基于层叠条件随机场模型的中文机构名自动识别

周俊生[1,2], 戴新宇[1], 尹存燕[1], 陈家骏[1]
(1. 南京大学计算机软件新技术国家重点实验室, 江苏南京 210093; 2. 南京师范大学计算机科学系, 江苏南京 210097)

摘 要: 中文机构名的自动识别是自然语言处理中的一个比较困难的问题. 本文提出了一种新的基于层叠条件随机场模型的中文机构名自动识别算法. 该算法在低层条件随机场模型中解决对人名、地名等简单命名实体的识别, 将识别结果传递到高层模型, 为高层的机构名条件随机场模型实现对复杂机构名的识别提供决策支持. 文中为机构名条件随机场模型设计了有效的特征模板和特征自动选择算法. 对大规模真实语料的开放测试中, 召回率达到 90.05%, 准确率达到 88.12%, 性能优于其它中文机构名识别算法.

关键词: 命名实体; 中文机构名识别; 条件随机场
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2006) 05-0804-06

Automatic Recognition of Chinese Organization Name Based on Cascaded Conditional Random Fields

ZHOU Jun-sheng[1,2], DAI Xin-yu[1], YIN Cun-yan[1], CHEN Jia-jun[1]
(1. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China;
2. Department of Computer Science, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: Automatic recognition of Chinese organization name is a very difficult problem in many NLP tasks. This paper presents a new algorithm of Chinese organization name recognition based on cascaded conditional random fields. In the proposed algorithm, the person name and location name are first recognized by the lower model. The result then is passed to the high model and supports the decision of high model for recognition of the complicated organization names. We experimentally evaluate the algorithm on large-scale corpus. In open test, its recalling rate achieves 90.05% and the precision rate 88.12%. The evaluation results show that the algorithm based on cascaded conditional random fields significantly outperforms previous methods.

Key words: named entity; Chinese organization name recognition; conditional random fields

1 引言

命名实体的识别是许多自然语言处理任务的基本要求, 如信息抽取、机器翻译、文本摘要、主题发现与跟踪等. 近年来, 中文人名、地名的识别研究已经取得了较大的进展, 而对中文机构名识别目前还未能获得较好的效果. 2004年度国家 863中文信息处理与智能人机接口技术评测的命名实体识别评测结果显示: 中文机构名识别的召回率仅为 57.41%, 准确率仅为 64.64%. 这表明对中文机构名的识别研究目前仍处在探索阶段.

相对于中文人名、地名的识别来说, 中文机构名的识别存在较大的困难. 目前, 有关中文机构名识别的研究相对较少, 主要使用一些规则方法和隐马尔可夫模型. 文[1]和文[2]则提出了基于启发式规则的机构名识别方法, 虽然取得了一定的效果, 但论文所报告的测试结果只是基于一个很小规模的测试数据集. 机构名由于种类繁多, 对各类机构名要总结出统一的识别规则, 这基本上是不可行的. 文[3]提出了一种基于隐马尔可夫模型的角色标注方法识别中文机构名. 但由于隐马尔可夫模型是一种产生性(generative)模型, 它存在一些固有缺陷与不足[4]. 在产生性模型中, 为保证推导的正确性, 需要作出严格的独立性假设. 事实上, 大多数序列数据都不能被表示成一系列独立的元素. 条件随机场 (Conditional Random Fields, CRFs) 则是一种新的概率图模型[5], 它具有表达元素长距离依赖性和交叠性特征的能力, 能方便地在模型中包含领域知识, 且较好地解决了标注偏置问题等优点. 文[6]显

收稿日期: 2005-04-08; 修回日期: 2005-12-12
基金项目: 国家 863高技术研究发展计划 (No. 2004AA117010-05); 江苏省教育厅基金 (No. 03KJD520117)

