

文章编号: 1000-8608(2010)05-0782-06

基于 SVM 和 CRF 的双层模型中文机构名识别

黄德根*, 李泽中, 万如

(大连理工大学 计算机科学与技术学院, 辽宁 大连 116024)

摘要: 提出了一种基于支持向量机(SVM)和条件随机场(CRF)的双层模型进行中文机构名识别的方法。第一层模型采用CRF识别简单机构名, 并将识别结果传至第二层辅助下一步的识别; 第二层采用基于驱动的方法, 将SVM和CRF结合进行复杂机构名的识别; 最后将两层的识别结果合并, 并通过一个后续处理对置信度较低的识别结果进行修正。大规模真实语料的开放测试表明, 精确率达到94.83%, 召回率达到95.02%, 证明了该方法的有效性。

关键词: 机构名识别; 条件随机场(CRF); 支持向量机(SVM); 双层模型

中图分类号: TP301.6 文献标志码: A

0 引言

命名实体识别是许多自然语言处理任务的基本要求, 其识别效果直接影响文本信息的深层次处理。机构名识别是命名实体识别的主要任务之一。与人名和地名相比, 机构名具有长度较长而且不固定、用词复杂并且未登录词较多、具有嵌套结构等特点, 因此其识别难度相对较大。

早期的机构名识别采用的多是基于规则的方法。文献[1]针对高校名称建立了一个规则模型, 而规则的获取往往依赖于特定的领域, 成为该类方法的瓶颈; 文献[2]采用决策树的方法进行命名实体识别, 但识别精度较低; 文献[3、4]采用隐马模型(HMM)进行命名实体识别, 该模型需要严格的独立性假设, 而事实上绝大多数的数据并不能表示为一系列独立的元素; 文献[5]采用SVM进行命名实体识别; 文献[6、7]采用CRF进行机构名识别, 识别效果比较理想, 但仍有改进的余地; 文献[8]提出了一个基于角色标注的方法, 不足之处是角色集对实验结果影响较大, 需要反复实验才可以确定合适的角色集; 文献[9]将机器学习和人工知识结合起来进行机构名识别。

本文将机构名分为简单机构名和复杂机构名两大类。简单机构名即仅由一个词组成的机构名, 如新华社、国安队、中共中央等; 复杂机构名即由

多个词组成的机构名, 可定义为 $P + S$ 的形式, P 为机构名前部词, S 为机构名特征词(如公司、大学等), 即复杂机构名是由一个或一个以上的机构名前部词加上机构名特征词组成的。

1 SVM 与 CRF

1.1 支持向量机(SVM)

假设原始输入空间 $X \subseteq \mathbb{R}^n$ (其中 n 为输入空间的维数), 定义训练集

$$M = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$$

其中 $x_i \in X$; $y_i \in \{-1, 1\}$ 是 x_i 的标记, 若 x_i 属于正类, 则 $y_i = 1$, 若 x_i 属于负类, 则 $y_i = -1$; l 为样本的个数。SVM 即寻找能够将训练数据划分为两类的最优超平面^[10], 该超平面可以通过求下面的凸二次规划方程的解得到:

$$\begin{aligned} \max & \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=i}^l \alpha_i y_i \alpha_j y_j k(x_i, x_j) \\ \text{s. t. } & \sum_{i=1}^l y_i \alpha_i = 0; 0 \leq \alpha_i \leq c, i = 1, 2, \dots, l \end{aligned} \quad (1)$$

其中 $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, 为 Kernel 函数, 其满足 Mercer 条件, $\phi(x)$ 为原始输入空间到高维特征空间的非线性映射; α_i 为与每个样本对应的 Lagrange 乘子; $c > 0$, 是自定义的惩罚系数。给定一个测试实例 x , 它的类别由下面的决策函数决定:

收稿日期: 2008-03-04; 修回日期: 2010-04-04。

基金项目: 中央高校基本科研业务费专项资金资助项目(DUT10RW202)。

作者简介: 黄德根*(1965-), 男, 博士, 教授, 博士生导师, E-mail: huangdg@dlut.edu.cn。

$$f(\mathbf{x}) = \text{sgn} \left[\sum_{\mathbf{x}_i \in \mathbf{s}_v} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right] \quad (2)$$

其中 \mathbf{s}_v 为支持向量, b 是分类阈值, 可用任一支持向量或通过两类中任一对支持向量取中值求得.

