结合 CRFs 的词典分词法①

张硕果 汪成亮 (重庆大学 计算机学院 重庆 400044)

摘 要: 逐字分词法是以汉语词典为基础对中文语句通过匹配进行切分的方法。该方法在分词中无法解决交叉歧义与组合歧义带来的问题。本文以词典分词为基础,从序列标注的角度,在逐字匹配过程中使用 CRFs 标注模型提供辅助决策,由此来处理歧义问题。经实验和分析,该方法较传统的 CRFs 模型分词法和词典分词,更适合对分词速率及正确率都有一定要求的系统。

关键词: 条件随机场; 分词; 交叉歧义; 组合歧义; 逐字匹配

Dictionary Chinese Word Segmentation Method Combined with CRFs

ZHANG Shuo-Guo, WAGNG Cheng-Liang

(Department of Computer, Chongqing University, Chongqing 400044, China)

Abstract: The Chinese Segmentation of matching literal based on Dictionary can not resolve the problem of segmenting ambiguousness and Combinatorial ambiguity. Based on the dictionary segmentation, this paper propose a method of Dictionary Chinese Word Segmentation combined with CRFs. It is proved that this method can have better performance than CRFs segmentation and traditional dictionary segmentation.

Keywords: CRFs; word segmentation; segmenting ambiguousness; combinatorial ambiguity; matching literal

中文自动分词是中文自动处理技术的基础。成功的分词方法需具有较高的准确性与快速的切分能力。其中前者需解决歧义切分、未登录词识别等问题,目前主要采用字符串频度统计、统计语言模型等机器学习方法;而快速切分则依靠设计高效的词典分词系统来实现。

经过对两者的分析,本文提出了一种结合字符标注的词典分词方案,来实现一种既有较快速度又有较高准确性的分词系统。第 1 部分介绍了常用的词典分词系统机制并分析其不足。第 2 部分简要描述条件随机场模型(CRFs)。第 3 部分介绍作者提出的分词方案。第 4 部分通过分析和实验说明新方案与传统的分词法相比,更适应于对速率和正确率都有要求的系统。第 5 部分提出总结与展望。

1 词典分词系统

1.1 分词机制

词典分词系统一般具有两个要素:词典构造的机制和词汇匹配的方法[1]。

本文分别采用了基于 TRIE 索引树[2]的词典机制与逐字的匹配方法。其中,TRIE 索引树是一种以树的多重链表形式表示的键树,基于 TRIE 索引树的词典机制由首字散列表和索引树节点两部分组成(如图 1 所示)。

逐字匹配法逐字读取当前处理语料中的字符,沿树链进行逐字匹配,直到构成一个完整词汇。

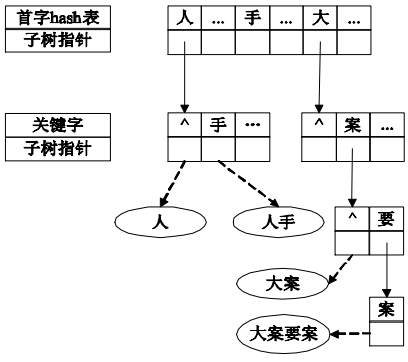


图 1 Trie 树结构图

① 基金项目:国家 863 项目(2007AA12Z306)
收稿时间:2010-03-10;收到修改稿时间:2010-04-12

1.2 机制分析

在当前语料与词典进行匹配的过程中，逐字分词法存在切分歧义的问题。而从构成形态上划分，切分歧义一般又分为交叉歧义和组合歧义[3]：

组合歧义：

例句 s_1 人手上有伤

正：　　人 / 手 / 上 / 有 / 伤

误：　　人手 / 上 / 有 / 伤

交叉歧义：

例句 s2 组合成分子时

正：　　结合 / 成 / 分子 / 时

误：　　结合 / 成分 / 子时

传统的逐字匹配方法不具备处理上述两种歧义的机制，可能对分词结果产生较大影响，因此有必要采用一些方法来解决该问题。

考虑从逐字标注的角度来看待匹配过程。定义标注集合{B,I}：B 表示多字词的首字或单字词；I 表示多字词中除首字以外的字符。比如,对例句：数码照相机的外观。读入首字‘数’，在 TRIE 树链中逐字匹配发现该字符可作为词汇首字，即可得出其标记为 B。再读入‘码’，查找到其出现在‘数’的子节点中，即可得出其标记为 I，依次下去，得出：

B I　　B I I　　B　　B I

数码 / 照相机 / 的 / 外观

由于逐字匹配无法处理歧义，从标注的角度来看，当发生歧义时，由该方法得出的标记就有可能出现错误。比如例句 s_2 中‘分’字可能得出‘B’或‘I’两种标记，若得出其为‘I’，结果出错。

反之，若能在词典匹配过程中，从标注的角度对字符标记进行正确判断，就能为逐字匹配提供辅助的决策，提高分词正确率。本文将采用 CRFs 标注模型来实现这一功能。

2 条件随机场[4](CRFs)

