TP18

基于后加词典利用句法语义知识的汉语词切分检纠错方法

杨 抒 伊 波

（南京大学计算机软件研究所）

摘 要

本文通过分析现有词典匹配汉语词切分法及相应切分错误检出与纠正方法的现状及不足，提出了一种基于后加词典，利用句法语义知识的汉语词切分检纠错方法，这种方法旨在将词切分作为汉语理解的有机组成部分，使得检纠切分错误更加有效，同时，利用后加词典，提高了词切分出错后重新切分的效率。

一、 引 言

汉语的最大特点之一是：句子由一串汉字线性排列起来，没有区分词的明显标记。而词是最小的能够独立活动的有意义的语言成份[1]，自然语言处理的各个领域，包括自然语言理解，机器翻译，文献检索等，都要以词作为基本语言单位，而汉语处理中一个重要的任务是词切分，即从由连续汉字串组成的句子中切分出逐个的词来。

现有的词切分方法很多，如[2][3][4]文中所提到的逐词遍历法，最大匹配法，逆向最大匹配法，运用切分标记法，正向最佳匹配法，逆向最佳匹配法等，这些方法本质上都是用一个词典去匹配待处理材料，故统称之词典匹配法。上述各词典匹配法的主要不同在于：词典的不同，指词典的构造和所包含的信息不同；处理文段的方向不同，即正向的或逆向的；匹配对象的截取方式不同，即减字方式或增字方式；确定初始结果的手段不同，即最大匹配或最小匹配[4]。如前述逐词遍历法，词典中的词由长到短顺序存放，称为顺序存放式；最大匹配法，词典采用首字索引式结构，处理文段的方向为正向，匹配对象的截取方式为减字方式，初始结果取最大匹配；逆向最大匹配法，与最大匹配法的不同在于处理文段的方向为逆向；正向和逆向最佳匹配法的词典是按照词频大小的顺序排列词条；切分标记法则在词典中的某些词上增加了切分标记信息。上述各法针对词典匹配法中各个因素的不同考虑，都是为了提高切分的效率和精度，由于后三种因素的可选择情况有限，因此，词典的设计就最为关键，事实上，词典的结构直接影响到切分算法，因而对切分效率影响最大。

上述方法本质上都是词典匹配法，因此，都可能产生切分错误，从而，基于词典匹配法进行词切分，必然要涉及如何检出和纠正切分错误的问题。在[2][5]中谈到两种检错和纠错方法，然而，这两种方法并不能圆满解决这个问题，特别是，这两种方法主要用于词频统计上，而本文希望研制一种作为汉语理解的有机组成部分的汉语词切分方法，现有方法就更显得不适宜了。下面，我们针对现有检纠切分错误方法之不足，提出

在汉语理解中利用词法，句法，语义等不同的语言知识来检出切分错误的思想，并根据错误切分源只可能是可重切词的特点，设计了后加词典，以提高纠错效率。在实现部分，主要介绍了后加词典的结构，并比较了这种词典与首字索引式词典的切分和纠错复杂性。

二、句法语义分析协同汉语词切分的检错与纠错方法

引言中提到，基于词典匹配法的汉语词切分，可能会引起切分错误，因而，必须进行检错与纠错。现有一些检纠错方法的思想是：首先，找出具有多种切分的句子或字段，再利用特定规则或词条知识来判断正确切分。比如，[5]中提出的切分出所有形式解方法，此法首先假定词典是充分的，因而，可以切分出当前句子的所有形式解(详见[5])，若形式解多于一个，则利用“分词知识”(即一些特殊词涉及的多义字段在某些上下文环境中的正确切分规则)来确定一种正确的切分。又如，[2]中提出的寻找交集字串法，根据[2]中分析存在多义切分的两种字段，交集字段和多义组合字段，并列用词尾字构词法找出句子中交段为1的交集字串，再利用“纠错知识”(即作者搜集的一些高频率多义字段的正确切分规则)来确定当前交集字串的正确切分。

