用于文本校对的分词与词性标注一体化算法

王永景,刘功申,李生红,荆 涛
(上海交通大学 电子工程系,上海 200240)

摘 要:分词和词性标注是中文处理中的一项基本步骤,其性能的好坏很大程度上影响了中文处理的效果。传统上人们使用基于词典的机械分词法,但是,在文本校对处理中的文本错误会恶化这种方法的结果,使之后的查错和纠错就建立在一个不正确的基础上。文中试探着寻找一种适用于文本校对处理的分词和词性标注算法。提出了全切分和一体化标注的思想。试验证明,该算法除了具有较高的正确率和召回率之外,还能够很好地抑制文本错误给分词和词性标注带来的影响。

关键词:文本校对;分词;词性标注;一体化算法
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2008)08-0001-03

One Combined Approach of Chinese Segment and Tagging for Proofreading

WANG Yong-jing, LIU Gong-shen, LI Sheng-hong, JING Tao
(School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Segment and part-of-speech tagging is two important procedures in Chinese processing. Use machine segment based on dictionary traditionally, but during the process of proofreading the errors in the input texts would deteriorate the result of segment and tagging, and then the errors' detection and correction would be made on base of the inexact output. In the paper, tried to find a method suitable for proofreading, and a combined of automatic segment and tagging approach was proposed, which was proved effective to minimize the influence of the errors with a high precise and callback rate.

Key words: automatic proofreading; automatic segment; tagging; combined approach

0 引 言

中文自动校对是近几年兴起的一个研究课题。已经被应用在出版业、语音输入、汉字识别、文本编辑、辅助教学等领域[1]。

就目前现有的与中文校对相关的文献来看,国内在自动文本查错方面主要采用三种方法:

①利用文本上下文的字、词和词性等局部语言特征、包括词性特征、同现特征或相互依存特征,甚至包括字形特征等;

②利用转移概率对相邻词间的接续关系进行分析[2];

③利用规则或语言学知识,如语法规则、词搭配规则等[3]。

1 现有算法存在的问题

(1) 现有大部分算法都需要事先对要处理的文本进行分词和词性标注,但是却没有考虑文本校对系统的输入文本的特殊性,即输入的并不是规范文本,而是含有各种各样的错误文本。

例如"今天天气不错",它可能被写作"近天天气不错",前向最大长度分词算法就会将其切分为"近|天天|气|不错"。

(2) 现在大部分的分词和词性标注都是分开进行的,但我们觉得词性标注对分词具有指导作用。而且这两个操作在过程上也是相关的,因此在算法中将它们结合在了一起。

故此提出了一种能够尽量减轻文本错误对分词和词性标注产生的影响的一体化算法。

收稿日期:2007-11-28
基金项目:国家自然科学基金资助项目(60402019, 60502032);教育部新世纪优秀人才支持计划项目(NCET-06-0393)
作者简介:王永景(1982-),男,江苏徐州人,硕士研究生,研究方向为自然语言理解、文本自动校对;刘功申,副教授,研究方向为内容安全、舆情分析、恶意代码防范;李生红,教授,研究方向为网络安全、计算机病毒、内容过滤;荆 涛,副教授,研究方向为信息安全、计算机通信网。

2 分词算法

根据是否利用机器可读词典和统计信息，可将汉语自动分词分为三大类：基于词典的方法、基于统计的方法和混合的分词方法。基于词典的分词方法的三个要素为分词词典、文本扫描顺序和匹配原则。文本的扫描顺序有正向扫描、逆向扫描和双向扫描。基于统计的分词方法所应用的主要的统计量或统计模型有：互信息、HMM模型和N元文法模型等。混合的分词算法使用了不止一种模型，例如王锡江提出了一种基于邻接知识的汉语自动分词系统[4]。赵铁军提出了一种提高汉语自动分词精度的多步处理策略[5]。

此处采用了基于词典的“全切分”的方法，为了适应一体化算法的需要，对其进行了改进。分词基于两个原则：

(1)保留尽量多的切分结果，以此来保证正确的切分结果包含在切分结果集合中。其基本思想是：根据词典，找出字串中所有可能的词，构成词语切分有向无环图。每个词对应图中的一条有向边，并赋给相应的边长(权值)。然后针对该切分图，在起点到终点的所有路径中求出长度值按严格升序排列依次为第1，第2，…，第 i，第 N 的路径集合作为相应的粗分结果集。如果两条或者两条以上的路径长度相等，那么它们的长度并列第 i，都要列入粗分结果集，而且不影响其它路径的排列序号，最后的粗分结果集合大小大于或等于 N。

例如：“当原子结合成分子时”可能的切分结果为“当|原子|结合|成分|子时”、“当|原子|结|合成|分子|时”、“当|原子|结合|成|分子|时”、“当|原子|结|合成|分|子时”、“当|原子|结|合|成|分子|时”、“当|原子|结|合|成|分|子时”等。

