文章编号:1671 - 6833(2011)04 - 0103 - 04

CRF 模型中参数 f 在字标注汉语分词中的适用性研究

赵晓凡,胡顺义,刘永革

(安阳师范学院 计算机与信息工程学院,河南 安阳 455002)

摘 要:汉语分词作为中文信息处理的首要环节,其精确度对后续步骤的准确度和处理速度成逐级放大性影响.如何提高分词的准确度和处理速度成为近年研究的重点.采用条件随机场模型进行汉语分词,通过定量分析 CRF 工具包训练参数 f,研究减少特征对分词准确度以及模型大小的影响程度,实验分别在国际汉语分词评测 Bakeoff2005 提供的北京大学和微软亚洲研究院两个语料上进行封闭测试,并对比采用不同模板时增加 f 参数值对分词性能的影响,最终得出实验结果:随着 f 参数值的增加,分词的准确度和生成的模型大小成正比,且 F 值减小的程度相对训练生成模型大小的减小程度要小得多.

关键词:汉语分词;字标注;f 阈值;模型大小;CRF + + 工具包

中图分类号:TP391 **文献标志码**:A

0 引言

随着信息化时代的到来,海量文本信息的挖掘特别是中文信息的处理问题摆到了我们面前,成为近 20 年来研究的重点,越来越多的研究者着手于利用计算机进行快速、准确的文本信息处理.中文分词是中文信息处理的第一步,广泛应用于诸如中文信息检索、文本分类、自动问答系统等领域[1,2].

与英文不同,中文文本是由连续的字串组成的,且词与词之间没有明显的分隔符,不同的字可以出现在不同词的不同位置.中文自动分词任务,就是由机器在中文文本中自动识别词边界,通俗的说就是要由机器在词与词之间自动加上分隔符.如句子"森林防火事关重大",应切分为"森林/防火/事关重大".中文分词技术的发展与日俱进,由传统的基于词典和规则[3,4]的方法,转变为基于统计的方法[5],后者又分为基于字标注的分词法[6]和基于字串标注的分词法[7].统计方案要优于基于人工规则的方法,它主要是利用统计语言模型对自然语言文本进行数学抽象、建模后再进行自动分词.一般分为训练、测试和评测 3 个过程,其中对文本的训练和测试都要经过分词、特征选择和训练分类 3 步.笔者用条件随机场模型对文本进行分词,并通过实验定量分析 CRF + + 0.53 工具包中的 f 阈值大小,统计同等条件下,由不同模板扩展出来的特征函数的个数对整个分词的效果以及模型大小的影响.

1 基于字标注的汉语分词

基于字的中文分词方法是由薛念文在 2003 年提出的,这种方法从词的结构来对词进行划分,将汉语分词任务转换为字序列标注的问题.根据每个字在形成一个特定意义的词语时所占据的不同但是确定的位置,可以把词标注为二词位(是否为词边界)、四词位(左边界、右边界、中间位和单字)和六词位(在四词位基础上,增加了两个中间位标识).在使用的特征方面,不仅限于字本身,可以是三字滑动窗口(即当前字的前一个和后一个字)和五字滑动窗口(即当前字的左右两个字).这样利用相对固定的字来推断相对不固定的字的位置信息,我们把字标注问题转换为了一个分类问题.笔者采用的是四词位的三字窗口进行实验,即 BEMS 标注集系统,B 代表词的词首,E 代表词的词尾,M 代表词的中间,S 代表单字词.那么下面句子(1)的分词结果就可以表示为句子(2)的逐字标注的形式:

(1)分词结果:/电力部/从/政企/合一/转变/为/只/行使/政府/的/行政管理/职能/;

(2)字标注形式:电/B 力/M 部/E 从/S 政/B 企/E 合/B 一/E 转/B 变/E 为/S 只/S 行/B 使/E 政/B 府/E 的/S 行/B 政/M 管/M 理/E 职/B 能/E;/S;

收稿日期:2011 - 02 - 01;**修订日期**:2011 - 04 - 07

基金项目:国家自然科学基金资助项目(60875081);河南省教育厅高等学校青年骨干教师资助项目(2009GGJS - 108).

作者简介:赵晓凡(1981 -),女,河南安阳人,安阳师范学院讲师,硕士,研究方向为自然语言处理,汉语分词,信息安全等.