上述检纠错方法存在以下问题：

a. 收集检纠错规则很困难，特别是，难以保证所收规则集的充分性，因而难以保证切分的正确性，这势必要影响理解机制的正确理解。

b. 对理解机制来说，切分与句法语义分析相分离，既增加了收集检纠错规则的开销，也浪费了理解机制中大量的句法语义知识。

c. 检纠切分错误时，先找出全部形式解，既增加了切分时的开销，也不符合人们检纠切分错误时的自然过程。

据此，为了使得检纠切分错误更加合理和有效，并适应汉语理解的需要，我们设计了句法语义分析协同汉语词切分的检错与纠错方法，其设计思想是：不设计专门的“切分规则”，而是利用理解机制所具有的词法，句法和语义知识来帮助检查切分错误，或判断切分的正确性。具体办法是，为了使得理解机制更加接近人的理解过程，分析人理解语言的特点，发现人在理解时，仅当遇到错误时，才去纠正，而不是先找出可能的不同理解，从中去确定一种正确理解。这样，我们在词切分时，不专门去找是否可能产生切分错误的字段，而是在词切分或句法语义分析过程中发现了不合法的情形时，再来进一步识别是否为切分错，若是，则纠正之。另一方面，若理解过程中没有发现错误，即说明切分结果“可理解”，那么就认为切分正确。由于在切分时，不必判断其切分结果是否正确，那么，当切分机制接受到一个切分失败的信息时，只需负责给出一种新的切分。

下面给出在词切分和句法语义分析中，利用词法，句法或语义知识判断切分错误的规则：

检错规则1 在切分时，若切分到一个非词字，则认为产生了切分错误。

检错规则2 在句法分析时，若遇到不合语法限制的词，且不存在一个不确定的句法分析路径时，表明不是句法分析的错，这时，亦判为切分错。

检错规则3 在语义分析时，若遇到不合语义限制的词，且不存在一个不确定的句法分析路径时，表明不是句法或语义分析可以纠正的错，这时亦判为切分错。

下面考察依据上述不同检错规则检出切分错误的例子：

a. 切分时，检出切分错误的例。

例：发展中国家兔养殖业。 (误)

发展中国家兔养殖业。 (正)

说明："兔"为非词字，故产生切分错误。

b. 句法分析时，检出切分错的例。

例：我让他从前门走。 (误)

我让他从前门走。 (正)

说明：切分时，先切出第一种情况，但无法发现切分错误，因为每个切分单位均为合法的词。但按照句法规则，"门"的位置出现名词是不合法的。

c. 语义分析时，检出切分错误的例。

例：他学会了解方程。 （误）

他学会了解方程。 （误）

他学会了解方程。 （正）

说明：第一次切分，由于不含语义限制，故重切，第二次切分仍不含语义限制，再重切，第三次切分，得正确切分的句子。

发现了切分错误，下一步就要进行纠正。由于我们利用句法语义分析来判断切分的正确性，因此，切分机制的纠错，仅仅进行一次新的切分，即找出另一种切分结果，而不必保证切分的正确性。重新切分，指从当前发生切分错误的地方回溯，找到最近的可重切词，从它开始重切。若是采用目前已有的词典构造和切分方法，须首先向前逐个退字查找可以重新切分的词，然后，再从它开始进行重新切分。若把可重新切分的词称为可重切词，便可根据可重切词的结构特点，构造一种后加词典，以提高查找可重切词的效率。为此，我们给出一组定义：

定义1 （联接∥）设S_1，S_2，S_3 均为汉字串，若$S_1 \parallel S_2 = S_3$，即表示将两个汉字串S_1，S_2联接成一个更大的汉字串S_3，则称S_3为S_1和S_2的联接，简记为$S_1 \parallel S_2$。

定义2. （子串） 设S，S'均为汉字串，若存在汉字串S_1，S_2，（S_1，S_2可为空串），使得$S_1 \parallel S' \parallel S_2 = S$，则称$S'$为$S$的子串。

定义3 （前缀子串） 设S'，S均为汉字串，若存在汉字串S_1（S_1可为空），使得$S' \parallel S_1 = S$，则S'为S的前缀子串。

定义4 （后加子串） 设S'，S均为汉字串，若存在汉字串S_1（S_1可为空），使得$S_1 \parallel S' = S$，则称S'为S的后加子串。

定义5 （前缀词） 设W_p为一个词，S为一个汉字串，若$W_p \parallel S = W$ 也为一个词，则称W_p为W的前缀词。

定义6 （直接前缀词）设W_p，W是词，W_p是W的前缀词，若对任意S'，S'是S的前缀子串，$S' \neq S$，$W_p \parallel S'$不是词，则称W_p是S的直接前缀词。

