基于条件随机域的中文命名实体识别

史树敏[1,2,3]，王志强[1]，周浪[1]，冯冲[2]，黄河燕[2]

（1. 南京理工大学计算机科学与技术学院，南京 210094；2. 中国科学院计算语言信息工程研究中心，北京 10083；3. 内蒙古师范大学计算机与信息工程学院，呼和浩特 010022）

摘 要： 本文基于条件随机域模型处理中文命名实体中的人名、地名、组织机构名识别，综合利用外部特征，研究了字一级包括外国译名在内的中文人名、地名、简单组织机构名的识别，采用了利用互信息获取外部统计词典并建立外部特征的方法。初步实验结果表明，外部特征的加入可以弥补训练规模的不足、显著提高识别效果。

关键词： 中文命名实体识别；条件随机域；统计词典

Chinese Named Entity Recognition Using Conditional Random Fields Model

Shumin SHI[1,2,3], Zhiqiang WANG[1],Lang ZHOU[1],Chong FENG[2],Heyan HUANG[2]

(1.Department of Computer Science & Technology, NanJing University of Science & Technology, Nanjing 210094; 2.Research Center of Language & Information Engineering, CAS, Beijing 10083; 3.Computer & Information Engineering College, Inner Mongolia Normal University, Huhhot 010022)

Abstract: This paper studies the recognition of Chinese person name, including translated name, location name, simple organization name on character level using conditional random fields model (CRFs), which used to deal with Chinese named entity recognition, while utilizing external feature synthetically, adopting an approach of making use of mutual information to obtain external statistical lexicon and establishing external feature. Experiment shows that joining external feature can offset the deficiency of training models and improve the effect of recognition prominently.。

Keywords: Chinese Named Entity Recognition; Conditional Random Fields; Lexicon

1 引言

命名实体识别(Name Entity Recognition：NER)作为信息抽取的一项基础子任务，在信息检索、机器翻译、数据挖掘、自动文摘等领域也发挥着重要作用。术语“命名实体（Named Entity）”首先出现于 MUC 系列会议上，20 世纪九十年代初就已经引起人们的关注。在 MUC-6/7 以及 MET-1/2 会议上都有对命名实体识别的评测[1]。2000 年 12 月正式启动的 ACE 会议一开始就将实体侦测与识别(EDT)作为评测的两大任务之一[2]。国内 863 评测于 2003 年首次将中文命名实体识别作为分词标注的子任务引入，2004 年将其作为一个独立的评测项目。

本文利用现有标注语料库，以条件随机域模型为基础研究中文命名实体中人名、地名、组织机构识别，综合

作者简介：史树敏（1978－），女，内蒙古包头人，博士研究生，E-mail:nmssm@126.com

利用外部特征，实现了字一级包括外国译名在内的中文人名、地名、简单组织机构名的识别。同时采用了利用互信息（Mutual Information）从现有的标注语料库资源中获取外部统计词典，并在模型的训练过程中利用统计词典引入外部特征的方法。实验表明外部特征的加入可以弥补训练规模的不足、显著的提高实体识别效果。

2 条件随机域

条件随机域模型(Conditional Random Fields, CRFs)用特征函数的方式综合使用各种互相影响的语言特征，集合了最大熵模型和 HMM 模型的特点，[3]回避了传统 HMM 方法处理长距离关联的不足和 MEMM 等模型中的标注偏置问题。目前 CRFs 在英文 POS 标注[4]，英文名词短语识别[5]，浅层分析[4]，语义角色标注[6]等领域取得了一定的成功。

CRFs 是一种无向图模型，对于指定的节点输入值，能够计算指定的节点输出值上的条件概率，其训练目标是使得条件概率最大化[5]。线性链是 CRFs 中常见的特定图结构之一，它是由指定的输出节点顺序链接而成。

定义 $X=x_1 \ldots x_t$ 为给定的输入观测序列，即无向图模型中 t 个输入节点上的值（如一个中文词序列）；定义 $Y=y_1 \ldots y_t$ 为一个长度与 x 相等的状态序列，即无向图中 t 个输出节点上的值。一个带参 $\Lambda=\{\lambda_1, \cdots, \lambda_k\}$ 的线性链 CRF 把给定输入序列 x 得到的状态序列 y 的条件概率定义为：$P_\Lambda(y \mid x)=\frac{1}{Z_\Lambda(x)} \exp(\sum_{i=1}^{T} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, t))$，式中 $Z_\Lambda(x)$ 是一个范化因子，使得在给定输入上的所有可能的状态序列的概率之和为 1；$f_k(y_{i-1}, y_i, x, t)$ 表示一个特征函数，通常取布尔值，λ_k 是训练中得到的、与每个特征 f_k 相关的权重参数，它的取值反映了特征函数所代表的事件发生的可能性。

