基于层叠隐马尔可夫模型的中文命名实体识别

俞鸿魁[1,2]，张华平[1]，刘群[1]，吕学强[2]，施水才[2]
(1.中国科学院 计算技术研究所，北京 100080; 2. 北京信息科技大学 中文信息处理研究中心，北京 100101)

摘 要：提出了一种基于层叠隐马尔可夫模型的中文命名实体一体化识别方法，旨在将人名识别、地名识别以及机构名识别等命名实体识别融合到一个相对统一的理论模型中。首先在词语粗切分的结果集上采用底层隐马尔可夫模型识别出普通无嵌套的人名、地名和机构名等，然后依次采取高层隐马尔可夫模型识别出嵌套了人名、地名的复杂地名和机构名。在对大规模真实语料库的封闭测试中，人名、地名和机构识别的 F-1 值分别达到 92.55%、94.53%、86.51%。采用该方法的系统 ICTCLAS 在 2003 年 5 月 SIGHAN 举办的第一届汉语分词大赛中名列前茅。

关键词：命名实体识别；角色标注；ICTCLAS
中图分类号：TP391.2 **文献标识码**：A **文章编号**：1000-436X(2006)02-0087-08

Chinese named entity identification using cascaded hidden Markov model

YU Hong-kui[1,2], ZHANG Hua-ping[1], LIU Qun[1], LV Xue-qiang[2], SHI Shui-cai[2]
(1. Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China;
2. Chinese Information Processing and Research Center, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: An approach for Chinese named entity identification using cascaded hidden Markov model, which aimed to incorporate person name, location name, organization name recognition into an integrated theoretical frame was presented. Simple named entity was recognized by lower HMM model after rough segmentation and complex named entity such as person name, location name and organization name was recognized by higher HMM model using role tagging. In the test on large realistic corpus, its F-1 measure of person name, location name and organization name was 92.55%, 94.53% and 86.51%. In the first international word segmentation bakeoff held by SIGHAN (the ACL Special Interest Group on Chinese Language Processing) in 2003. ICTCLAS, which name entity identification base on this model achieved excellent score.

Key words: named entity identification；role tagging；ICTCLAS

1 引言

命名实体识别(named entity identification)的研究是自然语言处理中的一项基本工作，不仅是分词和标注过程中的一个重要环节，而且在句法分析、机器翻译、信息检索、提取以及自动问答系统等领域中也有直接的应用。由于中文文本中词与词之间没有分隔符，中文文本的分词和中文命名实体的识

收稿日期：2005-11-15；**修回日期**：2005-12-20
基金项目：国家重点基础研究发展计划(“973”计划)基金资助项目(G1998030507-4,G1998030510)；计算所领域前沿青年基金资助项目（20026180-23）；国家自然科学基金资助项目（60272084）；北京市教育委员会科技发展计划重点项目（KZ200310772013）
Foundation Items: The National Basic Research Program of China(973 Program)(G1998030507-4,G1998030510)；The ICT Youth Fund（20026180-23）；The National Natural Science Foundation of China(60272084);The Scientific Research Key Program of Beijing Municipal Commission of Education (KZ200310772013)

别是互相缠绕、密不可分的。大部分的命名实体是未登录词，如果在中文分词过程中不考虑这些未登录词，不实施命名实体的识别，必然影响中文分词的质量。

一般来说，命名实体识别的任务就是对于一篇待处理文本，识别出其中出现的人名（person）、地名（location）、机构名（organization）、日期（data）、时间(time)、百分数(percentage)、货币(monetary value)这七类命名实体。其中人名、地名、机构名的识别是最难、也最重要的三类，本文将主要讲述这三类命名实体的一体化识别。

1.1 中文命名实体的特点

数量众多是各类命名实体的共同特点。根据对人民日报 1998 年 1 月的语料库(共计 2 305 896 字)进行的统计，共有人名 19 965 个，而这些人名大部分属于未登录词。

构成规律复杂是命名实体的另一共同特点。由于人名的构成规则各异，中文人名识别又可以细分为中国人名识别、日本人名识别和音译人名识别等。机构名的组成方式最为复杂，机构名的分类种类繁多，各有其独特的命名方式，用词相当广泛，只有结尾用词相对集中。

此外，一个命名实体经常和一些词组合成一个嵌套的命名实体，人名中嵌套着地名，地名中也经常嵌套着人名，嵌套的现象在机构名中最为明显，机构名不仅嵌套了大量的地名，而且还嵌套了相当数量的机构名。互相嵌套的现象大大制约了复杂命名实体的识别，也注定了各类命名实体的识别并不是孤立的，而是互相交织在一起的。