定义7 （后加词） 设W为词，S为汉字串，若$W \parallel S = W_s$也为词，则W_s为W的后加词。

定义8 （直接后加词） 设W，W_s是词，若W是W_s的直接前缀词，则W_s是W的直接后加词。

定义9 （可重切词） 若一个词具有前缀词，则这个词为可重切词。

定义10 （错误切分源）引起当前错误切分的第一个词为错误切分源。

命题 词典匹配切分法中，错误切分源只可能出现在可重切词中。

基此，可得如下启示，当发现切分错误时，似可以不必逐词退字前找错误切分源，若能直接判断可重切词，则效率将提高很多。如何能直接判断一个词是否为可重切词呢？

在引言中，我们谈到影响词切分效率的一个重要因素是词典构造，并谈到目前已有的三种构造方式，顺序存放式，按词频存放式，以及首字索引式。顺序存放式效率较低，不常用，词频存放式适用于对特定领域的大词汇量正文进行词频统计，而首字索引式较好地平衡了时空开销，是目前最常用的方法。但无论何法，都无法直接判断可重切词，一种最容易想到的办法是在首字索引式词典中的词上增加是否为可重切词的标记，但这不是最好的办法。这里，我们提出一种后加词典法。利用前面定义的概念，我们刻划后加词典的结构如下：

后加词典包括所有需要的单字（成词或不成词）及非单字的词，所有的单字顺序存放，以每个单字为根，构造一棵树，树上每个结点是一个词或词素，它的儿子是其所有的直接后加词，它的父亲是其直接前缀词（可能是不成词的字），可由图1描述：

其中，$((C_i)S_j)S_k$是$(C_i)S_j$的直接后加词。

很明显，利用这个词典，判断一个词是否为可重切词，只要判断其父结点是否为词即可，同时，可直接取得该可重切词的直接前缀词作为重新切分的结果。

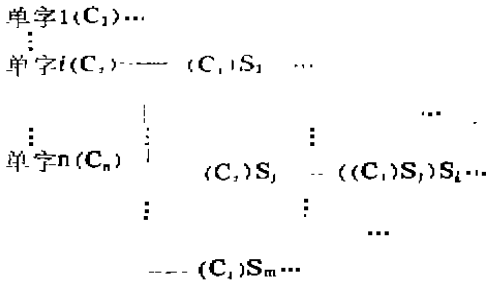

图1 后加词典结构示意图。

下面，再简单谈一下词切分与句法语义分析协同处理时，纠正了切分错误后，如何进行局部句法语义修正。

已有的切分方法，均是完全独立于句法语义分析的，其结果是，有些隐含的切分错误要到句法语义分析阶段才会发现，一旦句法语义分析中发现了切分错误，交给切分机制后，句法语义分析均要重做。由于切分错误没有向后延伸性[2]，纠正切分错误也只会影响一句话的局部，因而，句法语义分析也不必完全重做，可以只做局部修正，以节省开销。为此，我们提出词切分与句法语义协同处理的方法，即每切分一个词，就进行句法语义分析，一旦发现切分错误，一方面，给出重切结果(若存在的话)，另一方面，则要找最小的须进行局部句法语义修正的句子片断，进行局部重新句法语义分析。

确定重分析片断的规则：将包含发现切分错的那个词与当前须重切的可重切词的最小语法树作为重分析片断，若重切后的该句法子树不符合当前的句法语义分析规则，则重分析片断再扩大到包含该重切片断的最小句法树。

鉴于本文重点讨论词切分及其检纠错方法，句法语义分析的详细内容这里就不赘述了。

三、实现技术

1. 逻辑框图

图2给出了句法语义分析机制与词切分机制协同处理的逻辑框图，其中各组成部分

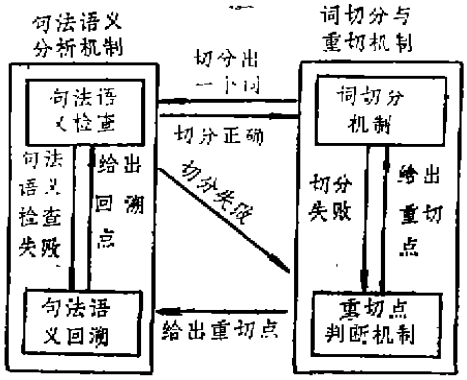

图2 句法语义分析与词切分协同处理逻辑框图

的功能及信息交互为：词切分机制切分出一个词交句法语义分析机制，若切分失败，则通知重切机制找出重切点；重切机制找出重切点通知切分机制重新切分，同时也通知句法语义分析机制找回溯点；句法语义检查机制对当前句子进行句法语义检查，若成功，则表明切分正确，请求切分机制切分下一个词，若失败，则通知回溯机制进行回溯，若不是句法语义检查中的问题，则认为切分错，通知重切点判断机制，请求给出重切点，以便重切和重分析；句法语义回溯机制的任务是找到重新进行句法语义分析的回溯点。

