DOI:10.19304/j.cnki.issn1000-7180.2011.10.017

基于条件随机场的汉语词性标注

于江德[1],葛彦强[1],余正涛[2]
(1 安阳师范学院 计算机与信息工程学院,河南 安阳 455002;2 昆明理工大学 信息工程与自动化学院,云南 昆明 650051)

摘 要:近年来条件随机场广泛应用于各类序列数据标注中,汉语词性标注中应用条件随机场对上下文建模时会扩展出数以亿计的特征,在深入分析特征产生机理的基础上对特征模板集进行了优化,采用条件随机场进一步研究了汉语词性标注中设定的特征模板集、扩展出的特征数、训练后模型大小、词性标注精度等指标之间的关系.实验结果表明,优化后的特征模板集在模型训练时间、训练后模型大小、标注精度等指标上达到了整体最优.

关键词:汉语词性标注;条件随机场;上下文;特征模板集;上下文特征
中图分类号:TP391 **文献标识码**:A **文章编号**:1000-7180(2011)10-0063-04

Chinese Part-of-speech Tagging Based on Conditional Random Fields

YU Jiang-de[1], GE Yan-qiang[1], YU Zheng-tao[2]
(1 School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China;
2 School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650051, China)

Abstract: In recent years, conditional random fields is widely used in various types of sequence data labeling, hundreds of millions of features will be extended out in the context modeling using CRFs for Chinese part of speech tagging, feature template set is optimized after in-depth analysis of the context features. We further studied the relations of the feature template set and the training model size, tagging accuracy for Chinese part-of-speech tagging via using maximum entropy model. Experimental results show that optimized feature set of templates is the overall optimum.

Key words: Chinese part-of-speech tagging; conditional random fields; context; feature templates; context feature

1 引言

条件随机场[1-2](conditional random fields, CRFs)是一种用于序列数据标注的条件概率模型[3].近几年来,CRFs已经被成功地应用到许多序列数据标注问题中,例如,汉语分词[3]、词性标注[4]、组块分析和短语识别[5]、命名实体识别[6]、词义消歧[7]等.在应用统计语言模型对序列数据标注的建模过程中,上下文扮演着解决问题所需语言知识和资源提供者的重要角色.在统计语言模型训练中,从特征模板扩展出的上下文特征起到了描述上下文的作用,并对模型的训练起重要作用.汉语词性标注中应用CRFs对上下文进行建模时按传统的特征模板集会扩展出数以亿计的上下文特征,如此多的特征会使得条件随机场的训练过程耗费时间过长,甚至导致现有的一些条件随机场工具包(例如,CRF++工具包)无法运行或崩溃.针对这一问题,本文首先深入分析了CRFs对汉语词性标注建模中上下文特征的表示方法和上下文特征产生的内在机理,在此基础上对采用CRF++工具包实现的汉语词性标注所使用的特征模板集进行了优化,实验结果表明,优化后的特征模板集在模型训练时间、训练后模型大小、词性标注精度等指标上达到了整体最优.

2 基于条件随机场的汉语词性标注

2.1 条件随机场简介

条件随机场是一种以给定的输入结点值为条件来预测输出结点值概率的无向图模型.CRFs通过

收稿日期:2011-04-27;**修回日期**:2011-06-30
基金项目:国家自然科学基金项目(60663004);河南省高等学校青年骨干教师项目(2009GGJS-108)

定义标记序列和观察序列的条件概率来预测最可能的标记序列.用于模拟序列数据标注的 CRFs 是一个简单的链图或线图(如图 1 所示),它是一种最简单也最重要的 CRFs,称为线链 CRFs.

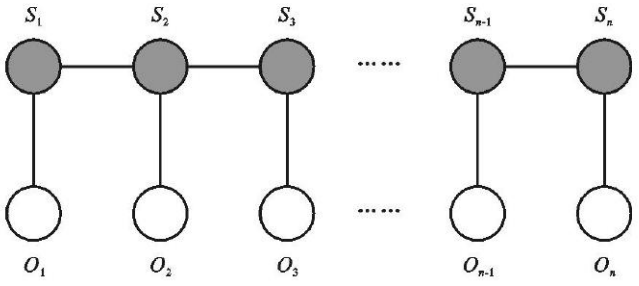

图 1 简单的链图或线图

设 $O=\{o_1,o_2,\cdots,o_T\}$ 表示被观察的输入数据序列,例如有待标注词性的词语.$S=\{s_1,s_2,\cdots,s_T\}$ 表示被预测的状态序列,每一个状态均与一个词性标记(例如动词 v、名词 n)相关联.这样,在一个输入序列给定的情况下,参数为 $\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$ 的线链 CRFs,其状态序列的条件概率为

$$P_\Lambda(S\mid O)=\frac{1}{Z_O}\exp\Big(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\Big) \quad (1)$$

