鞍山师范学院学报
Journal of Anshan Normal University
2008-04, 10(2): 42—45

正向最大匹配法在中文分词技术中的应用

胡锡衡
(鞍山师范学院 数学系, 辽宁 鞍山 114007)

摘 要: 分词是中文信息处理的一部分, 分词本身并不是目的, 而是后续处理过程的必要阶段, 是中文信息处理的基础技术. 正向最大匹配法是一种基于词典的分词方法, 它能够有效地实现对中文文档的扫描, 将文档分解成为词的集合. 从而实现中文文本结构化的表示.

关键词: 信息处理; 分词; 文档; 正向最大匹配; 文本结构化
中图分类号: TP391 文献标识码: A 文章篇号: 1008-2441(2008)02-0042-04

为了要进行中文的计算机处理, 首先必须把文档中的词与词分割开, 然后提取对过滤操作贡献大的词语并计算每个词在文本中重要的程度, 即进行特征提取和权重计算.

1 分词的概念

分词只是中文信息处理的一部分, 分词本身并不是目的, 而是后续处理过程的必要阶段, 是中文信息处理的基础技术[1]. 中文信息处理的是非结构化的自然语言文本, 汉语的书写是以汉字作为基础, 词与词之间没有明显的形态界限, 要进行中文的计算机处理, 首先要把词与词分割开来, 即分词. 通过对文档的扫描, 将文档分解成为词的集合. 这也是中文文本结构化表示的前提.

在印欧语系语言中, 词与词之间有空格作为固定的分隔符, 一般不存在分词问题. 在词汇数量上, 一般的印欧语种的词汇量最多为几十万词, 而汉语的词汇量高达几百万甚至上千万. 一个汉字序列可能有几种不同的切分结果, 产生歧义现象. 这些都给自动分词带来了极大的困难.

分词单位的选取一般以 1988年我国制定的《信息处理现代汉语分词规范》为准, 但还要考虑具体应用环境以及大规模语料库处理的特殊要求[2]. 因此, 分词的原则是:

(1)分词单位的选取必须有利于标注等后续过程的处理. 如"二分之一"、"五月一日"这样的词组, 按照分词规范规定: "构成分词单位的词组必须具备结合紧密的特征, 而分开后不改变原有组合意义的词组, 则一律加以切分", 但实际上, 这些词组在具体的上下文环境中常常合起来表达一个意思, 作为一个分词单位更能符合后续处理的需要, 这样就没有必要在分词阶段把它们分开, 然后又要在后边的分析阶段花费精力把它们合在一起.

(2)分词准确率是分词系统最重要的性能指标.

(3)为处理大规模的语料, 要求系统有较好的容错能力. 另外, 分词词典要有良好的可扩充性, 具备从语料库中自动学习的能力.

(4)分词系统还要有较好的可移植性.

2 分词的主要方法

现有的分词方法大体可以分为 3类: 基于词典的分词方法、基于理解的分词方法和基于统计的分词

收稿日期: 2007—10—10
作者简介: 胡锡衡 (1973—), 男, 辽宁鞍山人, 鞍山师范学院数学系讲师.

方法[3].

2.1 基于词典的分词法

这是一种应用最广泛的机械分词法, 依据一个分词词典和一个基本的切分评估原则, 目前最常用的规则是最长匹配原则. 它基于一个简单的思想: 一个正确的分词结果应该由合法的词组成, 这些词在当前待切分的句子中, 并且属于词典中的一个词. 分词过程, 按照扫描的方向不同, 匹配分词方法可以分为正向匹配和逆向匹配; 按照不同长度优先匹配的情况, 分为最长匹配和最短匹配; 按照是否与词性标注过程相结合, 可分为单纯分词方法和分词与标注相结合的一体化方法.

2.2 基于理解的分词法

分词系统最困难的就是如何消除歧义. 消除歧义需要很多额外的信息, 如句法、语义等, 而这些信息不是一部简单的词典能解决的. 基于理解的分词系统不仅要有好的词典, 而且还要加上句法和语义分析. 通过获得有关词、句子等的句法和语义信息来对分词歧义进行判断从而模拟人类对句子的理解过程. 如基于潜在语义索引的方法, 这种分词方法需要使用大量的语言知识和信息. 由于汉语语言知识的笼统和复杂性, 难以将各种信息组织成机器可直接读取的形式, 因此目前基于理解的分词系统还处于试验阶段.

