维普资讯 http://www.cqvip.com

中文分词及词性标注一体化模型研究

佟晓筠[1] 宋国龙[2] 刘 强[3] 张 俐[2] 姜 伟[1]

(哈尔滨工业大学(威海)计算机科学与技术学院 威海 264209)[1]

(东北大学信息科学与工程学院 沈阳 110004)[2] (辽东学院计算中心 丹东 118000)[3]

摘 要 本文应用 N-最短路径法,构造了一种中文自动分词和词性自动标注一体化处理的模型,在分词阶段召回 N 个最佳结果作为候选集,最终的结果会在未登录词识别和词性标注之后,从这 N 个最有潜力的候选结果中选优得到,并基于该模型实现了一个中文自动分词和词性自动标注一体化处理的中文词法分析器。初步的开放测试证明,该分析器的分词准确率和词性标注准确率分别达到 98.1%和 95.07%。

关键词 中文分词,词性标注,N-最短路径法

Research on the Model of Integrating Chinese Word Segmentation with Part-of-speech Tagging

TONG Xiao-Jun[1] SONG Guo-Long[2] LIU Qiang[3] ZHANG Li[2] JIANG Wei[1]

(School of Computer Science & Technology, Harbin Institute of Technology at Weihai, Weihai 264209)[1]

(School of Information Science & Engineering, Northeastern University, Shenyang 110004)[2]

(Center of Computer, Liaodong University, Dandong 118000)[3]

Abstract In this paper, we present a model integrating Chinese word segment with part-of-speech tagging. In the early stage, reserves the top N segmentation results as candidates. After Unknown words recognized and POS tagging finished, we get the final result by select form the top N segmentation candidates. We also develop a Chinese lexical analyzer based on this model. The primary experiment proved that the overall accuracy of the proposed analyzer is 98.1% for segmentation and 95.7% for POS tagging respectively.

Keywords Chinese word segmentation, Part-of-speech tagging, N-shortest paths method

1 引言

词是最小的、能够独立活动的、有意义的语言成分。但汉语是以字为基本的书写单位,词语之间没有明显的区分标记,因此中文词语分析是中文信息处理的基础与关键[1,2]。在基于字的自然语言词法分析中,一直以来习惯于将分词和词性标注分别处理。实际上,分词和词性标注有着密切的联系。分词中的切分歧义能用语法知识削解的就约占 90%以上,而涉及语义和语用知识的切分歧义则很少[3]。可见,有机地将分词过程和词性标注过程融合在一起,有利于消除歧义和提高整体效率。

本文应用 N-最短路径法,构造了一种中文自动分词和词性自动标注一体化处理的模型,并实现了一个中文自动分词和词性自动标注一体化处理的中文词法分析器。该方法通过对 N 个最有潜力的粗分结果,分别进行未登录词识别和词性自动标注,形成 N 个候选集,再通过评估判优选择最优结果,有效地将词形信息和词性信息结合在一起。

2 中文分词及词性标注一体化模型

2.1 中文分词及词性标注一体化问题描述

根据噪声信道模型,我们可以将中文分词及词性标注一体化问题描述为:一个已经被标注了词性的词串$\langle W,T\rangle=\langle w_1,t_1\rangle\langle w_2,t_2\rangle\cdots\langle w_n,t_n\rangle$,(其中,$\langle w_i,t_i\rangle$表示具有词性 t_i 的词 w_i),经过有噪声的信道,将词边界和词性信息丢失,在输出端输出为字序列 $C=c_1c_2\cdots c_n$。通过找出跟 C 相对应的$\langle W,T\rangle$,经比较得出具有最大概率的结果$\langle W,T\rangle^*$。

$$\langle W,T\rangle^*=\arg\max_{W,T}P(\langle W,T\rangle|C) \quad (1)$$

2.2 中文分词及词性标注一体化统计模型

式(1)从概率统计角度描述了分词及词性标注一体化的一般模型。为了将中文自动分词和词性自动标注一体化处理,并实现将词语信息和词性信息融合在一起作为最终结果的评价依据,本节引入了 N-最短路径法,用以产生包含 N 个最大结果的候选集。