式中,Z_O 是归一化因子,它是所有可能的状态序列的条件概率"得分"之和:

$$Z_O=\sum_{S}\exp\Big(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\Big) \quad (2)$$

$f_k(s_{t-1},s_t,o,t)$ 是一个任意的特征函数,通常是一个二值表征函数.λ_k 是一个需要从训练语料中学习的参数,是相应的特征函数 $f_k(s_{t-1},s_t,o,t)$ 的权重.

给定一个由式(1)定义的条件随机场模型,在已知输入数据序列 O 的情况下,最可能的词性标记序列可以由式(3)通过类似于隐马尔科夫模型中的韦特比算法动态规划求出:

$$S^*=\arg\max_S P_\Lambda(S\mid O) \quad (3)$$

2.2 条件随机场对汉语词性标注的建模

2.2.1 上下文特征的表示方法

用 CRFs 描述和表示上下文特征时,首先要考虑上下文范围开设大小的问题,其次要考虑上下文特征如何表示.通常情况下,上下文的选取是基于当前词左右一定范围进行的,这个固定的范围被称为"上下文窗口",该窗口表示对当前词进行词性标注时要考虑的上下文范围大小.图 2 是进行汉语词性标注时可能的上下文窗口的示意图.如果使用当前词前后各两个词作为上下文的范围,则上下文范围是"5 词窗口".如果使用当前词前后各一个词作为上下文的范围,则上下文范围就是"3 词窗口".

统计语言建模中上下文特征的描述和表示是通

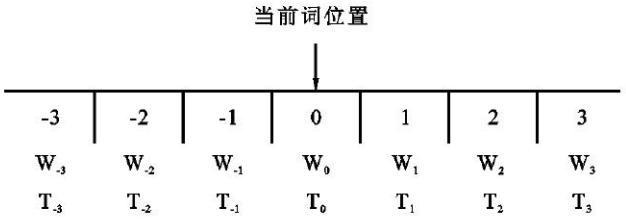

图 2 可能的上下文窗口

过设定合适的特征模板实现的.特征模板的主要功能是定义上下文中某些特定位置的语言成分或信息与某类待预测事件的关联情况.习惯上,特征模板可以看作是对一组上下文特征按照共同的属性进行的抽象.由于本文是根据一个词串序列中的当前词及其上下文来确定该词的词性,因此就由该词前后出现的词、词语组合等成分及这些语言成分出现的位置来描述上下文特征.如果限定上下文范围为"5 词窗口",根据模板中出现的词与当前词的距离属性将上下文特征抽象为 10 类,对应了 10 个特征模板.表 1 给出了这些特征模板及其表征的意义.其中,表中的 W_n 代表当前词或者和当前词相距若干位的词.例如,W_0 表示当前词,W_1 表示当前词的后一个词,W_{-1} 表示当前词的前一个词,依此类推.从表 1 可以看到,这 10 个特征模板的前 5 个是单个词语特征模板,即这些模板中只包含一个词语;后 5 个是双词语特征模板,每个特征模板都是上下文中两个词语的组合.表 1 中最后一行的特征模板是:$T_{i-1}T_0$,该模板用于表征上下文中相邻两个词的词性转移特征.表 1 中的第 3 列是上下文为"中国/ns 政府/n 顺利/ad 恢复/v 对/p"时相应特征模板扩展出的上下文特征.