针对机构名右边界识别任务来讲, 识别对象是存在于特征词表中的词, 对这些词提取出支持向量集. 每个向量均对应一个权值, 对其应用式(2)进行循环计算并求和, 得到的值即为 \mathbf{x} 到超平面的距离. 若该值大于 0, 表示该词识别为模型中定义的 +1 类, 即确定为右边界; 若小于 0, 表示该词识别为 -1 类, 即确定为非右边界.

1.2 条件随机场(CRF)

条件随机场是在给定输入节点值时计算指定输出节点值的条件概率的无向图模型^[11], 其中线性条件随机场(CRFs)是最简单的一种形式. 对于给定观测序列 $X = \{x_1, x_2, \dots, x_T\}$, CRFs 定义其对应的状态序列 $Y = \{y_1, y_2, \dots, y_T\}$ 的条件概率为

$$P(Y | X) = \frac{1}{Z(X)} \exp \left(\sum_{i=2}^T \sum_k \lambda_k f_k(y_{i-1}, y_i, x_i) + \sum_{i=1}^T \sum_k \lambda'_k f'_k(y_i, x_i) \right) \quad (3)$$

其中 $Z(X)$ 是归一化因子, 使得所有状态序列的概率和为 1, $f_k(y_{i-1}, y_i, x_i)$ 是关于观测序列和位置 i 及 $i-1$ 标记的转移特征函数, $f'_k(y_i, x_i)$ 是关于观测序列和位置 i 的标记的状态特征函数, λ_k 和 λ'_k 是与相应的特征函数相关的权值. 则最大可能的标记序列为

$$Y^* = \arg \max_Y \{P(Y | X)\} \quad (4)$$

对机构名识别的任务来说, 观测序列 X 为分词和词性标注后的序列, 对应的状态序列 Y 为标记集序列, 其中标记集的选择详见下一节. 例如在句子“呼市/jn/B 物资/n/I 集团/n/L 曾/d/O 有/vx/O 过/uo/O 辉煌/a/O 的/ud/O 历史/n/O”中, 对当前词“呼市”考虑其词形特征时定义特征函数如下:

$$f'_k(y_i, x_i) = \begin{cases} 1; & \text{位置 } i \text{ 的观测值 } x_i = \\ & "呼市" \text{ 并且 } y_i = "B" \\ 0; & \text{其他} \end{cases} \quad (5)$$

对“集团”考虑词性的组合特征时定义特征函数如下:

$$f_k(y_{i-1}, y_i, x_i) = \begin{cases} 1; & x_i = \{n, n\} \text{ 并且 } y_{i-1} = \\ & "I", y_i = "L" \\ 0; & \text{其他} \end{cases} \quad (6)$$

当特征函数取特定值时, 特征模板被实例化, 就可以得到具体的特征. 通过 CRF++(V0.49) 工具包的训练就可以得到特征函数对应的权值.

2 基于 SVM 和 CRF 的双层模型机构名识别

2.1 中文机构名识别所需要的资源

从训练语料中自动提取机构名识别所需的各词表, 详细介绍如下.

(1) 特征词表 D_f

特征词指的是机构名末尾具有一定表征意义的词, 如“厂、大学、公司”等. 对中文机构名的识别首先是从机构名右边界开始的, 所以建立该词表可作为机构名识别的触发条件.

(2) 前部词表 D_b

前部词是指机构名中除特征词之外的词, 地名名词和普通名词的比重较大, 但总体来说用词比较复杂, 有很强的随意性.