(2)合理地保留尽可能少的分词结果，来降低整个算法的复杂度。为此保留分词长度比较小的那些结果。上面的例子中只保留前四条结果。

3 词性标注算法

常用的词性标注模型有N元模型、隐马尔可夫模型[6]、最大熵模型[7]、基于缓存的模型、基于触发的模型、基于决策树的模型等等。其中，隐马尔可夫模型作为一种简单而有效的数学模型，在自然语言处理、语音识别、生物信息学很多领域得到了广泛的应用。这里就采用了隐马尔可夫模型(HMM)原理来进行词性标注。

现在假设 W 是分词后的词序列，T 是 W 某个可能的词性标注序列，其中 T^* 为最终的标注结果，即概率最大的词性序列，则有

$W = \{w_1, w_2, \cdots, w_m\}$，$T = \{t_1, t_2, \cdots, t_m\}$，$m > 0$，$T^* = \arg\max_T P(T \mid W)$

根据Bays公式，有

$$P(T \mid W) = P(T)P(W \mid T)/P(W) \quad (1)$$

对于一个特定的词序列来说，$P(W)$ 是一个常数，得到：

$$T^* = \arg\max_T P(T)P(W \mid T) \quad (2)$$

引入HMM来计算 $P(T)P(W \mid T)$，得

$$P(T)P(W \mid T) \approx \prod_{i=1}^{m} p(w_i \mid t_i)p(t_i \mid t_{i-1}) \quad (3)$$

故

$$T^* = \arg\max_T \prod_{i=1}^{m} p(w_i \mid t_i)p(t_i \mid t_{i-1}) \quad (4)$$

在大规模熟语料库中，根据大数定理，可以得到：$p(w_i \mid t_i) \approx C(w_i, t_i)/C(t_i)$，其中 $C(w_i, t_i)$ 表示 w_i 的词性为 t_i 时出现的次数；$C(t_i)$ 表示词性 t_i 出现的次数。

$$p(t_i \mid t_{i-1}) \approx C(t_i, t_{i-1})/C(t_{i-1}) \quad (5)$$

其中 $C(t_i, t_{i-1})$ 表示词性 t_{i-1} 到下一个词性为 t_i 的次数，$C(t_{i-1})$ 为词性 t_{i-1} 出现的次数。$C(w_i, t_i)$、$C(t_i)$、$C(w_i, t_i)$ 均可以通过对已经切分标注好的熟语料库的统计得到。

4 分词与词性标注一体化算法

如图1所示，将全切分用图1表示，每个词下面是此词对应的词性，每一条路径都是标注的一条结果，问题转化为有向无回图的最优路径问题。对含有错误的输入文本来讲，定义“正确的标注”为：分词和词性标注结果中除了出错的字词以外，其他部分都标注(包括分词)正确。为了尽可能使正确的结果被包括进去，函数取最优的 n 条路径，n 可以由用户根据需要指定。

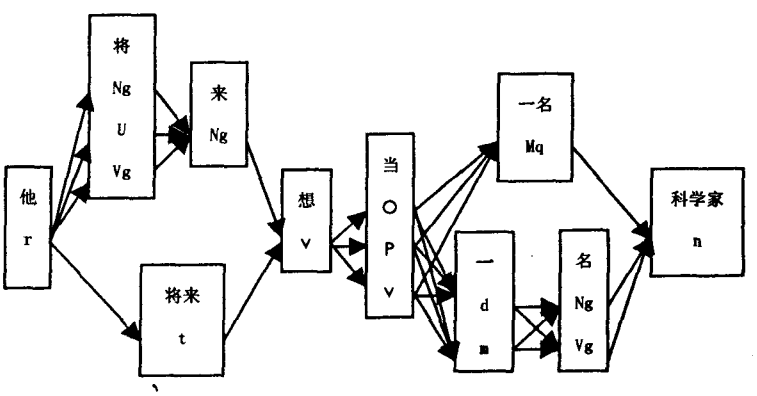

图1 全切分图

在这个算法中最重要的就是如何计算路径的权

重,设共有 N 个标注结果,分词结果为 $S_i = \{W_{i1}W_{i2}\cdots W_{ij-1}W_{ij}W_{ij+1}\cdots W_{iLi-1}W_{iLi}\}, i = 1,\cdots,N$。

对应的词性标注结果为 $T_i = \{t_{i1}t_{i2}\cdots t_{ij-1}t_{ij}t_{ij+1}\cdots t_{iLi-1}t_{iLi}\}, i = 1,\cdots,N$,其中 L_i 为标注的第 i 条结果中词的个数。现假设正确的标注结果为 T_k。假设词 W_{ij} 为第 i 个分词结果中的出错词。