示,该模型在解决英文命名实体的识别任务时,具有较好的性能.但中文机构名识别任务与英文机构名识别之间还存在较大的差异:首先,中文句子是未经切分的连续字符序列,对包含命名实体的中文句子进行切分经常会引发切分歧义;其次,中文机构名不具有任何明显的边界标记符号,而且汉语的表达方式相对英语而言,更加灵活而多变.针对中文机构名识别的困难与特点,本文提出了一种新的基于层叠条件随机场模型的中文机构名识别算法,对各粗分词串先在低层进行人名与地名的识别,将识别结果传递到高层模型,为高层机构名条件随机场模型对复杂机构名的识别提供决策支持.最后采用约束的前向后向算法对识别的结果进行可信度计算.实验结果显示,该算法明显优于其它中文机构名识别算法的效果.

2　条件随机场

条件随机场是一种用于在给定输入结点值时计算指定输出结点值的条件概率的无向图模型.若 O 是一个值可以被观察的"输入"随机变量集合,S 是一个值能够被模型预测的"输出"随机变量的集合,且这些输出随机变量之间通过指示依赖关系的无向边所连接.让 $C(S,O)$ 表示这个图中的团的集合,CRFs将输出随机变量值的条件概率定义为与无向图中各个团的势函数(potential function)的乘积成正比:

$$P_A(s|o)=\frac{1}{Z_o}\prod_{c\in C(s,o)}\Phi_c(s_c,o_c) \tag{1}$$

其中,$\Phi_c(s_c,o_c)$ 表示团 c 的势函数.当图形模型中的各输出结点被连接成一条线性链的特殊情形下,CRFs假设在各个输出结点之间存在一阶马尔可夫独立性,二阶或更高阶的模型可类似扩展.若让 $o=(o_1,o_2,\cdots,o_T)$ 表示被观察的输入数据序列,让 $s=(s_1,s_2,\cdots,s_T)$ 表示一个状态序列.在给定一个输入序列的情况下,线性链的 CRFs定义状态序列的条件概率为:

$$P_A(s|o)=\frac{1}{Z_o}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \tag{2}$$

其中,$f_k(s_{t-1},s_t,o,t)$ 是一个任意的特征函数,λ_k 是每个特征函数的权值.归一化因子

$$Z_0=\sum_s\exp\left(\sum_{t=1}^{T}\sum_k\lambda_k f_k(s_{t-1},s_t,o,t)\right)$$

3　基于层叠条件随机场的中文机构名识别算法

3.1　层叠条件随机场模型

由于中文机构名中存在较多的嵌套现象,即机构名中可能会包含有人名、地名,这种嵌套性给中文机构名的识别带来了一定的复杂性.一方面,这些被嵌套的人名、地名中的字或词很可能与上下文组合成词,会引发句子切分的歧义性;另一方面,在这些被嵌套的人名、地名中包含的单字或未登录词又会干扰机构名的识别准确性.因此,需要引入多个层次的条件随机场模型用于识别这类嵌套机构名.当前建立多层模型主要有两种不同的方法,一种方法是按层叠加建立模型[7,8],多个模型之间呈线性组合,称之为层叠模型;另一种方法采用递归方式建立模型[9],低层模型被嵌入为高层模型的一个子模型,称之为层次模型.相对于层叠模型,层次模型具有更复杂的数学模型,其训练复杂度和解码复杂度也远大于层叠模型.而在层叠模型中,不同层次模型间是一种松耦合的关系,各层模型可以独立地建立,整个模型的复杂度与句子的长度成线性关系.而且在层叠模型中,低层模型所产生的错误可以经过适当的过滤和调整后,再将结果传递到高层模型,从而可以避免错误的传播和扩散.基于以上考虑,本文提出了一种基于层叠条件随机场(CCRFs)模型(如图 1所示)的中文机构名识别算法.在 CCRFs模型中,低层的条件随机场模型仅以观察值为条件,用于人名、地名等简单命名实体的识别,识别的结果再传递到高层模型,这样高层模型的输入变量将不仅包含观察值,而且包含了来自低层模型的识别结果,从而为高层条件随机场模型对复杂机构名的识别提供了决策支持.各个层次条件随机场模型的训练都是以《人民日报》标注语料库作为训练语料,但需要设计相应的训练语料生成算法对原切分语料进行适当的改造以应用于不同层次条件随机场模型的训练.