表 1 特征模板列表

特征模板	模板表征的意义	扩展的上下文特征
W_{-2}	当前词的前面第二个词	中国
W_{-1}	当前词的前一个词	政府
W_0	当前词	顺利
W_1	当前词的后一个词	恢复
W_2	当前词的后面第二个词	对
$W_{-2}W_{-1}$	当前词的前面两个词组成的组合	中国政府
$W_{-1}W_0$	前一个词和当前词组成的组合	政府顺利
W_0W_1	当前词及其后一个词组成的组合	顺利恢复
W_1W_2	当前词的后面两个词组成的组合	恢复对
$W_{-1}W_1$	前一个词和后一个词组成的组合	政府恢复
$T_{-1}T_0$	相邻两个词的词性转移特征	n→ad

2.2.2　上下文特征产生的内在机理

在统计语言建模中,自然语言处理被看作是一个随机过程,图 3 是汉语词性标注作为一个随机过程的示意图.图中每个"时刻"上下文窗口中的内容是确定的,是该随机过程的一个样本,并且窗口是随"时间"逐词滑动的,图中示意了上下文窗口经过若干次滑动得到的另一个样本.上下文窗口大小和特征模板确定之后,模型训练过程中根据设定的特征模板集可以从训练语料中扩展出成千上万的上下文特征.显然,图 3 中示意的上下文窗口是"5 词窗口",对于第一个上下文样本"中国/ns 政府/n 顺利/ad 恢复/v 对/p"来说,当前词是"顺利",从该样本扩展出的上下文特征见表 1 第 3 列.从一个样本中每个特征模板将会扩展出一个特征,这产生大量特征.如果训练语料中不同的词语有 3 万个,对于单个词语特征模板来说,都会产生 3 万个上下文特征.对于双词语特征模板来说,在语料充足时理论上将会产生大约 9 亿个上下文特征.而对于 CRF++工具包来说,产生的上下文特征个数实质是特征函数个数,是上下文特征和语料中词性标记种类个数的乘积.由此可见,基于 CRFs 的汉语词性标注中将会产生数以亿计的特征,过多的特征导致模型训练时间过长,甚至无法进行或崩溃.

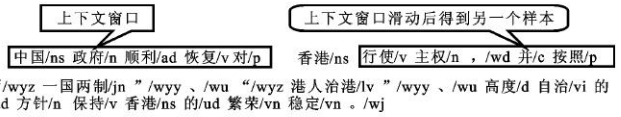

图 3　汉语词性标注作为一个随机过程的示意图

2.3　基于 CRFs 汉语词性标注中特征模板集的优化

在已有的汉语词性标注研究中,所使用的特征模板集不仅包括单个词语特征模板,也包括双词语特征模板[4],见表 2 中的序号为 1 和 4 的特征模板集.采用 CRF++使用它们实现汉语词性标注时,训练语料稍大则会出现无法训练的现象.针对这一问题,本文提出舍弃双词语特征模板,仅使用单词语特征模板的优化思路,并使用表 2 列出的其他特征模板集在 Bakeoff2007 的 PKU、NCC、CTB 三种语料上进行对比实验.表 2 中,后缀"Single"表示相应特征模板集中的只有单个词语构成的特征模板.

表 2　特征模板集列表

序号	模板集名称	包含的特征模板
1	TMPT-10+B	$W_{-2}, W_{-1}, W_0, W_1, W_2, W_{-2}W_{-1}, W_{-1}W_0, W_0W_1, W_1W_2, W_{-1}W_1, T_{-1}T_0$
2	T10-Single+B	$W_{-2}, W_{-1}, W_0, W_1, W_2, T_{-1}T_0$
3	T10-Single	$W_{-2}, W_{-1}, W_0, W_1, W_2$
4	TMPT-6+B	$W_{-1}, W_0, W_1, W_{-1}W_0, W_0W_1, W_{-1}W_1, T_{-1}T_0$
5	T6-Single+B	$W_{-1}, W_0, W_1, T_{-1}T_0$
6	T6-Single	W_{-1}, W_0, W_1

3　实验结果及其分析

我们设计了两组实验,分别从不同的角度对基于 CRFs 的汉语词性标注进行了对比研究.(1)模型训练过程反映出的"量"属性.本组实验关注的是不同特征模板集对模型训练的影响,主要从不同模板集扩展出的特征数、是否能进行训练、模型训练时间、训练出的模型大小等几个"量"化指标进行考察.(2)不同特征模板集对汉语词性标注性能的影响,采用了标注精度(Accuracy)评测指标.