(1) 词性的独立度。

定义:词性的独立度 Dep_{ij} 是指该词在语料库中以词性 t_{ij} 出现的概率,可以用该词以该词性出现的次数 $\mathrm{freq}t_{ij}$ 与该词出现的次数 $\mathrm{freq}w_{ij}$ 的比值计算。进行简单的数据平滑后的公式为:

$$\mathrm{Dep}_{ij} = \mathrm{freq}t_{ij} + 1/(\mathrm{freq}w_{ij} + 1), j = 1,\cdots,L_i \tag{6}$$

其中 freq_{ij} 表示该词在语料库中出现的次数,$\mathrm{freq}_{\mathrm{long}ij}$ 表示包含该词的所有长词出现的总次数。

(2) 归一化因子。

由于词数的增多会一定程度上增加路径权重,所以定义一个归一化因子

$$\lambda_i = \mathrm{sum}_{\min}/\mathrm{sum}_i \tag{7}$$

其中 $\mathrm{sum}_{\min}$ 为分词结果中最小的词数,sum_i 为该条结果的词数。

(3) 词性的一阶转移概率。

经过简单数据平滑后的转移概率公式为:

$$p(t_{ij=1} \mid t_{ij}) = \frac{\mathrm{freq}(t_{ij=1}, t_{ij}) + 1}{\mathrm{freq}(t_{ij}) + 1} \tag{8}$$

总的路径的权重函数为

$$\mathrm{Weight}_i = -\lambda_i \ln\left(\prod_{j=1}^{Li}\mathrm{Dep}_{ij}\prod_{j=2}^{Li}p(t_{ij=1} \mid t_{ij})\right) \tag{9}$$

算法选择总权重最小的 n 条路径。

5 实验结果

选择熟语料库中的7 617条记录,并对其中的100条记录318 714个词进行测试,得到如表1中的结果:

表1　分词词性标注一体化算法标注结果

	输入总词数	标注总词数	标注正确词数	正确率	召回率
分词结果	318714	328779	297840	0.905897	0.934506
词性标注结果	318714	328779	275282	0.837286	0.863727

通过表格可以发现,此算法最优结果具有不错的准确率和召回率,而后又选择常见的读音相似错误、字型相近错误、缺字错误、多字错误作为输入,发现最优的五条输出结果中包含正确标注结果的几率达到了71.6%,例如“今天天气不错”按照一体化算法最优的5条结果为:

今天-t | 天气-n | 不-d | 错-v
今天-t | 天气-n | 不错-a
今-r | 天天-d | 气-v | 不错-a
今-r | 天天-n | 气-v | 不错-a
今-t | 天天-d | 气-v | 不错-a

如果输入为“近天天气不错”最优的5条结果为:

近-a | 天-n | 天气-n | 不-d | 错-v
近-j | 天-n | 天气-n | 不-d | 错-v
近-nr | 天-nr | 天气-n | 不-d | 错-v
近-j | 天-j | 天气-n | 不-d | 错-v
近-v | 天-n | 天气-n | 不-d | 错-v

通过上面的例子,可以看出:错误的输入并没有对切分造成“污染”,使结果变为“近|天天|气|不错”。

6 结束语

文本纠错已经成为如今各种中文处理过程中很重要的一个方面,但是,之前提到一些分词和词性标注算法,没有考虑错误输入对处理带来的影响,不适合应用于文本纠错,所以提出了一种专门用于文本校对的分词和词性标注一体化的算法。通过试验,本算法对引入的字词错误,具有很好的驽钝性,不会扩散影响前后部分的处理。而且第一选择的正确率和召回率也很理想。

但是同时也看到,本算法相比其他算法,时间复杂度和空间复杂度都加大了,之后要做的就是优化该算法,使其满足实时信息处理的要求。

参考文献:

[1] 张　磊,周　明,黄昌宁,等. 中文文本自动校对[J]. 语言文字应用,2001(2):19-26.
[2] 张仰森,丁冰青. 基于二元接续关系检查的字词级自动查错方法[J]. 中文信息学报,2001(3):36-43.
[3] 张仰森,丁冰青. 中文文本自动校对技术的现状及展望[J]. 计算机应用研究,2006(6):8-12.
[4] 王锡江,王启祥,陈家俊. 基于邻接知识的汉语自动分词系统[J]. 计算机研究与发展, 1992, 29(11):54-58.
[5] 赵铁军. 提高汉语自动分词精度的多步处理策略[J]. 中文信息学报, 2001,11(1):23-26.
[6] 梁以敏,黄德根. 基于完全二阶隐马尔可夫模型的汉语词性标注[J]. 计算机工程, 2005, 31(10):177-179.
[7] 林　红,苑春法,郭树军. 基于最大熵方法的汉语词性标注[J]. 计算机应用, 2004,24(1):14-16.