(3) 左右指界词表

左指界词即出现在机构名前面的第一个词, 比如“代表”“考入”; 右指界词即出现在机构名后面的第一个词; 比如“局长”“主办”. 不同指界词对机构名边界的指示作用不同, 因此在统计指界词表时, 需同时统计出各词作为指界词出现的次数, 并根据次数将其分为不同的级别.

(4) 简单机构名表

主要用于简单机构名的识别, 存在于该词表中的词均被认为是简单机构名候选词.

2.2 标记集

机构名的识别最终可以转化为序列标注的任务, 首先要定义适合该任务的标记集合, 不同的标记集对识别结果也有一定的影响^[12], 通过分析和实验定义标记集, 如表 1 所示.

表 1 标记集的选择

Tab. 1 Selection of tag set

模型	标记集	意义
第一层	S,O	S 代表简单机构名,O 代表非机构名成分
第二层	B,I,L,O BS,IS,LS	B,I,L 分别代表复杂机构名的开始、内部、末尾成分,O 代表非机构名成分 BS,IS,LS 分别代表简单机构名作为复杂机构名的开始、内部、末尾成分

2.3 基于 SVM 和 CRF 的双层模型进行中文机构名识别

该机构名识别模型分两层,第一层采用 CRF 识别简单机构名,并将识别结果传至第 2 层;第二层采用基于驱动式标注的方法,结合 SVM 和 CRF 进行复杂机构名的识别,即用 SVM 识别机构名右边界,对识别为右边界词向前采用 CRF 进行前部标注。然后将两层的识别结果进行合并。图 1 为机构名识别转换为序列标注的实例,图 2 为模型结构。

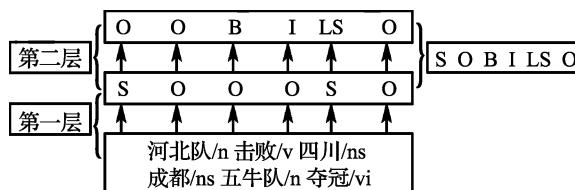


图 1 双层模型的识别过程

Fig. 1 Recognition process of cascaded model

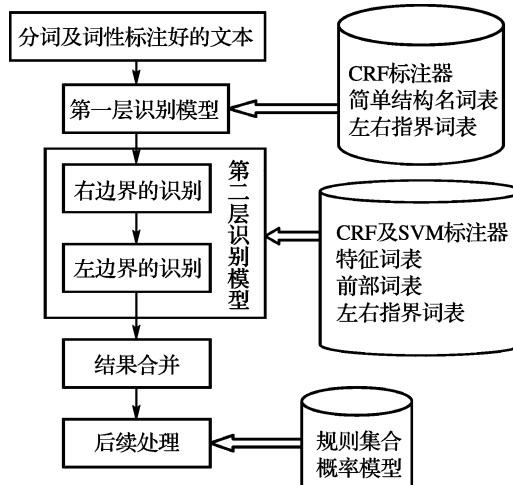


图 2 双层混合模型结构图

Fig. 2 Architecture of hybrid cascaded model

2.3.1 简单机构名识别 CRF 模型中非常重要的一步是针对特定的任务选择合适的特征集^[13]。原则上是选择的特征越多越好,但特征过多又会产生冗余信息,反而降低识别精度。通过对简单机构名分析确定的原子特征如表 2 所示。

其中 n 为表示位置的变量,取值为 $-2, -1, 0, 1, 2$ 。 $n = 0$ 表示当前位置, $n = -1$ 表示当前位置的前一位置, $n = 1$ 表示当前位置的后一位置,依此类推。

表 2 简单机构名识别的原子特征

Tab. 2 Atomic feature of simple organization name recognition

原子特征	意义
Word(n)	当前词的词形
Pos(n)	当前词的词性
L_spe(n)	若当前词前面第一个词在左指界词表中,则标为“Y”,否则标为“N”
R_spe(n)	若当前词后面第一个词在右指界词表中,则标为“Y”,否则标为“N”
Smp_org(n)	若当前词在简单机构名表中,则标为“Y”,否则标为“N”