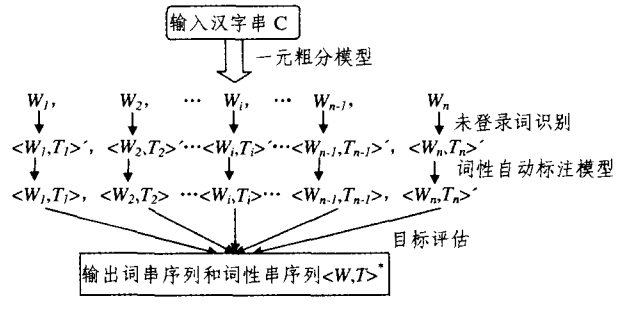

图 1 一体化模型处理过程

对于字序列 C,经过分词处理后,得到了 N 个概率最大的分词结果$(W_1,W_2,\cdots,W_n)$。对每个分词结果进行未登录

*)本课题得到 2002 年山东省科技发展计划项目基金资助(项目号:2002-276-022090104)。佟晓筠 副教授,博士生,研究方向:信息安全、Web 智能搜索。

维普资讯 http://www.cqvip.com

词的识别，并且保留了未登录词的词性，产生具有少量词性信息的词序列（$\langle W_1,T_1\rangle',\langle W_2,T_2\rangle',\cdots,\langle W_n,T_n\rangle'$），然后分别进行词性自动识别，得到具有完备词性信息的词序列（$\langle W_1,T_1\rangle,\langle W,T_2\rangle,\cdots,\langle W_n,T_n\rangle$）。最后，用评估函数进行判优，得到概率最大的结果$\langle W,T\rangle^*$。其处理过程如图1所示。

2.3 一元粗分模型

根据噪声-信道模型，我们可以认为一个词串经过有噪声的信道传送，由于噪声干扰而丢失了词边界标记，输出为一个汉字串。那么，自动分词就是已知一个汉字串，求跟它对应的、有最大概率的词串。即

$$W'=\arg\max_W P(W|C) \tag{2}$$

由贝叶斯公式知

$$W'=\arg\max_W P(W|C)=\arg\max_W \frac{P(W)P(C|W)}{P(C)} \tag{3}$$

式中，$P(C)$是汉字串的概率，它是一个常数，不必考虑。$P(C|W)$是词串到汉字串的条件概率。显然，从词串变为汉字串只有唯一的方式。因此，在已知词串的条件下，出现相应的汉字串的概率是1，也不必考虑。我们仅仅需要考虑的是$P(W)$，即词的概率。

上述公式可简化为

$$W'=\arg\max_W P(W) \tag{4}$$

词串概率采用一元语法进行求解，则

$$P(W)=\prod_{i=1}^{n} p(w_i) \tag{5}$$

这就是说，概率最大的词串便是最佳的词串[4~6]。

2.4 词性自动标注模型

在词性自动标注过程中，我们在已知词串W，寻求最大概率的词性标注列T'。

$$T'=\arg\max P(T|W) \tag{6}$$

根据贝叶斯公式：

$$P(T|W)=\frac{P(T)P(W|T)}{P(W)} \tag{7}$$

其中分母$P(W)$是词串W的概率，是一个常量，因此上式可简化为

$$P(T|W)=P(T)P(W|T) \tag{8}$$

我们假定词语之间是相互独立的，并且词语的出现只依赖于它本身的标注，则已知标记串T的条件下词串W的概率可近似地用每个词在已知标记时的条件概率的乘积来表示[7]：

$$P(W|T)\approx P(w_1|t_1)P(w_2|t_2)P(w_3|t_3)\cdots P(w_n|t_n) \tag{9}$$

我们假定一个标记的概率取决于出现在它前面的那个标记，那就可以用T中每个标记的概率的乘积来表示：

$$P(T)\approx P(t_1|t_0)P(t_2|t_1)P(t_3|t_2)\cdots P(t_n|t_{n-1}) \tag{10}$$

因为t_0是虚设的标记，所以$P(t_1|t_0)$实际上是$P(t_1)$。

综合起来，我们的词性标记的统计模型可以表示为

$$T'=\arg\max\prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i) \tag{11}$$