3 基于 CRF 的中文命名实体识别

针对人名、地名和组织机构名的不同特点，我们制定了不同的建模方法，主要体现在模型所考察的粒度大小上。对于这两种实体的建模是建立在字一级（character level）。设计是出于这样得考虑：第一，中文人名、地名的语言学特点决定了可以在字一级的基础上进行建模和识别，并显示出了一定的潜力[7]。第二，以往大多数利用统计方法进行的识别都是在词一级进行的，对于在字一级识别的研究尚少，值得探讨。第三，利用具有歧义的分词结果作为实体识别的基础会产生一定的干扰作用。

命名实体识别任务实际上就是序列标注任务，因此首先要确立序列标注集 L。在实际的操作中我们采取了 BIO 的标注方式，即把一个输入单元标注为 B(begin：开始)、I(internal 内部)和 O(other 其他)之一。对于字一级进行的人名地名、简单机构名识别任务定义了七种标记的集合，L={NB,NI, SB,SI,OB,OI,O}。其中，各个标记分别代表：人名开始 NB，人名内部 NI，地名开始 SB，地名内部 SI，简单机构名开始 OB，简单机构名内部 OI，其他 O。

4 特征函数

在定义特征函数的时候，我们首先构建观察值上的真实特征 b(x,i)的集合，这个特征集表现了训练数据的经验分布特性，同时也反映模型分布。在当前状态（对应于状态函数情况）或是前一状态与当前状态（对应于转移函数情况）有特定取值时，每个特征函数取值为一个观察特征 b(x,i)。例如：

$$f(y_{i-1}, y_i, x, i)=\begin{cases} b(x,i) & if\ y_{i-1}=\mathrm{NB}\ y_i=\mathrm{NI} \\ 0 & otherwise \end{cases}$$

，其中 b(x,i)表示某种真实观察值，出现时值为 1，其他情况为 0。对于 i 时刻的观察值，在实际应用中定义窗口大小为 2。实验中对纳入特征集（feature set）中的特征函数类型进行了细分，分为上下文特征（Content Feature）、词表特征（Lexicon Feature）、词性特征。我们将特征函数用

一种比较直观的形式表示：{yt-1=Label1,yt=Label2,xw=Content}，其中 yt-1 表示前一个标记，yt 表示当前标记，对于人名、地名识别任务来说 Label1,Label2∈{NB,NI,SB,SI,OB,OI,O}，对于组织机构名任务来说，Label1,Label2∈{ORGB,ORGI,O}，w 为考察的位置，w∈{-2,-1,0,1,2}，Content 表示 w 位置内容。对于状态特征函数($s_k(y_i,x,i)$)由于没有考虑前一位置的标注，于是在 yt-1 中用一个记号#代替标注以区别于转移特征函数($t_j(y_{i-1},y_i,x,i)$)。

上下文特征考察的观察值来自于输入观察序列本身（字面值），因此也可称为字面特征。字面特征比较直观，也涵盖了很多的语言信息。通过后面大量的实验表明，如果训练的规模比较大，只使用上下文特征依然可以获得比较好的效果。对于词性特征函数来说，它考察的观察值来自于输入观察序列的词性信息，主要应用于对于复杂组织机构名的识别模型。实验中复杂组织机构名的识别基于词一级，建立在分词和词性标注基础之上。限于篇幅，另文叙述。

CRF 模型除了能够综合利用包括字、词、词性在内的上下文信息，还能利用外部特征(External Feature)。这个特点使我们在很多方面有了尝试的空间，本文研究的一个重点就是将词表纳入外部特征。首先，在基于互信息的词典获取方面，从标注语料库获取了部分词表，配合网上获得的部分词表用于外部特征生成，重点针对人名识别任务。目前提取的外部特征列表是单字词二字词。其次，对于词典外部特征获取，考察的观察值来自于输入观察序列的上下文信息对应于词表中的位置。首先生成的词表序号，表示为 $L_0,L_1,..L_7$。分别表示中国人名常用字表、译名常用字表等。在词表中出现与否、单字还是 2-Gram 情况分别做标记，观察值查询出现在词表中的情况获得相应的观察特征，然后获得外部特征。