为更好地利用复杂的上下文信息,构建组合特征为 $Word(n-1)Word(n)Pos(n-1)Pos(n)$ 、 $L_spe(n-1)Smp_org(n)Smp_org(n-1)R_spe(n)$, 其中 $n = -1, 0, 1$ 。

2.3.2 SVM 和 CRF 结合识别复杂机构名

(1) SVM 确定机构名右边界

右边界确定是个二值分类问题,而 SVM 是优秀的二值分类器,因此采用 SVM 进行右边界确定。对于出现在特征词表中的词均作为右边界候选词,利用 SVM 进行筛选,确定是否确实为机构名右边界词。SVM 也需要针对特定的任务选择合适的特征,考虑到效率和识别效果两方面的因素,选择词形和词性这两个特征。定义的 11 维向量的格式如下:

$$(S \quad W(-2) \quad P(-2) \quad W(-1) \quad P(-1))$$

$$W(0) \quad P(0) \quad W(1) \quad P(1) \quad W(2) \quad P(2))$$

其中 $S \in \{-1, +1\}$ 表示类别,在右边界识别的任务中, $S = -1$ 代表该词不是机构名右边界, $S = +1$ 代表该词是机构名右边界。 W 表示词形, P 表示词性,数字表示所考核的词相对当前词的位置,0 表示当前词,1 表示当前词右侧第一个词, -1 表示当前词左侧第一个词。例如,在句子“呼市/jn 物资/n 集团/n 曾/d 有/vx 过/uo 辉煌/a 的/ud 历史/n”中,对“集团”构建向量如下:

$$(+1 \text{ 呼市 } jn \text{ 物资 } n \text{ 集团 } n \text{ 曾 } d \text{ 有 } vx)$$

通过 SVM_light 的工具包对向量集进行训练,即可以得到各向量对应的 Lagrange 乘子。

(2) CRF 进行前部标注

右边界确定后,用 CRF 进行前部标注。以往的识别方法都是对文本进行全标注,本文考虑到

机构名的比重较小,使用全标注策略会造成大量的资源浪费,决定采用驱动式标注,即以右边界为驱动,只对候选词进行标注。候选词的确定规则如下:假设最长的机构名的长度为 N ,每确定一个右边界,则该词直接标注为“L”,该词前面的 $N-1$ 个词就成为机构名候选词,除非碰到标点符号(其中“、”、“、”、“、”等除外)、另一个右边界或者一行的开头。然后对确定为候选词的词进行标注,其他的词均直接标注为非机构名成分。这一策略的采用,在一定程度上缩短了训练和标注时间,提高了识别的效率,并且由于减少了冗余信息,识别精度也有所提高。

此处选用的原子特征除了第一层中采用的 $Word$ 、 Pos 、 L_spe 、 R_spe 外还需如下特征,如表 3 所示。

表 3 前部标注增加的原子特征

Tab. 3 Additional atomic feature of tagging foreside

原子特征	意义
$Former_word(n)$	若当前词在前部词表中,则标为“Y”,否则标为“N”
$Is_smp(n)$	若当前词识别为简单机构名,则标为“Y”,否则标为“N”
$Is_candidate(n)$	若当前词确定为机构名右边界则标为“L”,若为机构名候选词则标为“U”,否则标为“O”

表中 n 的取值为 $-2、-1、0、1、2$,所有的地名不管是否存在于前部词表中,均标为“Y”。组合特征定义为 $Word(n-1)Word(n)Pos(n-1)Pos(n)L_spe(n-1)Former_word(n)Is_candidate(n-1)R_spe(n)L_spe(n-1)Is_smp(n)Former_word(n)Is_smp(n-1)Is_candidate(n-1)R_spe(n)$,其中 $n = -1, 0, 1$ 。

该方法比较适合于完整机构名的识别,针对不同的语料需要在方法上作一些调整。若文本中不完整的机构名占有一定的比重,则采用两种方法进行识别,第一种采用本文的方法,第二种直接用 CRF 进行识别,然后比较两个识别结果,对不同的识别结果选择置信度较高的作为最终结果。