2.1 CRFs 定义

CRFs 是一种无向图模型。常用作中文分词法。令 $x=(x_1, x_2, \cdots\cdots, x_n)$ 表示待标注序列，$y=(y_1, y_2, \cdots\cdots, y_n)$表示对应状态序列。其状态序列的条件概率为：

$$P_q(y|x)=\frac{1}{Z(x)}\exp\left(\sum_t\sum_k l_k s_k(y_{t-1},y_t,x,t)+\sum_t\sum_k m_k g_k(y_t,x,t)\right) \quad (1)$$

上式中 Z(X)为归一化因子，$S_k(y_{t-1},y_t,x,t)$是表示转移特征的(0,1)二值函数，$g_k(y_t,x,t)$是表示 t 位置状态特征的二值函数，λ_k，μ_k为权重。

为了表达上的统一，状态特征函数可以改写为：$g_k(y_i, x, t) = gk(y_{t-1}, y_t, x, t)$，这样状态转移和状态函数都可以表示为特征函数法的统一形式：$f_k(y_{t-1}, y_t, x, t)$则(1)式可改写为：

$$P_l(y|x)=\frac{1}{Z(x)}\exp\left(\sum_t\sum_k l_k f_k(y_{t-1},y_t,x,t)\right) \quad (2)$$

则输入序列 $x=(x_1, x_2, \cdots\cdots, x_n)$的最大可能标注序列为

$$Y^*=\arg_y \max P_l(y/x) \quad (3)$$

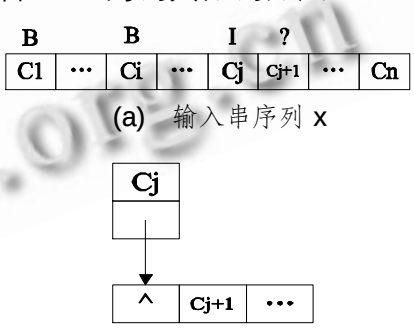
与隐马尔科夫模型相似 CRFs 使用维特比解码(Viterbi)[5]方法来得到最佳的标注结果序列。

Viterbi 方法是一种动态规划算法，将全局最佳的计算过程分解为阶段最佳解的计算，从中找到一条最优路径，使其路长最大。

2.2 CRFs 训练

CRFs 通过训练得出权重参数 λ，来使条件概率 $P_\lambda(y|x)$最大化。常用方法有 L-BFGS 等。

3 结合CRFs的词典分词方法



(a) 输入串序列 x

(b) C_j 与分支节点

图 2　序列与词典结构

首先，规定词典法为整个过程的主导者，CRFs 仅提供辅助决策，不能改变前者作出的决策(即不能改变由词典法产生出的标记)。然后规定产生歧义的条件：设有输入序列 X(图 2(a))，C_i 至 C_j 已能构成词，C_{j+1} 在 C_j 的分支节点中(图 2(b))，词典法无法判断出它是否能作为后继字符继续匹配。

算法描述如下:

(1) 从语料中读入字符，延 TRIE 树链进行匹配

(2) 判断是否满足产生歧义的条件，若满足转(3)，否则转(1)

(3) 使用 CRFs 对 C_{j+1} 进行标记：若为 B，则词典法将 C_{j+1} 作为首字，从首字 hash 表中开始新词的匹配；若为 I，则继续匹配。需要注意的是：CRFs 提供的标记仅供参考，并非最终结果。若词典根据 CRFs 模型得出的 C_{j+1} 标记不能匹配出一个完整词汇，则将 C_{j+1} 标记重置，然后转(1)继续匹配。比如：

B I B I B

数码 / 照相机 / 的 / 外观

假设其已满足产生歧义的条件，若“机”被 CRFs 标注为 B。但词典根据该标记无法匹配出完整词汇，便将其重置为 I，转(1)继续匹配。

(4) 结束匹配

4 实验与分析

4.1 CRFs 标注方案

本文中，CRFs 的标注任务与通常有两点区别：

其一，通常的 CRFs 标注任务是找出整个输入序列的最大可能标注序列，即输出满足(1)式的标注序列。而本文是将 CRFs 作为辅助决策，只需得出某位置 t 上单个字符的最大可能标记。

有文献[6]以边缘概率值来得出单个字符最大可能标记，文中将通过实验来考察 t 位置上，通过常规标注得出的标记与边缘概率最大值得出标记(两者可能相同也可能不同)对结果的影响。其中，边缘概率定义为：

$$P(y_t|X) = \sum_{y_1..y_{t-1}y_{t+1}..y_N} P(Y \mid X) \qquad (4)$$

其二，通常整个标注序列完全由 CRFs 生成，而本文中，在满足产生歧义条件前已有根据词典法生成的标注序列，CRFs 是否基于该序列继续标注也是值得考虑的问题。因此该模型在本文的使用可能需要作出调整。