2. 基于后加词典的词切分与重切机制

这里主要讨论后加词典的实现以及基于后加词典和首字索引式词典的词切分和纠错方法的复杂性比较。

后加词典用两个随机文件存放，一个文件存放所有的首字，存取采用Hash树方法[8]，另一个文件的变元是一个存放n个词条的组块，每个首字的所有后加词树对应一个或多个组块，对应于每一个词的所有直接后加词，仍按由长到短的顺序存放。

这样，切分一个词的匹配次数，用后加词典比用首字索引法略少，因为第二层以上的后加词数量较少。而寻找作为当前重切点的可重切词的匹配次数，后加词典明显优于首字索引法，并且，重切点与出错点相距越

(下转13页)

装等，实现程序的自动生成。

4.针对具体应用领域，设计有效的构件，带动通用软件重用的研究。

还有一些其他的相关问题，有待进一步研究，但是我们可以期望，在不久的将来，对于软件重用技术及其相关技术将有较快的发展，将会使软件生产方式产生一次革命。

参 考 文 献

〔CAVA83〕 Michael J. Cavaliere and philip J. cham beault, "Reusable Code at the Hartford Insurance Group" Proceedings of the Workshop on Reusability in programming (September 1983)

〔GOLD83〕 Adele Goldberg and David Robson, Smalltalk-80: The language and its implementation. Addison-wesley (1983).

〔HIRO88〕 Hiroyuki Tarumi, Kiyoshi Agusa, Yutaka Ohno, "A Programming Environment Supporting Reuse of Object-Oriented Software" Proc. Tenth Int'l Software Engineering Conf., (1988).

〔KERN84〕 Brain Kernighan, "The Unix System and Software Reusability." IEEE Transactions on Software Engineering, Vol.SE-10, No5, Sept. (1984).

〔LANE84〕 Robert G. Lanergan and Charles A. Grasso, "Software Engineering with Reusable Designs and Code." 同上.

〔RICE83〕 John R. Rice and Herbert D. Schwetman, "Interface Issues in a Software Ports Technology." Proceedings of the Workshop on Reusability in Programming (September 1983).

〔TED87〕 Ted Biggerstaff and Charles Richter, "Reusability Framework, Assessment, and Directions". Proceedings of the Twentieth Annual Hawaii International Conference on System Sciences, 1987.

〔NEIG83〕 J.M.Neighbors, "The Draco Approach to Constructing Software from Reusable Components." Proc. Workshop Reusability in Programming. ITT. Straitford, Conn. 1983.

（上接44页）

远，优越性越大。

三、小结

本文针对现有汉语词切分词典匹配法中检纠分错误方法的不足，特别是其独立于句法语义分析，难以作为汉语理解的有机组成部分的现状，提出了在汉语理解中，利用词法，句法和语义知识帮助检出切分错误的方法，初步解决了汉语理解与词切分结合起来的问题。同时，设计了一种后加词典，以提高纠错时判断重切点的效率。当然，这种结合句法语义知识检纠切分错误的方法尚待进一步研究，另外，对于遇到新词，计算机如何处理，也是颇有意义的研究课题。本文对汉语理解中的词切分方法，提出了一些初步想法和观点，有些已在计算机上实现。欠妥之处，尚请指正。

致谢：本文工作是南京大学计算机软件研究所研制的汉语会话系统NDTALK的一部分，这项工作始终得到徐家福教授的指导，NDTALK小组成员也提出了有益的建议，在此一并表示感谢！

参考文献

〔1〕 朱德熙，"语法讲义"，商务印书馆，1982.

〔2〕 梁南元，"书面汉语自动分词与一个自动分词系统——CDWS"，北京航空学院学报，1984.4.

〔3〕 梁南元，"书面汉语自动分词综述"，计算机应用与软件，1987.3.

〔4〕 杨春雨，刘源，梁南元，"论汉语自动分词方法"，中文信息学报，1989.1.Vol.3 No.1

〔5〕 曾纪文，谷新英，"结合上下文辅助分词的学习系统"，中文信息处理国际会议论文集，1983. 北京.

〔6〕 伊波，杨抒，"Hash树，——一种词典存放方法."1989.待发表.