我们首先分别使用表 2 中序号为 1～6 的六组特征模板集,在 PKU、NCC 和 CTB 三个语料集上进行了汉语词性标注的训练,表 3 中给出了使用这六组特征模板集的训练过程记录数据.从表 3 可见,优化后的四组特征模板集能够顺利进行训练,第二组实验则采用训练出的这些模型对测试语料进行词性标注,表 4 中给出了这四组特征模板集训练出的模型在相应的测试语料上的汉语词性标注性能.

综合分析表 3、表 4 中的数据可以得出如下结论:(1)优化后的特征模板集扩展出的特征数要比传统特征模板集扩展出的特征数少的多,也使模型能够训练.(2)同等条件下,训练出的模型大小与扩展出的特征数成正比.(3)对照 Bakoff2007 的评测结果,优化后的特征模板集在模型训练时间、大小、词性标注精度等指标上达到了整体最优.

表 3 PKU、NCC 和 CTB 语料上的训练过程记录数据

模板集序号	PKU 语料训练数据			NCC 语料训练数据			CTB 语料训练数据		
	特征数	训练时间/s	模型大小/MB	特征数	训练时间/s	模型大小/MB	特征数	训练时间/s	模型大小/MB
1	145 224 232	无法训练		74 761 652	无法训练		52 577 888	无法训练	
2	17 614 133	186 178.49	69.7	10 179 464	94 819.77	43.8	6 606 572	51 617.17	30.0
3	17 603 524	152 027.86	69.7	10 176 100	74 384.22	43.8	6 605 203	39 865.86	30.0
4	91 420 534	无法训练		47 666 314	无法训练		33 512 676	无法训练	
5	10 997 310	148 211.24	44.3	6 555 566	84 061.73	28.2	4 228 841	47 858.08	19.2
6	10 986 701	113 086.12	44.3	6 552 202	54 584.97	28.2	4 227 472	32 182.33	19.2

表 4 不同特征模板集的词性标注结果

模板集序号	特征模板集名称	PKU 语料上评测结果	NCC 语料上评测结果	CTB 语料上评测结果
		Accuracy/%	Accuracy/%	Accuracy/%
2	T10-Single+B	94.12	90.05	91.94
3	T10-Single	93.90	90.02	91.71
5	T6-Single+B	94.78	90.48	92.07
6	T6-Single	94.52	90.53	91.65

4 结束语

汉语词性标注中应用条件随机场对上下文建模时会扩展出数以亿计的上下文特征,如此多的特征使得模型训练的时间过长甚至无法进行.针对这一问题,本文通过深入分析上下文特征产生的内在机理,对采用 CRF++工具包实现的汉语词性标注所使用的特征模板集进行了优化,对比实验表明,优化后的特征模板集在模型训练时间、训练后模型大小、词性标注精度等指标上达到了整体最优.

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] 洪铭材,张阔,唐杰,等.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学,2006,33(10):148-155.

[3] Pereira L J, Mccallum F A. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc. of 18th ICML. San Francisco, USA: AAAI Press, 2001: 282-289.

[4] 于江德,睢丹,樊孝忠. 基于字的词位标注汉语分词[J]. 山东大学学报:工学版,2010, 40(5):117-122.

[5] 冯冲,陈肇雄,黄河燕,等. 基于条件随机域的复杂最长名词短语识别[J]. 小型微型计算机系统,2006,27(6):1134-1139.

[6] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006,34(5):804-809.

[7] 于丽丽,德鑫,曲维光,等.基于条件随机场的古汉语词义消歧研究[J].微电子学与计算机,2009,26(10):45-48.

作者简介:

于江德 男,(1971-),博士,副教授.研究方向为计算语言学、中文信息处理、文本信息抽取等.

(上接第 62 页)

[2] 程显毅,朱倩,王进.中文信息抽取原理与应用[M].北京:科学技术出版社,2010.

[3] 刘小东.自然语言理解综述[J].统计与信息论坛,2007,22(2):5-12.

[4] 任洁.自然语言与自然语言理解及其应用[J].科教文汇,2006:69-70.

[5] 高文利,李德华. 对自然语言理解的思考[J].牡丹江大学学报,2007,16(5):75-76.

[6] 程显毅,朱倩.文本挖掘原理[M].北京:科学技术出版社,2010.

作者简介:

潘 燕 女,(1988-),硕士研究生.研究方向为自然语言处理.

程显毅 男,(1956-),博士,教授.研究方向为人工智能.