汉语词切分及词性自动标注一体化方法*

白栓虎

电子工业部计算机与微电子发展研究中心

摘要： 本文从理论上探讨了汉语词切分和词性标注一体化的语言学模型，并给出了模型方法。在作者实现的系统中文本的处理步骤如下：首先用基于词典的最大匹配法初步切分句子，其次用基于词典的方法从每个多字词里发现切分歧义，最后用统计方法消除切分歧义和词性歧义。通过对 20 万字的语料(每个样本约 2000 汉字)初步的开放测试表明，系统的词条在只有约 25,000 词的情况下切出词的正确率在 99%以上，词性标注的正确率达 96%。

An Integrated Model of Chinese Word Segmentation and Part of Speech tagging

Bai, Shuanhu

Research Center of Computer and Microelectronics
Industry Development, M.E.I

ABSTRACT: In this paper the author studies the theoretic aspects and implementation methods of a statistics based integrated model for Chinese word segmentation and part of speech tagging. The system implemented by the author works by the following steps: First, the segmentation module segment character form Chinese sentences into word form sentences by using a dictionary based maximum-matching approach. Then the results are passed to the segmentation ambiguity detection module to find out ambiguity points by checking each n-gram(n> 1) words. Finally the ambiguity resolution process disambiguates segmentation and part of speech ambiguities based on part of speech cooccurrence probabilities. The primary test on texts of 200,000 Chinese characters (2, 000 Chinese characters for each sample) shows that, when the dictionary contains only 25,000 entries,the overall accuracy for segmentation is about 99%, the accuracy for part of speech tagging is about 96%.

一 序言

对汉语切词技术的研究已有十几个年头了，可以说这是汉语处理中一个永恒的话题。尽管这些年来有诸多的切词系统出现[2-5,9,10]，但使用起来不够理想。除了不同开发者对词有不同的认识，导致互相之间对切分的结果难以接受以外，建造切词系统所采用的方法也有很大的局限性。在过去的几十年里国内建造的汉语切词系统基本上采用了基于规则的方法[2,4]，这些系统虽然取得了较大的成就，但是规则系统的局限性也呈现了出来。基于规则的切词方法一方面需要各种各样大量的切分知识，另一方面也缺乏足够的灵活性。在海外有人使用纯统计方法建造切词系统[5]，这种系统除了需要占用较多的资源外，在效率和精度上也难以达到满意的效果。近年来也有一些结合规则和统计方法的切词系统出现[3,9,10]。汉语词性标注系统的出现可以说把汉语处理工作向前推进了一步，出现过基于统计[1]和基于规则[7]的系统。也有一些研究人员利用词性标记来校正切分歧义[3]，提供了汉语切词的另一条思路。一些利用分词和词性自动标注一体化方法的系统[6]也没有解决好

* 本项目由国家自然科学基金青年科学基金资助

分词和标注有机结合的问题。

汉语词切分要解决的重要问题之一是消除切分歧义。所有的切分歧义中能够用语法知识解决的约占90%以上，而涉及到语义和语用知识则很少[2]。在建造词性标注系统中，我们已经获得了汉语词类与词类同现的频度，同时也获得了一定规模的词条上带有词性标记的典，这都是汉语处理的宝贵语法资源。若能充分利用这些语法资源来有效地解决汉语切词中一些问题，那么必将减轻专门为汉语切词系统建造大量资源的负担，同时也可减小系统的整体规模。这正是我们的出发点。

针对这个目标，在后面几节中叙述的内容如下：第二节介绍基于统计的词切分和标注一体化模型和实现方法，第三节讨论切分歧义发现的技术，第四结给出切分标注的实例，第五节对全文作出总结。