2 基于条件随机场的字标注中f阈值分析

2.1 条件随机场理论

CRFs由Lafferty[8]在2001年提出来,是一种无向图模型或者马尔可夫随机域,它采用一阶链式无向图结构计算给定观察值条件下输出状态的条件概率如图1所示.

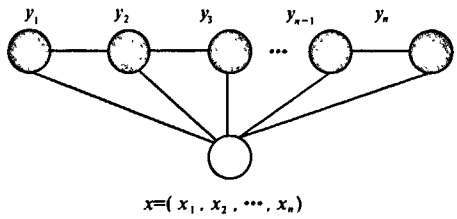

图1 线链CRFs的图形结构

Fig.1 The qraphics strnctuer of line CRFS

设$O=\{o_1,o_2,\cdots\cdots,o_T\}$表示被观察的输入字串序列;$S=\{s_1,s_2,\cdots\cdots,s_T\}$表示将被预测的词位标记序列,则在给定一个输入字串序列的情况下,对参数为$\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$的线链CRFs,其输出的词位序列的条件概率为:

$$P_\Delta(S|O)=\frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(S_{t-1},S_t,O,t)\right) \quad (1)$$

其中,Z_O是归一化因子,定义为:

$$Z_O=\sum_S\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(S_{t-1},S_t,O,t)\right) \quad (2)$$

$f_k(S_{t-1},S_t,O,t)$是一个任意的特征函数,用于表达上下文可能的语言特征,通常是一个二值表征函数,表示如下:

$$f_k(S_{t-1},S_t,O,t)=\begin{cases}1,\text{如果满足条件}\\0,\text{否则}\end{cases} \quad (3)$$

λ_k是一个需要被学习的参数,其对应于每一个特征函数的权值,取值范围可以是$-\infty$到$+\infty$.给定一个由式(1)定义的条件随机场模型,对任意的输入字串,其最佳词位标记序列应满足式(4):

$$S^*=\arg\max_S P_\Delta(S|O) \quad (4)$$

要求出使得$P_\Delta(S|O)$最大的标记序列S^*,可以使用Viterbi算法进行计算.

2.2 特征模板的选取

笔者主要研究训练过程中f参数对整个分词性能的影响,所以模板就选取了比较有代表性的Template6、Template6 + B、Template10和Template10 + B分别进行了实验.

3 实验及结果分析

3.1 实验设计

为了能够更好的验证实验结果,笔者采用了2005年SIGHAN举办的第二届国际中文分词评测所提供的MSRA和PKU两个单位的简体中文语料,包括训练语料和测试语料,分别进行了4个不同模板下的6组封闭测试.表1是这两个语料的相关统计信息.首先在MSRA语料进行了f值从1取到20时,采用模板Template6和Template6 + B分别进行测试,在看到大致的结果后,又在PKU语料上分别采用4个模板Template6、Template6 + B、Template10和Template10 + B,f值取值从1取到10,进一步验证之前的结果.

表1 Bakeoff 2005评测语料的相关信息

Tab. 1 Bakeoff 2005 corpus statistics

语料来源	训练语料大小/M	训练语料词数/K	测试语料大小/K	测试语料词数/K
MSRA	12.542	2 368	368	104
PKU	5.769	1 109	336	107

3.2 实验结果

整个实验过程都要经过训练,测试和评测三个阶段,而性能结果一般都出现在训练和评测阶段,所以将这2个阶段的结果都以文本的形式保存下来,分别以同样语料同样模板为准则,放到一起比较,得到6组实验结果,这里只列出在PKU语料上进行测试的部分实验数据,如表2所示,然后分别从不同角度来分析f阈值变化对它们的影响.