2.3 基于统计的分词法

从形式上看, 词是稳定的字的组合, 因此在上下文中, 相邻的字同时出现的次数越多, 就越有可能构成一个词[4]. 字与字相邻出现的频率或概率能够较好的反映词的可信度. 对语料中相邻出现的字的组合的频度进行统计, 计算他们的互现信息. 定义两个字的互现信息为:

$$M(X,Y) = \log\left(\frac{P(X,Y)}{P(X)P(Y)}\right),$$

其中, $P(X,Y)$是汉字 X和 Y的相邻共现概率, $P(X)$和 $P(Y)$分别是 X和 Y在语料中出现的概率. 互现信息体现了汉字之间结合关系的紧密程度. 当紧密程度高于某一个阀值时, 便可认为此字组可能构成了一个词. 如 n元分词法, 这种方法只需对语料中的字组频度进行统计, 不需要切分词典, 因而又叫无词典分词法或统计分词法. 这种方法也有局限性, 分词中经常会抽出一些出现频率高, 但并不是词的常用字组, 如“这一”、“我的”、“之一”等, 并且对常用词的识别精度差, 时空开销大.

针对基于理解的分词法与基于统计的分词法的不足, 采用一种基于词典的分词方法 正向最大匹配法来进行中文信息文档的分词操作, 其通过顺序扫描字符数组中每一存储单元, 达到快速准确的切分文档的目标.

正向最大匹配法的基本思想是: 假设自动分词词典中的最长词条所含汉字个数为 MaxLan, 则取被处理材料当前字符串中的 MaxLan个字作为匹配字段, 查找分词词典. 若分词词典中有这样的一个 MaxLan个字的词, 则匹配成功, 匹配字段作为一个词被切分出来; 若在分词词典中找不到, 则匹配失败, 匹配字段去掉最后一个字, 剩下的作为新的匹配字段, 进行新的匹配, 如此进行下去, 直至切分成功为止. 然后再按上面的步骤进行下去, 直到切分出所有词为止.

例如, C1 =“面向对象课程是五十一个课时”, C2 =“　”, 设定最大词长 MaxLan= 5. 分词词典中有: 面向、对象、课程、课时等词语. C1分词过程如图 1所示.

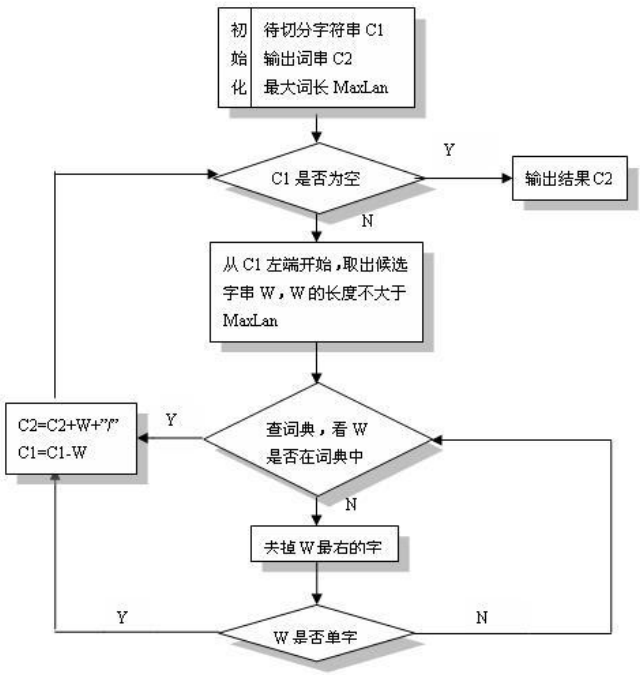

图 1　C1分词流程

(1)C2= “ ”; C不为空, 从 C1左边取出候选子串 W=“面向对象课”;

(2)查词表, “面向对象课”不在词表中, 将 W最右边一个字去掉, 得到 W=“面向对象”;

(3)W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“面向对”;

(4) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“面向”

(5) W不为单字, 查词表, W在词表中, 将 W加入到 C2中, C2=“面向 /” 并将 W从 C1中去掉, 此时 C1=“对象课程是五十一个课时”;

(6)C1不为空, 于是从 C1左边取出候选子串 W=“ 对象课程是”;

(7) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“对象课程”;

(8) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“对象课”;

(9) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“对象”;

(10)W不为单字, 查词表, W在词表中, 将 W加入到 C2中, C2=“面向 /对象”, 并将 W从 C1中去掉, 此时 C1=“课程是五十一个课时”;

(11) C1不为空, 于是从 C1左边取出候选子串 W=“ 课程是五十”;

(12) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“课程是五”;

(13) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“课程是”;

(14) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“课程”;

(15) W不为单字, 查词表, W在词表中, 将 W加入到 C2中, C2=“面向 /对象 /课程”, 并将 W从 C1中去掉, 此时 C1=“是五十一个课时”;

(16) C1不为空, 于是从 C1左边取出候选子串 W=“ 是五十一个”;

(17) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“是五十一”;