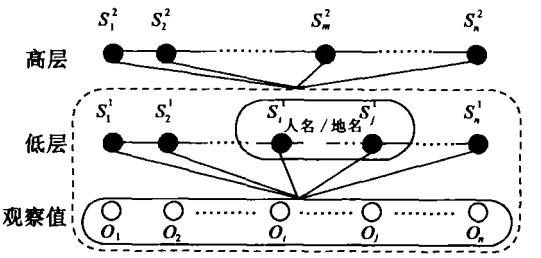


图 1　层叠条件随机场模型(图中实心圈表示状态,空心圈表示观察值)

3.2　基于 CCRFs的中文机构名识别算法框架

机构名识别算法的输入是未经切分的中文字符串,输出是所识别出的机构名以及人名与地名.图 2是基于 CCRFs模型的中文机构名识别算法流程图.由于中文机构名中包含的单字可能分别会与机构名的上、下文组合成词,为解决切分歧义问题,本文采用了 *N*-最短路径的切分排歧策略[10],即基于 *N*-最短路径的统计粗分模型对输入的未切分的句子进行粗分,得到一组 *N-best*的粗分词串序列(通过实验可知,N取值为 6较合适);然后在低层的条件随机场模型中对 N 个粗分词串序列分别进行人名和不包含复杂嵌套的地名识别,并对所识别出的人名或地名字串加上相应的标注(*nr*和 *ns*).由于在人名和地名的识别阶段会不可避免的引入一些识别错误,因此在进入机构名识别之前,基于转换的思想[11]引入一些规则对所识别出的人名、地名字串进行过滤操作,以减少错误的传播.在此基

础上再应用高层的机构名条件随机场模型进行机构名识别,最后采用约束的前向后向算法[12]对识别的结果进行可信度计算,选择具有最大可信度的词串作为最终的识别结果.

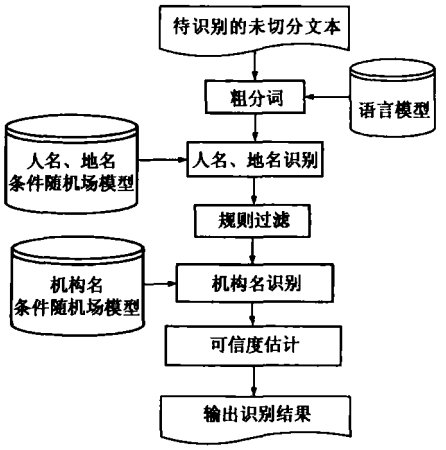


图2 基于层叠条件随机场模型的机构名识别流程

在基于层叠条件随机场模型的中文机构名识别算法中,由于低层的人名、地名识别模型相对机构名模型较为简单,限于篇幅,本文不再详细阐述.在人名、地名识别的基础上进行机构名识别时,将机构名识别问题转化为标注问题,应用训练好的机构名条件随机场模型对由低层模型识别输出的各个粗分词串分别进行标注.对于这个序列化的标注过程,我们采用了 Viterbi算法[13]推导出最大概率的状态序列,但这里需要将 Viterbi算法中的概率值的迭代公式修改为:

$$\delta_{t+1}(s_i)=\max_{s'}\left[\delta_t(s')\exp\left(\sum_k \lambda_k f_k(s',s,o,t)\right)\right] \quad (3)$$

