是，一个待切分字串中可能出现一到两个未登录词，为了不影响系统的效率，也为了提高系统的切分正确率，系统不会由于某个词无法切分而中断系统的进程，使以后的词都无法切分。采用伪双向切分时，如果两个方向的切分都遇到了生词而导致夹出的生词块（可能仅为一个生词，也可能两头为生词，但中间包括了可切分的单词）较大，说明可能有两个以上的生词。此时从待切分字串的前端逐一甩掉头字，直至正向切分能继续进行。甩出的字暂时合并为词，推测其词性（一般定为名词、动词或形容词）并以邻接表进行检验，之后给出标志，以待用户进行切分后的编辑。

4. 人工干预的方法。

切分系统的流程为：


输入待切分字串
根据绝对切分标志将
待切分字串断为词群
进行快速正向
最大匹配切分
改变词性组合
根据邻接表进行
相邻词的词性校正
No
词性正确
No
已穷尽相邻词
的词性组合
Yes
Yes
No
正向切分
已至词群尾
长词变短词以解决歧义
No
Yes
以邻接表校正
相邻词性，正确?
Yes
输出结果
No
进行伪逆向切分
根据邻接表进行
相邻词的词性校正
相邻词有词性
组合正确?
Yes
No
夹出生词，根据
邻接表推测词性

日语分词系统流程图

七、分词系统中 CACHE 技术的应用

在进行一篇文章的切分时，我们发现其中有些词的出现频度是非常高的，因此，我们在系统内部设立了一个固定长度的内部缓冲区。它可以按定长的格式存放 100 个词条，词条的信息包括源词条项、出现频率与它在压缩词典中的位置。根据高频词条在压缩词典中的位置，就可以不必查这些词的索引和对次首字进行改进的二分查找，因此能节省一定的系统运行时间。

对它的具体操作描述如下：

1. 当缓冲区存放的词条数不足 100 时，若此时有一未在缓冲区的新词，将它插入缓冲区的最后，并且距离它在压缩词典中的位置和初始化它的出现频率。若该词在缓冲区中出现，则将它的出现频率加一。

2. 当缓冲区中存放的词条数大于 100 时，若此时有一未在缓冲区的新词，则选择一个出现频率最小的词，在这个位置存入该新词。这样，我们将基本上使缓冲区中出现的词为在一篇文章中出现的高频词。

参考文献

[1]. 陈利人："实用型日语自动分词系统的研究和初步实现"，毕业设计论文，1993 年 6 月。

[2]. 陈群秀："国内汉语自动分词研究进展"，《计算机世界》，1992 年 4 月。

[3]. 张国煊、王小华、周必水："汉语自动分词系统研究"，《计算机世界》，1992 年 16 期。

[4]. 马晏、黄昌宁："一种基于评价函数的汉语分词算法"，全国人工智能与智能机学术会议（NJCACTI'91）论文，1991 年 9 月。

[5]. 王启祥、王锡江等："日汉机器翻译中词的自动切分技术"，《中文信息学报》，第二卷第三期，1988 年 9 月。

[6]. 白拴虎："汉语词切分及词性自动标注一体化方法"，《计算语言学进展与应用》，1995 年 9 月。

[7]. 候敏、孙建军、陈肇雄："汉语自动分词中的歧义问题"，《计算语言学进展与应用》，1995 年 9 月。

[8]. 揭春雨、刘源、梁南元："论汉语自动分词方法"，《中文信息学报》，第三卷第一期，1989 年 3 月。

[9]. 李国臣、刘开瑛、张永奎："汉语自动分词及歧义组合结构的处理"，《中文信息学报》，1988 年。

[10]. 张普、张光汉："现代汉语'有穷多层列举'自动分词方法的讨论"，《语言和计算机》，第三辑。