我们用隐马尔可夫模型来描述词性自动标注问题：以词串$W=w_1w_2\cdots w_n$作为观察到的输出序列，以对应的词性标注串$T=t_1t_2\cdots t_n$作为隐藏的状态转移序列，将词语以某词性出现的概率变相作为状态转移的发射概率。在求解过程中，应用Viterbi算法。Viterbi算法有三步：(1)初始化；(2)推导；(3)终止和路径读出[8]。

(1)初始化

第一个词处于状态j(标注为j)的概率

$$\delta_1(t^j)=P(t^j|w_1) \tag{12}$$

(2)推导

词$i+1$处于状态j(标注为j)的概率

$$\delta_{i+1}(t^j)=\max_{1\leqslant k\leqslant T}[\delta_i(t^k)\times P(w_{i+1}|t^j)\times P(t^j|t^k)],1\leqslant j\leqslant T \tag{13}$$

词$i+1$处于状态j(标注为j)时，词i所处的最有可能的状态(标记)为

$$\Psi_{i+1}(t^j)=\arg\max_{1\leqslant k\leqslant T}[\delta_i(t^k)\times P(w_{i+1}|t^j)\times P(t^j|t^k)],1\leqslant j\leqslant T \tag{14}$$

(3)终止和路径读出

其中$t_1,\cdots,t_n$是我们为词语序列$w_1,\cdots,w_n$选择的标记：

$$t_n=\arg\max_{1\leqslant j\leqslant T}\delta_n(t^j) \tag{15}$$

$$t_i=\Psi_{i+1}(t_{i+1}),1\leqslant i\leqslant n-1 \tag{16}$$

$$P(t_1,\cdots,t_n)=\max_{1\leqslant j\leqslant T}\delta_{n+1}(t^j) \tag{17}$$

2.5 目标评估函数

一个词串对应的汉字串是唯一的，即

$$P(C|W)=1\rightarrow P(CW)=P(W) \tag{18}$$

$$\begin{aligned}P(W,T|C)&=P(T|CW)P(W|C)=P(T|W)P(W|C)\\&=P(T)P(W|T)/P(W)\times P(W)/P(C)\\&=P(T)P(W|T)/P(C)=P(T)P(W|T)\end{aligned} \tag{19}$$