对于标注集,我们并没有采用常规的"BIO"标注方法,而是引入了4个标注符号:B,I,F,O,其中,B为机构名开始,I为机构名内部,F为机构名尾部,O为其他.例如粗分词串"河流镇 /ns出让原镇政府大礼堂建起的阳信 /ns占峰餐具有限公司,"(标注 ns表示由低层模型识别出的地名),用这4个符号标注得到下面的结果:

河流镇 /ns	出让 /O	原 /O	镇政府 /O	大礼堂 /O	建 /O	起 /O
的 /O	阳信 /B	占 /I	峰 /I	餐具 /I	有限公司 /F	

3.3 训练数据的生成算法

机构名条件随机场模型的训练语料来自于切分标注好的《人民日报》语料.在该语料库中,机构名全部被显式标注了.按照标注的形式,其中的机构名可以被分为两类:简单机构名(如"致公党 /nt")和复合机构名(如"[美国 /ns加利福尼亚 /ns 理工学院 /n] nt").由于机构名识别的输入是句子的粗分词串,其切分粒度与标注语料中的机构名切分存在较大差异,尤其标注语料中的简单机构名经常在粗分词串中被切成多个片段,如"致公党"会被切成"致 / 公 党",因此在改造标注语料生成训练语料时,应根据机构名字串可能的粗分粒度对标注语料中的机构名切分进行相应的调整,具体算法描述如下:

(1)从标注语料中依次读入按词性标注好的句子,根据机构名标注 nt定位出机构名;

(2)依据标注 nt前面的符号,区分当前机构名的类别,若是简单机构名,转步骤(3),否则转步骤(4);

(3)判断当前简单机构名是否为已登录词,若是,将其词性标记改为 B;否则,判断当前简单机构名是否包含机构名特征词,若包含特征词,则取出最长特征词,并将其尾部的词性标注改为 F,且对剩余字串按正向最大匹配法进行切分;若不包含特征词,则按正向最大匹配法直接对该简单机构名进行切分,在切分后的第一个词后面插入标注 B,其它词后都插入标注 I,然后转步骤(6);

(4)取出当前复合机构名中的最后一个词,首先判断该词是否为机构名特征词,若是,直接将该词的词性标注改为 F;否则,判断当前词是否包含机构名特征词,若包含特征词,则取出最长特征词,并将其尾部的词性标注改为 F,且对该词中的剩余字串按正向最大匹配法进行切分;若不包含特征词,则按正向最大匹配法对该词进行切分,在切分后的各个词后面都插入标注 I;

(5)按从后往前的次序依次取出复合机构名的各个词,首先判断当前词是否为人名或地名或已登录机构名,若是,增加相应的类别标记,同时插入标注 I(若是第一个词,则插入标注 B),否则,判断当前词是否为已登录词,若是,在该词的后面插入标注 I(若是第一个词,则插入标注 B),否则,按正向最大匹配法对当前词进行切分,在切分出的各个词后面插入标注 I(若是第一个词,则插入标注 B);

(6)将机构名以外的其它词的词性标注全部替换为标注 O.

3.4 基于约束前向后向算法的机构名识别可信度估计

前向后向算法可用于在给定观察值序列的条件下计算所有可能的状态序列的概率.在条件随机场中需要定义一个作相应修改的"前向变量"$\alpha_t(s_i)$,递归定义如下:

$$\alpha_{t+1}(s)=\sum_{s'}\alpha_t(s')\exp\left(\sum_{k=1}^{K}\lambda_k f_k(s',s,o,t)\right) \quad (4)$$