2.4 后续处理

后续处理包括两部分,第一部分为构建概率

模型,对识别结果中置信度低于某阈值的字串计算其可信度,并通过实验选择一个合适的阈值,可信度高于该阈值的确定为机构名,否则确定为非机构名。机构名的可信度 $T(org)$ 包括机构名特征词可信度 $T(S)$ 和机构名前部词可信度 $T(P)$,计算如下:

$$T(S) = \frac{\log(N_s + 2)}{\sum_{y \in D_f} \log(N_y + 2)}$$

$$T(P) = \frac{\log(N_p + 2)}{\sum_{y \in D_b} (N_y + 2)}$$

$$T(org) = \frac{C_n \times \sqrt[3]{n} \times (\sum_{i=1}^n T(F_i) + T(S))}{n + 1}$$

其中 N_s 为建立机构名特征词表时特征词 S 出现的次数; N_p 为建立机构名前部词表时前部词 P 出现的次数; C_n 为调整系数, n 为机构名前部词的个数。

第二部分为构建规则模型,主要用于识别不完整的机构名和兼类机构名,并修正一些明显的识别错误。规则举例如下。

(1) 并列关系词(如:和、与、及其、“、”;“等)前后的标注应保持一致,出现不一致的情况时将标注结果统一为置信度较高的一方。

(2) 从训练语料中提取机构名框架,比如:(考入、应聘到等)+机构名+(上学、读书、工作等),并根据出现次数进行精简,对置信度低于某阈值的识别结果进行匹配,能匹配上的确定为机构名,否则确定为非机构名。

(3) 体育新闻中经常出现和地名兼类的机构名,比如“中国对巴西”中的中国和巴西应标为机构名。首先提取一个体育新闻常用词表,比如半决赛、锦标赛等,当句子中出现“地名”对“地名”、(小胜、平等)+地名这一类的模式时,向前搜索,看前 N 个词中是否存在体育新闻常用词,若存在,则把该处的地名标为机构名。该规则正确修正了一些兼类词的识别错误,但同时也把一些地名错误地标成了机构名。

根据语料的不同,还有一些其他的规则,在此不再一一赘述。

3 实验分析

本文选取的语料是北大标注的《人民日报》2000 年 1~4 月和 9~10 月语料,所需资源是从 1~4 月及 9 月的语料中提取的,SVM 和 CRF 的训练语料是 1 月份的语料,约 9.51 MB,测试语料是 10 月份的语料,约 8.66 MB。

本文方法的实验结果如表 4 所示。

表 4 识别结果

Tab. 4 Recognition result %

实验	精确率	召回率	F 值
简单机构名识别	99.83	97.31	98.55
右边界识别	93.95	95.54	94.73
复杂机构名识别	94.36	94.56	94.46
整体识别	94.83	95.02	94.93

针对复杂机构名采用不同的方法进行实验,实验结果比较如表 5 所示。

表 5 复杂机构名的识别结果

Tab. 5 Results of complicated organization name recognition %

识别方法	精确率	召回率	F 值
CRF	93.59	93.84	93.72
全标注的 SVM+CRF	94.08	94.50	94.28
驱动式标注的 SVM+CRF	94.36	94.56	94.46

从实验结果可以看出,驱动式标注的 SVM+CRF 的识别效果最好,虽然相对于全标注的 SVM+CRF 在精度上的提高不太明显,但由于冗余信息的减少而使训练时间有所减少。

文献[6]采用基于层叠 CRF 的方法进行中文机构名识别,精确率和召回率分别为 88.12% 和 90.05%,本文的识别结果好于文献[6]的识别结果,但是由于本文的识别是基于正确的分词和词性标注之上的,而实际上分词的错误会降低识别精度。

文献[7]采用的训练语料和本文一样,测试语料是北大 1998 年的语料,该方法也是基于正确的分词和词性标注之上的,精确率和召回率分别为 94.20% 和 93.11%。