与其它类型的命名实体相比，长度和边界难以确定使得机构名更难识别。中国人名一般二至三个字，最多不过四个字，常用地名也多为二至四个字，但是机构名长度变化范围极大，少到只有两个字的简称，多到长达几十个字的全称，在人民日报的真实文本中，由十个以上的词构成的机构名占了相当一部分的比例。

而且中文不像英文那样在命名实体中有形态的变化。中文命名实体的识别困难重重，归根到底还是由机构名的自身特点造成的。

1.2 已有的工作

目前汉语命名实体识别的研究有很多[1~7]，从方法上来说，主要是规则和统计这两种方法。基于规则的方法一般采用特征字（词）触发的方式来进行命名实体识别，比如用中国人名的姓氏用字来触发中国人名的识别，或者利用机构名的结尾关键词相对集中的特点来触发机构名的识别。基于统计的方法主要是通过对大规模语料库内的命名实体以及上下文进行统计分析，构建统计模型来进行命名实体的识别，解决方案有隐马尔可夫模型、最大熵模型、基于 Agent 的方法和基于类的三元语言模型等。

虽然目前汉语命名实体识别的研究有很多，如人名识别、地名识别、译名识别以及机构名识别等，但很多是专门针对于某一类命名实体的识别。从上节所述中文命名实体的互相嵌套的特点可知，中文命名实体识别的不应是孤立的。如何在一个集成的框架下进行各类命名实体的识别并达到一个整体的最优效果，将是整个命名实体识别过程的关键所在。

本文提出了一种基于层叠隐马尔可夫模型（cascaded HMM, cascaded hidden markov model）的方法，旨在将人名识别、地名识别以及机构名识别等命名实体识别融合到一个相对统一的理论模型中。首先在词语粗切分的结果集上，采用底层隐马尔可夫模型识别出普通无嵌套的人名、地名和机构名等，然后依次采取高层隐马尔可夫模型识别出嵌套了人名、地名的复杂地名和机构名。中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS[8](institute of computing technology Chinese lexical analysis system，该系统全部的源码和文档均可在中文自然语言处理开放平台 www.nlp.org.cn 中自由下载，免费使用) 采用的就是基于层叠隐马尔可夫模型的命名实体识别，该系统在 2003 年 5 月 SIGHAN 举办的第一届汉语分词大赛中名列前茅。

2 基于层叠隐马模型的命名实体识别方法

隐马尔可夫模型是一种在自然语言处理领域中被广泛应用的统计模型。中文命名实体识别中的人名识别、地名识别、译名识别以及机构名识别等都可以用隐马尔可夫模型来解决。本文提出的层叠隐马尔可夫模型（cascaded HMM, cascaded hidden markov model）就是试图在统一的隐马尔可夫模型中识别各类命名实体，并在这些隐马尔可夫模型中建立起一定的联系，以形成一个一体化的命名实体识别系统。

整个命名实体识别的层叠隐马尔可夫模型由三级互相联系的隐马尔可夫模型构成，自底向上分

别为人名识别 HMM、地名识别 HMM 和机构名识别 HMM，每一级都是以隐马尔可夫模型作为基本的算法模型，整个算法的时间复杂度和隐马尔可夫模型的时间复杂度相同，分析时间随着输入串长度的增长而线性增长，速度非常快。各层隐马尔可夫模型之间以如下两种方式互相关联，形成一种紧密的耦合关系：

(1) 每一层隐马尔可夫模型都采用 N-Best 策略，将产生的最好的若干个结果送到词图中供高层模型使用；

(2) 低层的隐马尔可夫模型通过词语的生成模型为高层隐马尔可夫模型的参数估计提供支持。

其中第一层人名识别的输入为粗切分的分词序列，每一层隐马尔可夫模型都采用改进的 Viterbi 算法（N-Best）,输出最好的若干个结果作为高一级隐马尔可夫模型的输入。最高级隐马模型将在人名识别和地名识别的基础之上进行机构名识别。

2.1　基于角色标注的命名实体识别

我们采用一种统一的策略——基于角色标注的隐马尔可夫模型对各类命名实体进行识别。角色标注的基本思想是，根据各类命名实体的构成和用词特点各自制定一套角色标记集，采取 Viterbi 算法对切分结果进行角色标注（类似于一个简单的词性标注过程），在角色序列的基础上进行简单的模式识别，最终实现中文命名实体的自动识别。识别过程中，只需要某个词作为特定角色的概率以及角色之间的转移概率。该方法的实用性还在于，这些角色信息只要对真实语料库稍加改动，就可以得到。