为了估计在每一个粗分词串中所识别出机构名的可信度,对前向后向算法作如下约束:每条路径均应经过满足约束 $C=\langle s_i, s_{i+1}, \cdots\rangle$的子路径,其中,$s_i\in C$或者是一个正约束(该序列应该经过 s_i)、或者是一个负约束(该序列不应经过 s_i).在机构名识别中,约束 C对应于被识别出的机构名,C中的正约束表示机构名的内部,而负约束则指定机构名的边界.在约束前向后向算法中,同样前向变量值也应该满足约束 C,基于条件随机场模型,对所有的 $s_q\in C$,定义相应的约束前向变量如下:

$$\alpha_q(s_i)=\begin{cases}\max_{s'}\left[\alpha_{q-1}\exp\left(\sum_k \lambda_k f_k(s',s,o,t)\right)\right], & \text{if } s_i\subset s_q\\ 0, & \text{otherwise}\end{cases} \quad (5)$$

式中, 符号 $s_i \supset s_q$ 表示 s_i 满足约束 s_q. 如果 $\alpha'_{t+1}(s_i)$ 是一个约束前向变量, $Z'_0 = \sum_i \alpha'_T(s_i)$ 就是约束格(lattice)的值, 则对一个句子粗分词串的机构名识别可信度可表示为 Z'_0 的归一化, 即 Z'_0 / Z_0, $Z_0 = \sum_i \alpha_T(s_i)$.

4 特征的自动选择与特征参数的训练

4.1 特征模板的构造

条件随机场模型中一个非常重要的因素是如何针对特定的任务为模型选择合适的特征集合, 用简单的特征表示复杂的语言现象. 由于中文机构名的构成方式非常复杂, 从机构名的结构上看, 大多数机构名除了具有机构名特征词(如公司、学校等等)这一特征外, 并无其它共性的结构特征. 而机构名作为一种专有名词, 一般有一定的上下文语言环境. 机构名的上下文信息主要包括指界词(主要为动词和副词)和称谓词(如"局长"等). 通过对大量机构名语料的分析可知, 上下文信息中的各种指界词与机构名同现的概率相差较大, 这说明不同指界词对出现机构名的指示信息的强弱程度相差较大. 为此, 本文依据指界词对出现机构名的指示信息强度将左、右指界词各分为二级(如表 1所示), 并构造了四个相应的指界词词表. 对指界词的选取与分级, 本文采用了基于互信息计算的方法. 在自然语言处理中, 互信息 $I(x, y)$ 常被作为描述两个字或两个词之间关联程度大小的量度. 而本文则利用互信息计算各个指界词与机构名类别的关联强度, 从而实现了对指界词的有效选取与分级.

表 1　左、右边界信息的分级

类型	级别	举例(互信息值)
左边界	1级	历任(6.0006)
指界词	2级	接管(3.1161)
右边界	1级	管辖(5.4531)
指界词	2级	规定(2.0135)

表 2　原子特征模板

序号	原子模板(函数)	模板意义
1	CurWord	当前词
2	LocationName	当前词是否为地名
3	PersonName	当前词是否为人名
4	KnownOrganization	当前词是否为已登录机构名
5	OrganizationFeature	当前词是否为机构名特征词
6	ScanFeatureWord_8	扫描当前词的后 8个词中是否含机构名特征词
7	LeftBoundary1_2	当前词的前面 2个词中是否含 1级左指界词
8	LeftBoundary2_2	当前词的前面 2个词中是否含 2级左指界词
9	RightBoundary1_+2	当前词的后面 2个词中是否含 1级右指界词
10	RightBoundary2_+2	当前词的后面 2个词中是否含 2级右指界词
11	ForbiddenWord	当前词是否为机构名禁止用词
12	CurTag	当前词的标注类别
13	CurTag_1	当前词的前一个词的标注类别

为提高机构名识别的准确率, 我们还引入了机构名禁止用词特征. 有些词是不允许在机构名中出现的, 主要是一些形容词和副词等, 如"即使"、"当然"等. 因此我们把它们集中起来建立了"机构名禁止词表". 根据以上考虑, 我们定义了模型中的特征模板, 如表 2所示. 在这个表中每个模板只考虑一种因素, 故称之为原子模板.

为增加对上下文信息的描述, 还需将上述各特征模板分别进行 -2, -1, 1, 2四个位置的偏移. 这些特征可以表示为二值特征函数的形式.