4 结语

本文建立了一个基于 SVM 和 CRF 的双层

模型进行机构名识别,根据简单机构名和复杂机构名的不同特点,在不同的层次中分别采用不同的方法进行识别。复杂机构名中经常包含有简单机构名,因此两层的识别不是孤立的。首先在第一层采用 CRF 进行简单机构名识别,并将结果传至下一层,在第二层采用驱动的 SVM 和 CRF 进行复杂机构名识别,然后将两层的识别结果进行合并,最后通过后续处理对置信度较低的结果进行修正。

实验表明该方法有较好的中文机构名识别效果,不足之处就是对复杂机构名的识别依赖于右边界的确定,因此无法识别不含特征词的机构名。此外,对于不完整的机构名、地名与机构名兼类的识别还有待进一步的深入研究。

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] ISOZAKI Hideki. Japanese named entity recognition based on a simple rule generator and decision tree learning [C] // Proceedings of the 39th Annual Meeting Association for Computational Linguistics. San Francisco: Morgan Kaufmann, 2001:314-321
- [3] ZHOU Guo-dong, SU Jian. Named entity recognition using an HMM-based Chunk Tagger [C] // Proceedings of the 40th Annual Meeting Association for Computational Linguistics. San Francisco: Morgan Kaufmann, 2002:473-480
- [4] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2):87-93
- [5] TAKEUCHI Koichi, COLLIER N. Use of support vector machines in extended named entity recognition [C] // Proceedings of the 6th Conference on Natural Language Learning. Morristown: Association for Computational Linguistics, 2002:167-170
- [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] ZHANG Su-xiang, ZHANG Su-xian, WANG Xiao-jie. Automatic recognition of Chinese organization name based on conditional random fields

- [C] // Natural Language Processing and Knowledge Engineering. Washington D C: IEEE Signal Processing Society, 2007:229-233
- [8] YU Hong-kui, ZHANG Hua-ping, LIU Qun. Recognition of Chinese organization name based on role tagging [C] // 20th International Conference on Computer Processing of Oriental Languages. Beijing: Tsinghua University Press, 2003:79-87
- [9] WU You-zheng, ZHAO Jun, XU Bo. Chinese named entity recognition combining statistical model with human knowledge [C] // Proceedings of the ACL Workshop on Multilingual and Mixed-language Named Entity Recognition. Morristown: Association for Computational Linguistics, 2003:65-72
- [10] 李丽双, 黄德根, 陈春荣, 等. 基于支持向量机的中文文本中地名识别[J]. 大连理工大学学报, 2007, 47(3):433-438
(LI Li-shuang, HUANG De-gen, CHEN Chun-rong, et al. Identification of location names from Chinese texts based on support vector machine [J]. Journal of Dalian University of Technology,
- 2007, 47(3):433-438)
- [11] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] // Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publisher Inc., 2001:282-289
- [12] ZHAO Hai, HUANG Chang-ning, LI Mu, et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling [C] // The 20th Pacific Asia Conference on Language, Information and Computation. Beijing: Tsinghua University Press, 2006:87-94
- [13] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons [C] // Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Morristown: Association for Computational Linguistics, 2003: 188-191

Chinese organization name recognition using cascaded model based on SVM and CRF

HUANG De-gen*, LI Ze-zhong, WAN Ru

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: A cascaded approach of Chinese organization name recognition based on support vector machine (SVM) and conditional random fields (CRF) is proposed. The simple organization name is recognized in the first level with CRF, and the result is used to support the decision of the second level. Then, a drive-based method is proposed in the second level for recognition of the complicated organization name combining SVM and CRF. Finally, the results of the two levels are combined, and a post-processing to correct those results with low confidence is adopted. The results show that this approach based on SVM and CRF is efficient in recognizing organization name through open test for large-scale real linguistics, and the recalling rate achieves 95.02% and the precision rate achieves 94.83%.

Key words: organization name recognition; conditional random fields (CRF); support vector machine (SVM); cascaded model