各类命名实体角色标记集的选取不仅需要根据其自身特点，结合专家知识科学地设定，而且还要通过不断地实验，对角色集进行调整。一般来讲，进行人名识别都会想到利用人名的姓氏、姓名常用字以及上下文信息等，但是姓氏、姓名常用字和上下文等常常能组合成词，根据对《人民日报》1998 年一月的语料库进行统计，在 15 890 个中国人名中，共出现了 1 131 次姓与名或名与名成词的情况，而姓名与上下文成词的情况很少，仅为 23 次。为了防止设计与实际情况相脱节，我们有针对性地添加了表 1 中的 I、J、K 三种角色。表 1～表 3 列举了命名实体识别过程中用到的部分角色，完整的角色表、角色的自动标注过程以及角色集是如何筛选的，文献[8～11]中有详细的描述，这里就不再重复。

表 1　人名识别角色简表

角色	意义	示例
A	人名的上文	又/来到/于/洪/洋/的/家
B	人名的下文	新华社/记者/黄/文/摄
C	中国人名的姓	张/华/平/先生；欧阳/修
D	双名的首字	张/华/平/先生
E	双名的末字	张/华/平/先生
F	单名	张/浩
I	姓与双名的首字成词	记者/盛世/良
J	姓与单名成词	著名/学者/何方/同志
K	双名本身成词	新华社/记者/兰/红光/摄
X	连接词	邵/钧/林/和/稽/道/青/说
Z	其它非人名成分	

表 2　地名识别角色简表

角色	意义	示例
A	地名的上文	我/来到/中/关/园
B	地名的下文	刘家村/和/下岸村/相邻
C	中国地名的首部	石/河/子/乡/
D	中国地名的中部	石/河/子/乡/
F	中国地名的末部	石/河/子/乡/
G	中国地名的后缀	海/淀区
X	连接词	刘家村/和/下岸村/相邻
Z	其它非地名成分	

表 3　机构名识别角色表

角色	意义	例子
A	上文	参与/亚太经合组织/的/活动
B	下文	中央/电视台/报道
X	连接词	北京/电视台/和/天津/电视台
C	特征词的一般性前缀	北京/电影/学院
G	特征词的地名性前缀	交通/银行/北京/分行
H	特征词的机构名前缀	中共中央/顾问/委员会
I	特征词的特殊性前缀	中央/电视台
D	机构名的特征词	国务院/侨务/办公室
Z	其它非机构名成份	

2.2　命名实体的概率估计：基于角色的词语生成模型

在复合命名实体识别过程中，如果用到了某个简单命名实体 *wi* 作为其组成部分，相应的角色标记是 *ti*，就会出现无法从语料库中统计得到输出概率 $p(wi|ti)$的情况，这是因为 *wi* 本身是个未登录词，从来没有在词典和语料库中出现过。为此我们需要引入一个新的模型来估计这个输出概率，我们称这

个模型为基于角色的词语生成模型。基于角色的词语生成模型和基于角色的隐马尔可夫模型是一一对应的。对于每一个角色隐马尔可夫模型而言，都需要一个相应的角色生成模型，用于计算其所识别出的命名实体的输出概率，而且使两层 HMM 之内形成一种联系。

假设识别出来的未登录词为 w，类别为 c，利用隐马过程可以得到

$$p(w|c)=\prod_{j=0}^{k}p(w_{p+j}|r_{p+j})p(r_{p+j}|r_{p+j-1}) \quad (1)$$

k是多大合适呢？

其中 w_i 由第 $p,p+1,\cdots p+k-1$ 个初始切分单元组成。

我们可以看到，这个模型中总共有两个参数：$p(w_{p+j}|r_{p+j})$ 和 $p(r_{p+j}|r_{p+j-1})$，都可从语料库中统计得到。

2.3　角色信息的自动抽取

$p(w_i|t_i)$ 和 $p(t_i|t_{i-1})$是两个关键的角色信息参数。其中 $p(w_i|t_i)$指的是角色为 t_i 的 Token 集合中 w_i 的概率；$p(t_i|t_{i-1})$表示的是角色 t_{i-1} 到角色 t_i 的转移概率。在大规模语料库训练的前提下，根据大数定理，可以得到