表2 PKU语料训练过程记录

Tab. 2 Record of training process on pku

阈值f	Template6 模型大小/MB	F	Template6 + B 模型大小/MB	F	Template10 模型大小/MB	F	Template10 + B 模型大小/MB	F
1	33.813	0.911	33.813	0.929	52.338	0.923	52.338	0.929
2	15.970	0.911	15.971	0.929	25.074	0.923	25.074	0.928
3	10.689	0.908	10.689	0.928	16.954	0.922	16.954	0.927
4	8.089	0.907	8.089	0.926	12.916	0.920	12.916	0.927
5	6.525	0.904	6.525	0.925	10.490	0.919	10.490	0.926
6	5.478	0.903	5.478	0.925	8.857	0.918	8.857	0.925
7	4.722	0.900	4.722	0.923	7.665	0.916	7.665	0.924
8	4.153	0.898	4.153	0.921	6.768	0.916	6.768	0.922
9	3.710	0.896	3.710	0.921	6.070	0.915	6.071	0.922
10	3.352	0.894	3.352	0.92	5.501	0.914	5.502	0.921

从实验结果分析可以得到:(1)随着f阈值的增加,特征数,模型大小和F值都有一定程度的减小,且相对于模型大小的减小程度来说,F值的减小并不是很大;(2)训练时间和迭代次数基本上不受f阈值的影响,并没有呈相应的递减变化;(3)6组实验都有特征数减少,模型减小,但F值保持不变的情况,基本上在f值取到10的时候都会有1到2次出现;(4)加入Bigram模板后,整体

效果都要比不加好.

3.3　f 阈值对模型大小以及 F 值的影响

综合考量几组实验数据可以看出,由于 f 阈值参数的加入,使得出现次数少于 NUM 的特征数被舍去,所以特征数在逐次减少.但是由于要统计各个特征出现的次数,所以对训练时间来说,并没有逐次减少,反而有增加的情况,但增加的时间不长,基本上维持在原训练时间范围内,迭代次数和训练时间有相同的变化.受 f 阈值影响较大的是模型大小的变化,如图 2,3 所示,以 MSRA 语料的 Template6 + B 训练数据为例,在 f 阈值增加的过程中,模型大小是逐渐减小的,而且减小的幅度较大,当 f 阈值从 1 取到 20 变化时,模型大小从 52 M 减小到 3 M,大致减小 94%,同时从图 3 可以看到,在同样的条件下,F 值从 0.963 降到了 0.941,仅仅下降 2.2 个百分点.

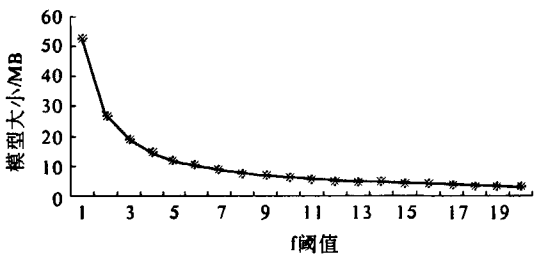

图 2　模型大小随 f 阈值变化曲线

Fig. 2　The variation curve of the model size

3.4　F 值相同结果比较

在实验过程中,发现有相邻两个 f 阈值变化时,对 F 值没有影响,但是模型减小的情况,当 f

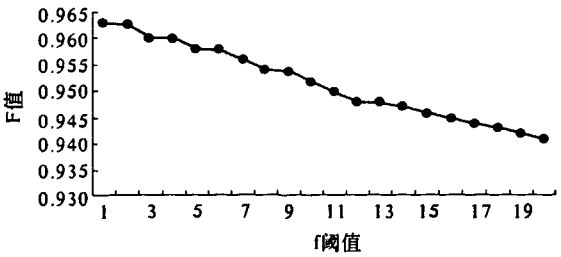

图 3　F 值随 f 阈值变化曲线

Fig. 3　The variation curve of F value

阈值取到 10 时,PKU 语料上 Template6 出现 1 次,其他 3 个模板集各出现 2 次;当 f 阈值取 20 时,MSRA 语料上 Template6 出现 2 次,Template6 + B 出现 5 次.经过比对这些数据发现,当相邻两个 f 阈值变化出现相同 F 值时,模型大小至少可以减小 5.96 个百分点,好的情况下模型大小甚至可以减小 52 个百分点,平均下来模型大小要减小 24 个百分点,这种好的情况一般出现在 f 阈值在前面的变化中,如 f NUM 值取 1 或者 2 和取 5 或者 6 的情况下,肯定是取模型大小较小的 2 和 6 作为参数 f 的值较好,这样放到实际的应用系统中,在保证分词准确率不变的条件下,加载模型的时间就可以大大降低,整个应用系统的运行效率就可以大大提高.表 3 是所有相同 F 值出现时的数据对比情况以及模型大小的减小率.