利用隐马模型展开$P(T)P(W|T)$，并引入共现概率

$$P(\langle W,T\rangle|C)=\prod P(t_i|t_{i-1})P(w_i|t_i) \tag{20}$$

$$P^*(W,T)=\ln P(W,T)=\sum\ln P(t_i|t_{i-1})+\sum\ln P(w_i|t_i) \tag{21}$$

评估函数如下：

$$R^{\#}=\arg\max_{W,T}[\sum P(t_i|t_{i-1})+\sum P(w_i|t_i)] \tag{22}$$

3 中文分词及词性标注一体化系统的设计与实现

这个中文分词及词性标注一体化系统由5部分组成：预处理模块、自动分词模块、未登录词识别模块、词性自动标注模块和评估模块，其系统结构如图2所示。

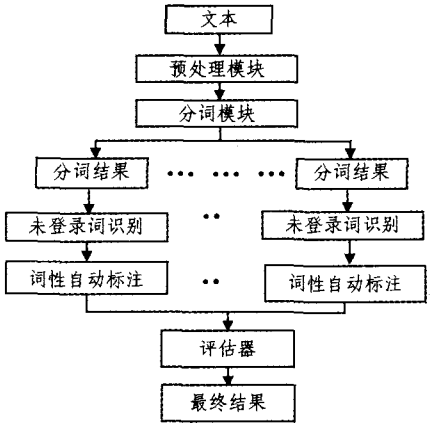

图2 中文分词及词性标注一体化系统结构图

3.1 预处理阶段

在大陆，一般使用国标码(GB2312-80)。在该编码体系中，中文字符都用两个字节表示，而且每个字节的ASCII码都大于127，这可以作为区分汉字字符与非汉字字符的依据。

在预处理过程中，我们需要进行两次扫描。第一次，扫描待处理文本，根据标点符号进行断句。第二次，扫描通过断句

(下转第212页)

维普资讯 http://www.cqvip.com

参 考 文 献

1 Elmesbahi J, Bouattane O, Benabbou Z. theta (1) time quadtree algorithm and its application for image geometric properties on a mesh connected computer (MCC). IEEE Transactions on Systems, Man and Cybernetics, 1995, 25(12)
2 Shneier M. Calculations of Geometric Properties Using Quadtrees. Computer Graphics and Image Processing, 1981, 16 (3): 296～302
3 Dyer C R. Computing the Euler Number of an Image From its Quadtree. Computer Graphics and Image Precessing, 1980, 13 (3): 270～276
4 Samet H. Connected Component Labeling Using Quadtrees. Comm ACM, 1981, 28 (3): 487～501
5 Gargantini I. An Effective Way to Represent Quadtrees. Comm ACM, 1982, 25(12): 905～910
6 Chang F, Chen C J, Lu C J. A linear-time component-labeling algorithm using contour tracing technique. Computer Vision and Image Understanding, 2004, 93(2): 206～220
7 Suzuki K, Horiba I, Sugie N. Linear-time connected-component labeling based on sequential local operations. Computer Vision and Image Understanding, 2003, 89(1): 1～23
8 Dillencourt M B, Samet H, Tamminen M. A general approach to connected-component labeling for arbitrary image representations. J ACM, 1992, 39(2): 253～280
9 Shima Y, Murakami T, Koga M, et al. A high-speed algorithm for propagation-type labeling based on block sorting of runs in binary images, In: Proceedings of 10th International Conference on Pattern Recognition, 1990. 655～658
10 Nicol C J. A systolic approach for realtime connected component labeling. CVGIP: Image Understanding, 1995, 61 (1): 17～31

(上接第 175 页)

得出的每一句话，区分汉字和非汉字字符，并对非汉字字符用自动机识别数字和英文字串。汉字串作为自动分词模块的输入序列。

3.2 分词及未登录词识别阶段

自动分词过程中，需要找出 N 个概率最大的切分词串（N 的值可以根据需要通过用户界面进行设定）。计算概率时，由于每个词的概率是一个很小的正数（小于 1），最后词串的概率接近于 0，无法在计算机上表示。为了解决这个问题，我们用费用代替概率，词串的费用按如下公式计算：

$$Fee(W)=\sum_{i=1}^{n}-\log P(w_i) \qquad (23)$$

处理时，首先从左到右扫描输入汉字串，按其在句子中的先后顺序列出所有候选词，并保留在数组中。扫描候选词序列，采用动态规划的方式，计算每个候选词的 N 个最佳前驱词，即累计费用最小的前一词，并计算当前词的费用，保留这 N 个最佳前驱词信息。如果当前词是终点词，通过回退，得到 N 个费用最小（概率最大）的词串。

未登录词的识别包括姓名识别和地名识别两部分，识别过程如图 3 所示。姓名的识别采用基于概率统计的方法。姓名的费用为姓氏费用与人名费用的和。如果一个汉字串的姓名费用小于一个阈值，就认为它是一个姓名。地名的识别才用基于规则的方式，主要通过检查其后缀来识别。识别出的未登录词保留其词性信息。