$$p(w_i|t_i)\approx C(w_i,t_i)/C(t_i) \quad (2)$$

其中，$C(w_i,t_i)$：w_i 作为角色 t_i 出现的次数；$C(t_i)$：角色 t_i 出现的次数。

$$p(t_i|t_{i-1})\approx C(t_{i-1},t_i)/C(t_{i-1}) \quad (3)$$

其中，$C(t_{i-1},t_i)$：角色 t_{i-1} 下一个角色是 t_i 的次数。

$C(w_i,t_i)$、$C(t_i)$、$C(t_{i-1},t_i)$均可通过对已经切分标注好的熟语料库进行学习训练、自动抽取得到。

由于命名实体的角色标注是在分词粗分结果的基础上进行的，为了得到最真实的角色信息，我们并没有用语料库作为惟一的训练语料，所有的角色训练是在分词系统实际的分词结果上进行的。

如图 1 所示，整个命名实体识别系统的训练过程共分四步：

第一步，根据《人民日报》语料库，制作分词词典，统计各分词的出现频率。核心分词词典中，除简单地名、机构名和高频的人名外，不包括任何的命名实体。

第二步，利用词法分析系统，使用第一步生成的核心分词词典，对语料库的原始文本进行词语的粗切分，得到无任何命名实体识别的切分结果。与语料库中标注好的人名进行比对得到人名的角色语料库。例如，如果粗切分的结果为“老挝/ 国会/ 主席/ 会见/ 何/ 鲁/ 丽/”，语料库为“[老挝/ns 国会/n]nt 主席/n 会见/v 何/nr 鲁丽/nr”，则相应的人名角色语料为“老挝/Z 国会/Z 主席/Z 会见/A 何/C 鲁/D 丽/E”。再对角色语料进行训练，最终得到人名的角色词典和各个角色之间的角色转移概率。

第三步，词法分析系统对语料库的原始文本重新进行切分，利用人名角色词典在粗切分的基础之上进行人名识别，得到经过人名识别的切分结果。与语料库中标注好的地名进行比对，得到地名的角色语料库。地名中经常嵌套有人名，为了识别含有人名的复合地名，人名识别 HMM 会把识别出来的人名作为一个人名类输送到地名识别 HMM，人名类的输出概率可由基于角色的词语生成模型得出，为了统计人名类的转移概率，我们将角色语料库中所有的人名转换为<PER>。例如，如果切分结果为“周/ 恩来/ 纪念馆/ 门前/ 排/ 起/ 了/ 长/ 队/”，语料库为“[周/nr 恩来/nr 纪念馆/n]ns 门前/s 排/v 起/v 了/u 长/a 队/n”，则相应的地名角色语料为“<PER>/C 纪念馆/F 门前/B 排/Z 起/Z 了/Z 长/Z 队/Z”。

第四步，与前两步相似，对语料库的原始文本再切分后得到经过人名和地名识别的切分结果，在

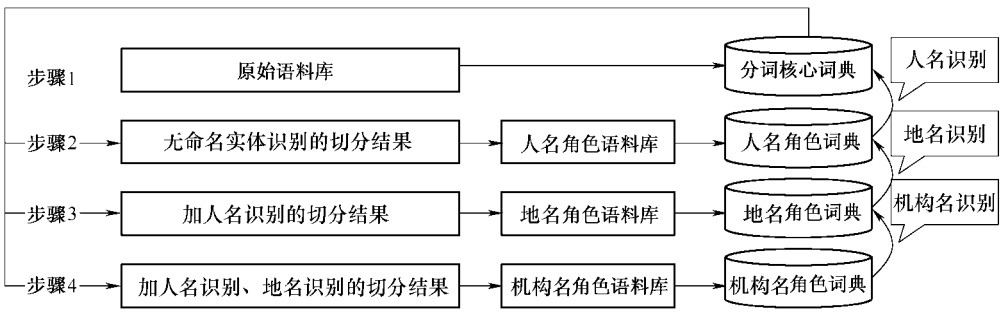

图 1　角色词典训练流程

与语料库中标注好的机构名进行比对，基于与第三步同样的理由，我们将所有的人名转换为<PER>，机构名转换为<LOC>，得到机构名的角色语料库。

最终，我们得到命名实体识别所需的三个角色词典和角色转移概率。

从某种程度上讲，角色训练过程是分词系统的自适应过程，通过对未进行命名实体识别的切分结果进行角色学习，从切分结果中统计规律，从而不断地提高自身性能。多层次的自适应性训练也是ICTCLAS 的显著特点之一。