表 3　相同 f 值记录

Tab. 3　Recond of the same f

MSRA 语料					PKU 语料				
模板集	f 阈值	模型大小/MB	F	模型减小率/%	模板集	f 阈值	模型大小/MB	F	模型减小率/%
Template6	5	12.116	0.939	14.61	Template6	1	33.813	0.911	52.77
	6	10.346	0.939			2	15.97	0.911	
Template6 + B	14	4.846	0.925	5.96	Template6 + B	1	33.813	0.929	52.77
	15	4.557	0.925			2	15.971	0.929	
	1	52.334	0.963	48.53		5	6.525	0.925	16.05
	2	26.935	0.963			6	5.478	0.925	
	3	18.822	0.96	21.95	Template10	1	52.338	0.923	52.09
	4	14.69	0.96			2	25.074	0.923	
	5	12.116	0.958	14.61		7	7.665	0.916	11.70
	6	10.346	0.958			8	6.768	0.916	
	8	8.042	0.954	9.81	Template10 + B	3	16.954	0.927	23.82
	9	7.253	0.954			4	12.916	0.927	
	12	5.59	0.948	7.12		8	6.768	0.922	10.30
	13	5.192	0.948			9	6.071	0.922	

4 结论

用统计语言模型来进行汉语分词当下已经是主流的处理方式.训练过程中的特征选择也已经产生了好多方法,对于CRF++工具包中的自带特征选择参数的效果却还未进行过深入研究,通过这次实验证明,选择适当的f参数的值,可以在保证F值在一定的准确率范围内得到较小的训练模型,这样对于一些实际应用领域,如网络搜索引擎,模型的减小可以降低加载时长,提高相应的运行效率.下一步需要对相邻两个相同F值的评测结果进行比对,找到被减掉的特征,统计数据并分析其规律,期待在减去后对分词准确率没有影响的特征中,能研究出更加适用的特征选择方法.

参考文献:

[1] 姜维,王晓龙,关毅,等. 基于多知识源的中文词法分析系统[J]. 计算机学报, 2007,30(1):137-145.

[2] 刘群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8):1421-1429.

[3] DAVID D P. A trainable rule-based algorithm for word segmentation[C]. spain: Proceedings of ACL 1997, 321-328.

[4] CHENG K S, YOUNG G H, WONG K F. A study on word-based and integral-bit Chinese text compression algorithms[J]. Jounal of the American Society for Information Science, 2001, 50(3):218-228.

[5] SPROAT R. A stochastic finite-state wod segmentation algorithm for Chinese[J]. Computational Linguistics. 1996, 22(3):377-404.

[6] XUE Nian-wen. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1):29-48.

[7] ZHANG Rui-qiang, GENICHIRO K, EIICHIRO S. Subword-based tagging for confidence dependent Chinese word segmentation[C]. Australia. Proceedings of the COLING/ACL, 2006:961-968.

[8] RABINER L R. A tutorial on hidden markov models and selected applications in speech recognition [J]. Proceedings of IEEE. 1989, 77(2): 257-286.

Research on the Applicability of Parameter f in Character-based Tagging Approach of Chinese Word Segmentation

ZHAO Xiao-fan, HU Shun-yi, LIU Yong-ge

(School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China)

Abstract: As the first and foremost part of Chinese information processing, the accuracy of Chinese word segmentation direct lead to magnified effect of the accuracy and processing speed in the following steps. In recent years, more and more researchers focus on how to improve the accuracy and processing speed of Chinese word segmentation. In this paper, the conditional random field model is used to segment Chinese word. Through quantitative analysis of the parameter f in CRF training process, a lot of experimental are done to find out whether the reduction of features can affect the accuracy of Chinese word segmentation and the size of model. Closed evaluations are performed on PKU and MSRA corpus provided by the second international Chinese word segmentation Bakeoff-2005 with the different templates compare to the different experimental data on increasingly parameter value f for one to ten or one to twenty. The final results show: Increase of f parameter value, the accuracy of Chinese word segmentation is always proportionate to the model size, and the decrease of F is far smaller than the model size which generated by training process.

Key words: Chinese word segmentation; character tagging; parameter f; model size; conditional random fields toolkit