(18) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“是五十”;

(19) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“是五十”;

(20) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“是五”;

(21) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“是”;

(22) W为单字, 将 W加入到 C2中, C2=“面向 /对象 /课程 /是”, 并将 W从 C1中去掉, 此时 C1=“五十一个课时”;

(23) C1不为空, 于是从 C1左边取出候选子串 W=“ 五十一个课”;

(24) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“五十一个”;

(25) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“五十一”;

(26) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“五十”;

(27) W不为单字, 查词表, W在词表中, 将 W加入到 C2中, C2=“面向 /对象 /课程 /是 /五十”, 并将 W从 C1中去掉, 此时 C1=“一个课时”;

(28) C1不为空, 于是从 C1左边取出候选子串 W=“ 一个课时”;

(29) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“一个课时”;

(30) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“一个课”;

(31) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“一个”;

(32) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“一”;

(33) W为单字, 将 W加入到 C2中, C2=“面向 /对象 /课程 /是 /五十 /一”, 并将 W从 C1中去掉, 此时 C1=“个课时”;

(34) C1不为空, 于是从 C1左边取出候选子串 W=“ 个课时”;

(35) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“个课”;

(36) W不为单字, 查词表, W不在词表中, 将 W最右边一个字去掉, 得到 W=“个”;

(37) W为单字, 将 W加入到 C2中, C2=“面向 /对象 /课程 /是 /五十 /一个”, 并将 W从 C1中去掉, 此时 C1=“课时”;

（38）C1不为空, 于是从 C1左边取出候选子串 W=“ 课时 ”;

（39）W不为单字, 查词表, W在词表中, 将 W加入到 C2中, C2=“面向 /对象 /课程 /是 /五十 /一 /个 /课时 ”, 并将 W从 C1中去掉, 此时 C1=“ ”;

（40）C1为空, 输出 C2=“面向 /对象 /课程 /是 /五十 /一 /个 /课时 ”;

为解决分词歧义的问题, 可以用最大匹配法进行扩展: 增加歧义词表, 规则等知识库. 对于某些交集型歧义, 可以通过增加回溯机制来改进最大匹配法的分词结果. 据统计最大匹配法错误切分率为 1/169. 目前最大匹配法作为一种基本的方法被肯定下来.

3 结束语

影响分词精度的主要因素是歧义切分问题, 而解决歧义切分的主要方法是利用构词知识[5]. 但是由于知识库中知识的增多, 知识之间的相互影响会增大, 主要是知识处理的错误, 因为每一条知识都有副作用, 反过来又降低了分词精度; 再者由于知识的增多, 严重影响了分词的速度, 因此, 不能增加过多的知识.

自动分词是依据分词词典进行的, 分词词典中没有的词系统是切分不出来的. 在大规模真实文本的处理过程中, 会遇到许多不能由词典识别的词汇, 包括人名、地名、时间、术语等, 这些词总称为未登录词. 未登录词现象往往成为影响分词系统准确率的主要因素. 但不能因此而盲目地增大分词词典, 因为这样会增加歧义字段, 如增加“国是 ”词, 就会增加“中国是 ”一类的歧义字段, 被方位词单独构词知识错误地切分成 “中 /国是 ”, 而影响分词精度. 因此, 在系统设计过程中, 考虑建立根据 “标准分词词典 ”来进行分词, 这样可能会减少一些由于分词词典中收词不标准或不足而造成的错误.

参考文献:

[1] 唐培丽, 胡明. 基于中文文本主题提取的分词方法研究[J]. 吉林工程技术师范学院学报, 2005, (2): 23—26.
[2] 孙茂松, 左正平. 汉语自动分词词典机制的实验研究[J]. 中文信息学报, 2000, (1): 58—61.
[3] 梁南元. 书面汉语自动分词系统-CDWS[J]. 中文信息学报, 1998, (2): 82—86.
[4] 黄昌宁. 中文信息处理中的分词问题[J]. 语言文字应用, 1997, (1): 76—79.
[5] 刘开瑛. 歧义切分与专有名词识别软件[J]. 语言文字应用, 2001, (3): 41—44.

Application of Maximum Matching Method in Chinese Segmentation Technology

HU Xi-heng
(Department of Mathematics, Anshan Normal University, Anshan, Liaoning 114007, China)

Abstract: Segmentation is a part of Chinese information processing. It's not the aim, but the necessary stage of follow-up processing. It's the basic technology of Chinese information processing. The MM(Maximum Matching Method) is a method based on dictionary and can scan Chinese document effectively and decompose the document into collections of words. Thus Chinese structured text is achieved.

Key words: Information Processing; Segmentation; Documentation; MM(Maximum Matching); Structured text

（责任编辑: 张冬冬）