二　基于统计的切词和标注一体化模型及实现方法

当把切词和标注一体化以后，输入的是汉字字符串，输出的结果则是带有词性标记的汉语词串。在处理过程中要充分利用词性标注的资源，来消除切分歧义。

令 $C=c_1c_2\cdots c_m$ 是输入的由 m 个汉字字符组成的汉字字符串，$W=w_1w_2\cdots w_n$ 是把 C 切分后得到的由 n 个词组成的词串，其中 w_1 和 w_n 是两个没有切分和词类歧义的词(如标点)。$T=t_1t_2\cdots t_n$ 是对 W 进行标注后所得的一个标记串。我们希望 W 是 C 的正确切分，同时 T 又是 W 的正确标注。我们先考虑切词部分。P(W|C)是在给定输入字串 C 的条件下所产生的输出词串 W 的概率。如果用最小错误率决策，我们的目的就是寻找这样的词串 W'，使得：

$$P(W'|C)=\max_{W}P(W|C) \tag{1}$$

根据贝叶斯公式，我们可以有如下结果：

$$P(W|C)=\frac{P(W)P(C|W)}{P(C)} \tag{2}$$

在(2)式中，C 是给定的，P(C)是一个确定的值，在计算中不起作用。而 P(C|W) 是在给定词串的情况下字串出现的概率，可以认为是1。所以我们可以得出如下结论：

$$P(W|C)\approx\max_{W}P(W) \tag{3}$$

这就表明，基于统计的词切分过程，可以认为是寻找具有最大概率值的词串的过程，并且可用词的 n 元语法模型来实现。词的 n 元语法模型看起来容易，但存在的固有的缺陷是，即当词汇量较大的情况下训练语料不足所带来的数据稀疏问题，随之可能导致切词系统“领域转移困难”，或出现“倾向性”问题。我们认为，语言的用词具有很大的灵活性，而语言的语法结构具有相对的稳定性，所以尽管词的分布会随着领域的变化而变化，词类在句子中的分布具有比较普遍的意义，不会象词的分布那样变化明显。词类的这种相对稳定分布特征是有效的资源。让我们还是把 P(W)放到标注模型中来考察，找出词的出现和词类的关系。根据贝叶斯公式，有

$$P(W)=\frac{P(T)P(W|T)}{P(T|W)} \tag{4}$$

从(4)式可以看出，其分子代表了词性标注的统计模型。如果一个句子在切词的时候没有歧义，那么简单地对其进行标注即可。对有切分歧义的句子来说，就不那么简单。对(3)和(4)式用二元语法模型化简，用下式计算：

$$P(W)\approx\max P(t_1)\prod_i\frac{P(t_i|t_{i-1})P(w_i|t_i)}{P(t_i|w_i)}\qquad i=2,\ldots,n \tag{5}$$

在计算(5)式时，可对选定的一条切词路径(这时就限定了词条信息的参数)，用词性标注的模型计算具有最大概率的词性标记串，这时也就得到了对应的标记路径．然后再对经过这样选择的每一条切词路径所对应的标记串的概率除以 $\prod P(t_i|w_i)$，就得到了切词路径的概率，最后选择具有最大概率的切词路经的词串．此时所选择的词串所对应的词性标记串已经选择出来了．这样词性标注就成为切词的附产品，大大地节省了计算工作量．在这里选择 i 从 2 开始是因为我们已经在前面限定了 W_1 和 W_n 是两个没有切分和词类歧义的词．

三　切分及发现切分歧义

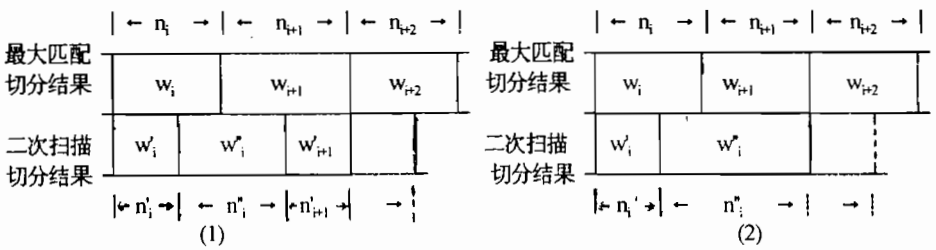
以上我们对基于统计的汉语切词和词性标注模型作了一些介绍．模型的核心是用定量的手段去消除切分歧义和词类歧义，其主要内容是计算一种切分或标注的概率．不过，切词的核心问题不仅仅是消除切分歧义，如何发现切分歧义也是比较重要的问题之一．