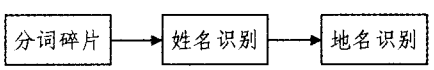

图 3 未登录词识别模块路流程图

3.3 词性标注及评估判优阶段

词性自动标注的对象是一句话的词串，我们的目标是寻找一条概率乘积最大的词性序列。将词串中的每个词的所有词性及其费用添入到该节点中（未登录词的词性信息由前面的处理获得，其余从词典中获得）。从左到右扫描词串，计算该词在某一词性的费用和前一词最有可能的词性。当前词为终点词时，进行回退，得出该词串的词性序列。

对每个分词结果分别进行未登录词识别和词性标注之后，形成了 N 个具有词性信息的词串。我们根据式(22)对每个路径进行打分，费用最小者为最终结果。在这个模块中，我们还进行叠词识别和合并部分未登录词的操作。

4 实验与分析

我们所实现的中文分词及词性标注一体化系统使用了词性标记集为 39 个词性标记的北大标记集，词典中包括 108784 条记录的词语表、114758 条记录的词性表和 2328 条记录的姓名频率表。本文对从 1998 年 1 月人民日报提取的包含 5221 词的文章进行了开放测试，结果如表 1 所示。

从表 1 中数据可看出：(1)该一体化系统的性能良好，开放测试分词准确率和词性标注准确率高于文[9]的 96% 和 94%。(2)引用 N-最短路径法后，分词准确率和词性标注准确率都得到提高，证明前期保留多个粗分结果是合适的。(3)词性信息的引入提高了分词的准确率。(4)随着 N 值的增大，系统性能提升变得不明显。针对这种情况，并且考虑到系统的运行效率，我们一般选取 N 值为 3。

表 1 系统性能测试表

N 值	总词数	自动分词		词性自动标注	
		正确切分词数	准确率 (%)	正确标注词数	准确率 (%)
1	5221	5090	97.49	4920	94.23
2	5221	5190	97.85	4956	94.92
3	5221	5121	98.08	4964	95.07
4	5221	5123	98.12	4964	95.07

结论 本文应用 N-最短路径法，构造了一种中文自动分词和词性自动标注的模型，并实现了一个中文自动分词和词性自动标注一体化处理的中文词法分析器。经测试，该系统具有较高的分词准确率和词性标注准确率，初步证明该方法是有效的。

未登录词的识别和概率字典的构建是影响本系统性能的重要因素。如何处理好这些问题，还有待进一步的研究。

参 考 文 献

1 张华平，刘群．基于 N-最短路径的中文词语粗分模型．中文信息学报，2002，16(5)：1～7
2 ZHANG Hua-Ping, LIU Qun, Zhang Hao, et al. Automatic Recognition of Chinese Unknown Words Recognition. In: First SIGHAN Workshop attached with the 19th COLING, 2002. 71～77
3 何克抗，徐辉，孙波．书面中文自动分词专家系统设计原理[J]．中文信息学报，1991，5(2)：1～14
4 陈小荷．现代中文自动分析．北京语言文化大学出版社，2000
5 Yuan S C, Henry T. Probability Theory. Springer-Verlag New York Inc, 1978. 324～338
6 Manning C D, Hinrich S. Foundations of statistical natural language processing, MIT press, 1999. 197～202
7 Ma Qing, Isshara H, Maoson S A. Mnhi-neuro Tagger Applied in Chinese Texts. In: Proceedings 1998 International Conference on Chinese Information Processing, Beijing, 1998-11-18/20. 200
8 Fine S, Singer Y, Tishby N. The hierarchical Hidden Markov Model: Analysis and applications. Machine Learning, 1998, 32 (1): 41
9 白栓虎．中文词切分及词性自动标注一体化方法[J]．计算语言学进展与应用(JSCL－95)，1995．56～61