4.2 特征参数的训练方法

条件随机场的训练目标是在给定一个训练数据集 $D = \{\langle o, l\rangle^{(1)}, \cdots, \langle o, l\rangle^{(j)}, \cdots, \langle o, l\rangle^{(N)}\}$ 的条件下, 最大化训练集的对数似然(log-likelihood):

$$L_\Lambda = \sum_{j=1}^{N} \lg(P_\Lambda(l^{(j)} \mid o^{(j)})) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \qquad (6)$$

式中的第二项是用于提供平滑处理的特征参数的高斯先验值, σ^2 表示先验方差. 本文使用 L-BFGS算法实现对目标函数的优化求解. L-BFGS[14]是一种充分利用以前的梯度和修改值来近似曲率值的二阶方法, 可以避免准确的 Hessian矩阵的逆矩阵的计算. 因而使用 L-BFGS 算法进行 CRF训练只要求提供似然函数的一阶导数, 假定第 j个训练实例的标注使它的状态序列不产生二义性, 且 $s^{(j)}$ 表示那条路径, 则训练数据集的对数似然的一阶导数为:

$$\frac{\partial L}{\partial \lambda_k} = \left(\sum_{j=1}^{N} C_k(s^{(j)}, o^{(j)})\right) - \left(\sum_{j=1}^{N}\sum_{s} P_\Lambda(s \mid o^{(j)}) C_k(s, o^{(j)})\right) \frac{\lambda_k}{\sigma^2} \qquad (7)$$

式中, $C_k(s, o)$ 表示特征 f_k 在串 s中各个位置 t的和, 式中的前两项相应于特征 f_k 的经验期望值 $E[f_k]$ 与关于模型的期望值 $E_\Lambda[f_k]$ 的差, 对它们的计算, 可采用动态规划算法高效地实现.

4.3 特征的自动选择算法

根据特征模板可以自动从语料中生成相应的特征集合, 然而这个特征集合中的特征数量十分庞大. 在这些特征中不可避免地会存在一些"噪声"特征数据, 影响模型的训练和识别速度与效果, 因而需要设计有效的特征选择算法. 文[15]设计了一种用于条件随机场的特征自动选择和归纳算法, 该算法是在初始原子特征集合的基础上, 先通过对候选特征的初选和高增益特征间的连接构造复合特征, 形成候选特征集, 在此基础上再基于增益同时进行原子特征和复合特征的选择. 但是针对中文机构名的识别问题, 我们在实验中发现该算法计算量非常巨大, 几乎实际不可行. 因此我们在分析了大量的基于原子特征的识别错误实例的基础上, 设计了如表 3所示的复合特征模板, 以构造有效的复合特征来表示复杂的机构名上下文环境. 并在初始的原子特征和复合特征的基础上, 我们设计了基于近似增益计算的特征自动选择算法, 这样就产生一个几乎没有计算浪费的无噪声特征集合. 具体算法描述如下:

(1)将模型的初始选择特征集设置为空;

(2)计算每一候选特征的增益;

(3)选出增益值大的一组(200个)候选原子特征或复合特征, 并将它们加入到选择特征集中;

(4)使用 L-BFGS算法重新调整选择特征集中的各特征参数值,为避免部分特征训练过拟合,将 L-BFGS算法的迭代次数设定为 10次;

(5)对步骤(2)~(4)迭代,直至收敛.

为避免计算机量过大,候选特征增益的计算采用近似计算的方法[15],对于新增候选特征,增益的近似计算公式如下:

$$G_{\Lambda}(g,\mu)=\sum_{i=1}^{M}\log\left(\frac{\exp(\mu g(s_{(i)},o(i),t(i)))}{Z_{O(i)}(\Lambda,g,\mu)}\right)-\frac{\mu^2}{2\sigma^2} \tag{8}$$

式中 σ^2 为高斯先验方差,参数 μ 的最优值可通过牛顿方法求解,然后再代入上式,即求得候选特征 g 的增益.