习惯上人们把切分歧义的两种类型叫交集型歧义和组合型歧义．简而言之，交集型歧义就是切分字段中出现词和词交错的现象，组合型歧义就是切分字段中出现词中含词的现象．如果采用基于词表的匹配方法，所有这些歧义都是由词表的构成引起的．词表是汉语词法知识最直接的资源，展示了汉语的构词规则，支持汉语切词的相当一部分资源就是词表的词法知识．

汉语词切分方法中有两种极端情况，一种是正向或逆向最大匹配法，另一种是全切分．前者是用文本中的汉字串到词表中匹配最长的词串，这种方法掩盖了所有的切分歧义；后者是用文本中的汉字串匹配词典中所有可能成词的片段，这种方法确实发现了所有可能的切分歧义，但同时也会产生许多“切分垃圾”．绝大多数的切分系统还是采用了二者折中的办法来尽可能地发现歧义，提高切分效率．最大匹配法会掩盖切分歧义，而能否从全切分算法产生的众多的切分结果中挑选出正确的结果也是值得怀疑的．理想的办法是通过实现多种切分算法，来比较每种方法的优劣．根据一些初步的统计，汉语的交集型歧义约占 84%，而组合型歧义仅占 16%[2]，相对较少．处理好交集型歧义，就解决了切分的大多数歧义了．在我们的系统中采用了不同的方法，来分别发现这两种不同类型的歧义．整个系统首先采用最大匹配法对文本中的句子初步切分，然后使用一些歧义发现的策略来发现切分歧义．

1. 交集型歧义的发现

发现交集型歧义的过程大致上是这样的．首先对一个汉字串用最大匹配法切分，然后逐词扫描寻找双字及其以上的词．当寻找到以后对该多字词从第二个字开始逐字用最大匹配法匹配词库中的词，方法如图一．



图一　发现交集型歧义的两种情况

在图一中 w 是词，其上下标代表了其顺序。和 w 具有相同上下标的 n 表示其长度。假设 w_i 是多字词，把 w_i 从第二个字开始逐字经向后最大匹配法切分后切成 w'_i 和 w''_i 两个字串，其结果分为三种情况:

(a) w''_i 未跨越 w_i，此时 $n'_i+n''_i < n_i$。在这种情况下没有交集型歧义。

(b) w''_i 跨越了 w_i 且未跨越 w_{i+1}，此时 $n_i < n'_i+n''_i < n_i+n_{i+1}$，如上图(1)。在这种情况下如果 w'_i 是词表中的词，则从 w''_i 的末尾开始再向前看一步进行最大匹配切分，切出 w'_{i+1}。此时判断有歧义的标准是 w'_{i+1} 达到或跨越了 w_{i+1} 的边界。

(c) w''_i 跨越 w_i 并达到或超过了 w_{i+1} 的边界，此时 $n'_i+n''_i \geq n_i+n_{i+1}$，如上图(2)。在这种情况下判断有歧义的标准是 w'_i 是词表中的词。

下列表一中例子是从处理过的语料中抽取的句子，我们分别来说明各种情况。

原句子	他开门见山地讲解.	
最大匹配切分结果	他 开门见山 地 讲解 .	
假歧义	他 开门见 山地 讲解 .	
	(1)	
原句子	他已清清楚楚地表明了他的态度.	
最大匹配切分结果	他 已 清清楚楚 地表 明了 他 的 态度 .	
二次扫描切分结果	他 已 清清楚楚 地 表明 了 他 的 态度 .	
	(2)	
原句子	有机会见面.	作对称性试验.
最大匹配切分结果	有机 会见 面 .	作对 称 性 试验 .
二次扫描切分结果	有 机会 见面 .	作 对称 性 试验 .
	有机 会 见面 .	
	(3)	(4)