以上考虑的特征都是原子特征，不足以表示上下文出现的各种现象。通过模板的复合，可将更多的因素纳入考虑范畴。如考虑人名＋动词，介词/动词+地名等情况。原子特征的组合扩展了模型的特征空间，通过实验比较最终采取定义复合模板的方式生成复合特征。特征模板的作用就是为特征函数的生成提供一个统一的模式，实验中，模板分为原子特征模板和复合特征模板。本文使用基于设定特征频率阈值 K 的方法从候选特征集中选择出现了 K 次以上的特征作为模型特征。对于上下文特征、外部词表特征、词性特征分别制定原子特征模板。在人名地名识别任务中的复合特征模板：x[-1,0]/ x[0,0]、x[0,0]/ x[1,0]、x[1,0]/ x[2,0]、x[-1,1]/ x[0,1]、x[0,1]/ x[1,1]、x[1,1]/ x[2,1], x[0,1]/ x[2,2], x[-1,1]/ x[0,2] x[-2,2]/ x[0,1]。

5 实验

实验语料主要面向新闻领域，由北大富士通人民日报 98 年 1 月的语料和兰开斯特汉语语料库构成，网上获得译名常用字词典。从人民日报语料库中选取不同规模训练语料分别为 5000 句(Train1)，10000 句(Train2)，20000 句(Train3)，40000 句(Train4)。为了考察词表外部特征对于识别的作用，封闭测试语料规模固定，从训练使用语料中随机抽取；开放测试语料部分为人民日报语料，部分为兰开斯特语料。

表 1　不同规模训练集得到的对比结果

人名识别		P(%)	R(%)	F(%)
Train1	封闭测试	95.3	94.8	95.1
	开放测试 1	87.4	67.9	69.8
Train2	封闭测试	94.9	94.0	94.5
	开放测试 1	87.9	72.2	79.3
Train3	封闭测试	95.3	95.2	95.3
	开放测试 1	88.3	77.9	82.8
Train4	封闭测试	95.1	95.2	95.2
	开放测试 1	90.4	85.3	87.8

表 2 外部词表特征对比实验

训练规模	类型	P (%)		R (%)		F (%)	
		不带	带	不带	带	不带	带
10000	开放测试 1	89.1	87.2	64.1	72.2	74.6	79.3
	开放测试 2	72.4	71.1	55.6	64.9	62.8	67.8
40000	开放测试 1	91.5	90.4	83.2	87.7	87.2	89.1
	开放测试 2	82.6	81.9	71.5	77.6	76.6	79.7

四个不同规模的训练集通过预处理、训练模块后生成 model1，model2，model3， model4，包含词表外部特征。其中训练过程中我们设定用于特征选择的阈值为 4，σ 的大小为 2.0。分别对这四个模板文件借助于开源资源解码模块和测评模块得到实验结果如表 1 所示。为了考察词表外部特征对于识别的作用，针对人名识别进行带词表和不带词表两种方式进行对比实验，对比实验结果见表 2。分别选择训练规模为 10000 句和 40000 句。本文只考察开放测试结果。

随着实验规模的增大，人名识别的性能指标，特别是召回率随训练语料规模的增大提高的速度较快。通过是否带词表外部特征的对比实验可看出，外部特征的引入能够在几乎不影响准确率的情况下，显著得提高识别的召回率；并在一定程度上克服训练数据规模上的不足。

6 结束语

本文研究以条件随机域模型为基础，进行字一级人名、地名、简单组织机构名的识别，利用条件随机域模型的良好特性，通过互信息统计的方法从现有的语料库中获得大部分外部词表，借助于外部词表的使用，同时在模型训练过程中利用统计词典引入外部特征，目前外部特征主要针对人名、地名识别任务。在下一步的工作中我们将在组织机构名识别任务中尝试使用外部特征改善识别效果。

参考文献：

[1] Nancy A Chinchor, Overview of MUC-7/MET-2. In Proceedings of seventh Message Understanding Conference, 1998.

[2] http://www.nist.gov/speech/tests/ace/

[3] Viola P, Mukund Narasimhand. Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar[C]. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2005:330~337

[4] Sha F and Pereira F. Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol.1. Morristown(USA):ACL, 2003:134~141.

[5] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of 18th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2001:282–289

[6] T. Cohn, P. Blunsom. Semantic Role Labeling with Tree Conditional Random Fields. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005:169–172

[7] Hanna M. Wallach. Conditional Random Fields: An Introduction. University of Pennsylvania CIS Technical Report MS-CIS-04-21, 2004.