表 3 复合特征模板

序号	复合模板
1	LeftBoundary_-1 & LeftBoundary_-2
2	RightBoundary_+1 & RightBoundary_+2
3	LeftBoundary_-1 & ScanFeatureWord_8
4	LeftBoundary_-1 & LeftBoundary_-2 & ScanFeatureWord_7
5	OrganizationFeature & RightBoundary_+1
6	OrganizationFeature & RightBoundary_+1&RightBoundary_+2

5 实验结果与分析

根据测试集和训练集的不同关系,可以将评测分为封闭测试和开放测试.为了能够客观评价基于层叠条件随机场模型的中文机构名识别算法的识别效果,我们做了四组识别实验,其中前两组实验是基于层叠条件随机场模型的封闭测试与开放测试,后两组是基于单层模型的封闭测试与开放测试.我们作封闭测试和开放测试所用的训练语料都是《人民日报》九八年一月的语料,封闭测试时所用的测试语料仍是《人民日报》九八年一月的语料,开放测试时所用的测试语料是九八年二月的语料.测试语料中既含有大量包含机构名的句子,同时也包含有大量不含机构名的句子,接近真实的语言环境.测试的结果采取了常用的 3个评测指标,即准确率(P)、召回率(R)和综合指标 F 值(F)来评测机构名识别的结果.

表 4 层叠条件随机场模型与单层模型的实验结果比较

	测试类型	实际机构名数	识别出的机构名数	正确识别数	准确率(%)	召回率(%)	$F1$值(%)
层叠条件随机场模型	封闭测试	10586	10746	10311	95.95	97.40	96.67
	开放测试	10605	10836	9549	88.12	90.05	89.07
单层模型	封闭测试	10586	10720	9896	92.31	93.48	92.89
	开放测试	10605	10762	9156	85.07	86.34	85.70

从实验结果可以看出,基于层叠条件随机场模型的识别效果相对于单层模型有了较大的改进,改进的效果主要是体现在对以未登录人名、地名开头的机构名的识别中.如机构名"伏尔加格勒罗特尔队"、"巴伊亚州府萨尔瓦多市立天主教大学"、"门头沟区东辛房街道生活服务总站创创科技开发部"等在单层模型中就无法被正确的识别出来,单层模型对这种类型的机构名识别往往无能为力.另外在实验中我们也发现,由于我们的特征模板和特征自动选择算法设计较合理,使得模型不仅适用于简单机构名识别,对一些长的复杂机构名也特别有效,如机构名"马鞍山钢铁总公司利民公司铁路建筑工程公司"、"联合国销毁伊拉克化学、生物和核武器特别委员会"、"意大利国家科研委员会国际基因和生物物理研究所"等均能被正确识别.

据我们所查到的资料,对大规模真实语料进行了中文机构名识别实验的主要有文[3,16],所用训练和测试语料也均为九八年人民日报语料.文[16]中设计了一个多层最大熵模型并将其应用于中文机构名的识别实验,识别准确率为 82.1%,而召回率为 53.8%.文[3]中报告的封闭测试召回率为 87.14%,准确率为 89.60%,开放测试准确率为 88.39%,召回率为 75.82%,他们采用的测试语料和我们的实验基本相同,而我们实验的训练语料仅用了一个月的标注语料,实验结果却明显高于他们的指标.

6 结论

中文机构名的识别一直是一个比较困难与有挑战性的问题.本文针对中文机构名的特点,提出了一个基于层叠条件随机场模型的机构名自动识别算法,并设计了有效的特征模板和条件随机场特征自动选择算法,通过对大规模真实语料的封闭与开放测试显示,该方法取得了相当好的效果.但我们在实验中也发现当前的条件随机场参数估计算法存在对训练数据的过拟合问题,这是由于现有的条件随机场参数估计算法都是以最大化训练数据的对数似然值为准则,结果导致训练过拟合问题.下一步我们将基于大间隔(Margin)的思想设计新的条件随机场的参数训练算法,以进一步提高算法识别的效果.