表一 几个句子的最大匹配切分和发现歧义后的结果

在表一(1)中虽然“山地”是一个词，但是“开门见”不是词，我们不认为此句有歧义。在表一(2)中“表明”跨越了“地表”和“明了”两个词，属图一中(1)。根据(b) 我们判断“地”是词，同时用最大匹配法向前匹配“了他的态度”，结果只匹配出词“了”，所以可以断定在表一(2)中“地表”一词中从“表”开始存在长度为两字的歧义结构。同样，表一(3)属于图一(1)，是根据(b)判断的结果；表一(4)属于图一(2)，是根据(c)判断的结果。

上述三种情况作为判断交集型歧义的标准，始终贯穿了最大匹配法。从现在看来在所处理的语料中还没有由算法引起的遗漏歧义的现象。

2. 组合型歧义的发现

相对于交集型歧义而言，组合型歧义出现的机会较少，但这是一种不可预测的歧义类型。虽然有一些词象“不是”、“就是”等引起切分歧义比较明显，绝大多数组合歧义的出现与处理的语料有密切的关系。很难想象由“位置”这个词可引起组合型歧义，在处理计算机领域文本时且会出现，例如“表示正数时高位置 0，表示负数时高位置 1。”中由“位置”引起的组合歧义。全切分可以彻底地发现这类歧义，但过多切分结果的干扰因素引起错误的机会可能会更多。

我们解决这种问题的办法是在词上加标记位与别的词加以区别。这种方法的最大局限性在于不可能穷举这类词，造成有些组合型歧义被遗漏的现象。经常遇到的引起组合型歧

义的词有:

不是　就是　个人　从来　将来　将就　去向　所在　现在　有的　一头
一次　一时　一下　一行　最好　人为　有为　有数　难为　会见　才能
人选　与其　不如

表二　几个常引起组合型切分歧义的词

另外，绝大多数复合词会引起组合型歧义，如“研究所”、“机器翻译”等词。

从上面的切分过程来看，用向前最大匹配法初步切分，二次扫描用最大匹配发现交集型歧义和用标记位法发现组合歧义，其结果可以是一棵树或一个图，消除歧义只不过是对树的剪枝或对图寻找从起点到终点的最佳路径。

四　消除歧义实例

在分词和词性自动标注一体化的系统中，消除歧义有两个方面的内容。其一是切分歧义的消除，其二是词性歧义的消除。在前面我们已经提到，切分过的句子可以用一棵树来表达。在这棵树中从根结点到叶子结点的每一条由词组成的路径都代表一种切分结果。切分歧义消除的结果是找出一条从根结点到叶子结点的词的路径，词性歧义消除的结果是找出对应于句子当中词的词类序列。由(5)知，我们对树中的每一条从根结点到叶子结点的路径用词性自动标注的办法(如动态规划或 Forward-Backward 算法)来寻找一个切分结果所对应的最佳或候选标记串，然后根据(5)来计算每一条切分路径的评价值。

对词性歧义的消除方法在[1]已经作过介绍，在我们采用的分词和标注一体化模型中是对每一个切分结果进行一次词性歧义消除，这里不再作专门介绍。这里我们把句子“结合成分子时，”的处理过程作一分析来说明整个流程。

1* 结合(vg vgn vgo)　成分(ng)　子时(t qt ng dr)　，
2* 结合(vg vgn vgo)　成(vc vgn dr vgo mab)　分子(ng)　时(t qt ng dr)　，
3* 结(vgn vgo vg)　合成(vgo vg vgn)　分子(ng)　时(t qt ng dr)　，

1# 结合(vg<普通动词>)　成分(ng<普通名词>)　子时(t<时间词>)　，
2# 结合(vg<普通动词>)　成(vc<补语动词>)　分子(ng<普通名词>)　时(t<时间词>)　，
3# 结(vg<普通动词>)　合成(vgn<动词带体宾>)　分子(ng<普通名词>)　时(t<时间词>)　，

表三　“结合成分子时，”经切分和歧义发现后的结果

表三中每个词后的圆括号里的符号是词类代码。在整个过程中，系统运行的步骤如下:

1) 用最大匹配法切出结果 1*;
2) 用歧义法现策略切出结果 2*, 3*;
3) 用词类歧义消除策略分别对 1*, 2*, 3*消除词类歧义，得 1#, 2#, 3#;
4) 用公式(5)来消除切分歧义，最后得到的结果是 2#。

最初的切分歧义消除的办法是对每一个输入的句子产生一个输出词的序列，这种办法具有彻底地消除歧义的优点。尽管这种方法能达到很低的错误率，可是仍然存在给出错误的切分结果而丢失了正确的切分结果的情况。这种错误对句法分析或其他高层次的处理来说往往是致命的。

从词性自动标注产生多个候选结果的技术得到启示，在词切分时也可能产生多于一个的结果，并且可以给出每个结果的评价值。初步的统计结果表明，如果切分结果输出增加

13%左右。切分错误率将降低 7%。

五 结束语

本文主要描述了基于词表的词切分和基于统计的切分和词类歧义消除的模型及其实现方法。初步的试验表明，结合词性标注的方法和资源进行汉语词切分是成功的，用基于词表的切分和基于统计的歧义消除策略有较高的效率，同时也极大地提高了系统的灵活性。通过对实现的模型的初步测试，可得出如下结论:

1) 词性自动标注的资源可以用来消除切分歧义，汉语词切分和词性自动标注的方法可以结合使用提高系统的整体性能。消除切分歧义时采用总体最优的原则，这种方法可以比较容易地应用于其它切分方法之上。

2) 实践证明我们采用的机制是简单实用而效率较高的。在我们的系统中词类同现频率矩阵占用很小的空间(目前只有 113 × 113 × 2 字节)，在 486-50 机 windows 环境下处理速度约 120 字/秒，执行代码(包括训练和维护工具)不到 75K 字节。切分正确率按词计算(在切分过的语料中正确的词与切出的总的词数之比)在 99%以上，而词性标注的正确率在 96%以上。 同时在我们的系统中引入有效的切分和标注生词识别的机制，将有另文介绍。

3) 现有的资源和手段决定了在切词和词性自动标注时仍然有一些问题不可能彻底解决，在一定范围内的输出多项选择是提高系统灵活性的重要手段。

参考文献

[1] 白栓虎，夏莹，黄昌宁，“汉语语料库词性自动标注方法研究”，《机器翻译研究进展》，陈兆雄主编、电子工业出版社，1992 年 8 月.

[2] 何克抗，徐辉，孙波，“书面汉语自动分词专家系统设计原理”，《中文信息学报》，第 5 卷第 2 期，1991 年.

[3] 黎邦洋等，“一种主要使用语料库标记进行歧义校正的、最大匹配汉语自动分词策略”，Proceedings of ROCLING IV, pp135-146，台湾，1991.

[4] 梁南元，“再论汉语自动分词和切分知识”，中文信息处理国际会议论文集，1987 年 8 月，北京.

[5] 张俊盛等，“限制式满足及机率最佳化的中文断词方法”，Proceedings of ROCLING IV，pp147-165，台湾，1991.

[6] 周强，俞士汶，“一种切词和词性标注相融合的汉语语料库多级加工法”，《计算语言学研究与应用》，陈力为主编，北京语言学院出版社，1993 年十月.

[7] 周莉娜，郑家恒，刘开瑛，“汉语词类标注规则的获取技术”，《计算语言学研究与应用》，陈力为主编，北京语言学院出版社，1993 年十月.

[8] Meteer, et al, POST: Using Probabilities in Language Processing, IJCAI'91, P960-965.

[9] Nie, Jian-Yun, Jin, Wanying and Hannan, Marie-Louise,Ahybrid approach to unknown word detection and segmentation of Chinese, Proceedings of International Conference on Chinese Computing' 94(ICCC94), 1-4 June 1994, Singapore.

[10] Sun, M.S, Lai, T.B.Y, Lun, S.C. and Sun,C.F. , Some Issueson the Statistical Approach to Chinese Word Idetification, Proc. of 3rd International Conference on Chinese Information Processing, pp. 246-253, 1992.