本文将根据实验结果来确定最合适方案。

实验中，词典法将结合 4 种标注方案：

标注方案 A：以根据词典法得出的前 i-1 个标记(B…BIII)为基础,以边缘概率最大的标记作为 y_i*的结果，得出满足下式的标记 y_i*:

$$P\left(y_i^{'}/X\right) = \sum_{B...B...III y_i^* y_{i+1}...y_n} P\left(Y/X, y_i^{'}\right)$$

$$y_i^* = \arg\max P\left(y_i^{'}/X\right) \qquad (5)$$

标注方案 B：以词典法得出的前 i-1 个标记为基础，以常规 CRFs 标注方法得出从 i 到 n 的最大可能标注序列，其中 y_i 即为所需的辅助决策标记。用到算式如下：

令 $B(y_{t-1}, y_t) = \exp\left(\sum_k l_k f_k (y_{t-1}, y_t, x, t)\right)$

则(2)式可改写为 $P_l(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{n} B(y_{t-1}, y_t)$ (6)

前 i-1 个标记由词典法得出(如图 2(a))，其标记序列的值为 α，则(6)式写为:

$$P_l(y|x) = \frac{1}{Z(x)} a \prod_{t=i}^{n} B(y_{t-1}, y_t) \qquad (7)$$

由于 Z(x)与 y 无关[7]，α 值已确定,因此

$$Y^* = \arg\max \prod_{t=i}^{n} B(y_{t-1}, y_t) \qquad (8)$$

标注方案 C：不以词典法得出的标记为基础，在整个序列上用 Viterbi 算法，计算出 CRFs 边缘概率，得出满足(4)式的 y_i*。

标注方案 D：使用常规 CRFs 标记法得出序列 1 到 n 的满足(3)式的最大可能标记序列，单取 y_i*作为 CRFs 作出的决策。

4.2 试验结果与分析

本文采用 Bakeoff2005 提供的语料进行训练与测试，以准确率、召回率、F-值来衡量试验结果，其中：

准确率(P) = 正确切分结果数/切分结果总数*100%

召回率(R) = 正确切分的结果数/标准切分中的结果数*100%

F-值 = (2PR)/(P+R)*100%

表 1 词典结合 4 种标注方法的实验结果

	准确率	召回率	F-值
A	0.913	0.922	0.917
B	0.952	0.947	0.949
C	0.889	0.897	0.893
D	0.901	0.915	0.908

F-值代表对切分结果的综合评价。

从表 1 可以看出，基于词典得出的标注结果的 A、B 方案在试验中优于根据整体标注的 C、D 方案。而采用 Viterbi 方法得出最大可能标注序列，单取 yi*为决策的 B 方案优于采用边缘概率的 A 方案。

所以文中采用 B 方案，作为分词方法。

下表是本文方案、基于传统的 CRFs[8]方法结果、传统的词典分词法结果比较

表 2　三种方法比较

	本文方案	CRFs	词典分词
F-值	0.949	0.960	0.843

从表 2 可以看出本文的分词效果优于纯粹的词典分词法，而与基于传统 CRFs 模型的分词法相比 F-值稍低。主要原因在于本文提出的方案没有解决词典的未登录词识别问题，而这正是基于传统 CRFs 模型分词的优势。但从分词时间看，传统的 CRFs 分词模型采用 Viterbi 算法，时间复杂度随序列长度呈指数级增长[9]，分词速度较慢。而本文方案仅在句中出现歧义无法对后继字符进行判断时调用 CRFs 模型提供辅助决策。据统计，一篇汉语文档中出现交叉歧义或组合歧义的句子平均大约[10]在 21%左右，约 79%的句子可以直接使用词典正确分词，因此本文提出的方案在一般情况下能比传统 CRFs 分词模型更快完成分词任务。

5　结论与展望

本文提出了一种结合 CRFs 模型的词典分词法。经实验与分析，该方法与基于传统 CRFs 模型的分词法以及词典分词法相比，更适合于对分词速率和准确性都具有一定要求的实际应用。而下一步的工作将针对未登录词的识别问题进行研究。

参考文献

1 吴晶晶,荆继武,聂晓峰,王平建.一种快速中文分词词典机制.中国科学院研究生院学报, 2009(9):703－711.

2 王思力,张华平,王斌.双数组 Trie 树算法优化及其应用研究.中文信息学报, 2006(5):24－30.

3 刘群,张华平,俞鸿魁,程学旗.基于层叠隐马模型的汉语词法分析.计算机研究与发展, 2004(8):1421－1429.

4 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. the Eighteenth International Conference on Machine Learning, ICML'01}, Williams College: Morgan Kaufmann Publishers Inc., USA, 2001.282－289.

5 The Viterbi algorithm. Proceedings of the IEEE,61(3): 268－278.

6 罗彦彦,黄德根.基于 CRFs 边缘概率的中文分词.中文信息学报, 2009(9):3－7.

7 洪铭材,张阔,唐杰,李涓子.基于条件随机场(CRFs)的中文磁性标注方法.计算机科学, 2006,33(10):148－150.

8 李双龙,刘群,王成耀.基于条件随机场的汉语分词系统. 微计算机信息,2006,22(28):178－180.

9 Bai SH, Xia Y, Huang CN. Automatic Part-of-Speech Tagging System of Chinese.Technical Report.

10 Bai Shuanhu. An Integrated Model of Chinese Word Segmentation and Part-of Speech Tagging. Advanced and Applications on Computational Linguistics. 1995. 56－61.