参考文献:

[1] 张小衡,王玲玲.中文机构名称的识别与分析[J].中文信息学报,1997,11(4):21-32

[2] Wang Houfeng, Shi Wuguang. A simple rule-based approach to organization name recognition in chinese text[A]. Proc of 5th CIPS-ICCLing[C]. LNCS 3406, Heidelberg, German: Springer Verlag, 2005. 769-772.

[3] Hongkui Yu, Huaping Zhang, Quan Liu. Recognition of Chinese organization name based role tagging[A]. Proc of Advances in Computation of Oriental Languages[C]. Beijing: Tsinghua University Press, 2003. 79-87.

[4] McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation[A]. Proc of 17th ICML[C]. Stanford, California, USA: Morgan Kaufmann, 2000. 591-598.

[5] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and

labeling sequence data[A]. Proc of the 18th ICML[C]. San Francisco: Morgan Kaufmann, USA: 2001. 282 - 289.

[6] Andrew McCallum, Wei Li. Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons[A]. Proc of the 7th CoNLL[C]. Edmonton, Canada: Morgan Kaufmann, 2003. 188 - 191.

[7] Thorsten Brants. Cascaded Markov models[A]. Proc of EACL'99[C]. Bergen, Norway: European Chapter of the Association for Computational Linguistics, 1999. 118 - 125.

[8] 刘群, 张华平, 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.

[9] M Skounakis, M Craven, S Ray. Hierarchical hidden markov models for information extraction[A]. Proc of the 18th International Joint Conference on Artificial Intelligence[C]. Acapulco, Mexico: Morgan Kaufmann, 2003. 427 - 433.

[10] 张华平, 刘群. 基于 N 最短路径的中文词语粗分模型[J]. 中文信息学报, 2002, 16(5): 1 - 7.

[11] Eric Brill. Transformation based error driven learning and natural language processing: A case study in part of speech tagging[J]. Computational Linguistics, 1995, 21(4): 543 - 566.

[12] Aron Culotta, Andrew McCallum. Confidence estimation for information extraction[A]. Proc of HLT-NAACL'04[C]. Boston, Massachusetts: Association for Computational Linguistics, 2004. 109 - 112.

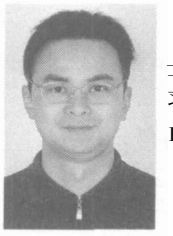
[13] Rabiner L A. tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 - 286.

[14] J Nocedal, S J Wright. Numerical Optimization[M]. Springer, 1999.

[15] Andrew McCallum. Efficiently inducing features of conditional random fields[A]. Proc of the 19th Conf on Uncertainty in Artificial Intelligence[C]. Acapulco, Mexico: Morgan Kaufmann, 2003. 403 - 410.

[16] Deyi Xiong, et al. Qun Liu. Tagging complex NEs with maxent models: Layered structures versus extended tagset[A]. Proc of the 1th IJCNLP[C]. Hainan Island: Springer, 2004. INAI 3248: 537-544.

作者简介:



周俊生　男, 1972年 3月生于安徽枞阳县, 博士研究生, 主要研究方向为自然语言处理、机器学习、信息抽取.
E-mail: zhoujs@nlp.nju.edu.cn

戴新宇　男, 1979年 2月生于江苏盱眙县, 博士研究生, 主要研究方向为机器翻译、信息检索.

尹存燕　女, 1976年 6月生于南京, 博士研究生, 主要研究方向为自然语言处理、机器翻译.

陈家骏　男, 1963年 10月生于南京, 教授, 博士生导师, 主要研究方向为自然语言处理、机器翻译、软件工程.