2.4　自动识别的最终实现

识别的过程就是在已经角色标注好的序列上进行的。以最为复杂的机构名自动识别为例，识别的策略就是找出满足“[CFGHIJ]D”的子串。

角色标注好的文本一般如下：

“在/Z 1998 年/Z 来临/Z 之际/Z ，/Z 我/Z 十分/Z 高兴/Z 地/Z 通过/A 中央/I 人民/I 广播/C 电台/D、/X 中国/G 国际/I 广播/C 电台/D 和/X 中央/I 电视台/D，/B 向/Z 全国/Z 各族/Z 人民/Z ，/Z 向/Z 香港/Z 特别/Z 行政区/Z 同胞/Z、/Z 澳门/Z 和/Z 台湾/Z 同胞/Z 、/Z 海外/Z 侨胞/Z，/Z 向/Z 世界/Z 各国/Z 的/Z 朋友/Z 们/Z，/Z 致以/Z 诚挚/Z 的/Z 问候/Z 和/Z 良好/Z 的/Z 祝愿/Z ！/Z”。

应用上述的策略，识别出的潜在机构名为“中央人民广播电台”、“中国国际广播电台”以及“中央电视台”。

3　实验结果与分析

下面给出ICTCLAS在不同条件下的测试结果，并介绍第一届国际分词大赛中的比赛情况。

3.1　对《人民日报》语料库的命名实体识别实验

在这里，按照惯例引入如下评测指标：命名实体识别的准确率 P 和召回率 R，以及 F 值。它们的定义分别如下：

P=正确识别该类命名实体数/识别出该类命名实体总数×100%

R=正确识别该类命名实体数/该类命名实体总数×100%

$F=\dfrac{R\times P\times(1+\beta^2)}{R+P\times\beta^2}$，这里取 β=1，称为 $F-1$ 值。

这里我们训练和测试使用的是在北京大学计算语言学研究所加工的《人民日报》语料库的基础上进行了改造的语料库，分词标准完全一致，只是在词性标注集上进行了细分。该套词性标注集也把命名实体进一步分类，不仅把人名细分为中国人名、音译人名和日本人名，还将中国人名的姓和名分别标注；地名划分为中国地名和音译地名两大类。

以《人民日报》1998 年一月至五月总共五个月的新闻语料库为训练语料库，《人民日报》1998 年一月语料为封闭测试语料，我们进行了以下四种条件下的性能测试：

(1) BASE: 基准测试，即仅仅做隐马分词和词性标注，不引入其他层面的 HMM。

(2) +PER: 在 BASE 的基础上引入人名识别 HMM。

(3) +LOC: 在+PER 的基础上引入地名识别 HMM。

(4) +ORG: 在+LOC 的基础上引入机构名识别 HMM。

表 4～表 7 给出了 4 种条件下，词法人名识别、地名识别以及机构名识别的准确率 P、召回率 R 和 $F-1$ 值。

表 4　BASE 基准测试命名实体识别性能指标

BASE	P	R	$F-1$
中国人名	21.70	18.41	19.92
音译人名	85.34	14.95	25.44
日本人名	37.88	26.88	31.45
Σ人名	24.36	17.92	20.65
中国地名	74.18	90.92	81.70
音译地名	85.57	94.65	89.88
Σ地名	76.00	91.57	83.07
机构名	87.72	37.44	52.48

从表 4 中可以看到：除了地名识别因为分词词典中已含有大量常用地名的缘故，其它各类命实体识别的指标都很低。

表 5　　+PER 人名识别性能指标

+PER	P	R	$F-1$
中国人名	90.34	94.63	92.43
音译人名	85.33	92.60	88.82
日本人名	84.24	91.94	87.92
Σ人名	89.43	94.26	91.78

随着引入人名识别，分词系统的切分正确率也由 BASE 基准测试时的 96.47%上升为 97.88%。

登录词，命名实体占相当大的比例，OOV rate 指测试集中未登录词所占比率，OOV Recall 指未登录词的识别召回率；IV（in vocabulary）是指训练语料中已出现的词，IV Recall 指已登录词的识别召回率。

ICTCLAS 分别参加了简体的所有四项任务（CTBc、CTBo、PKc、PKo）和繁体的受限训练任务。其中在宾州树库受限训练任务（CTBc）中综合得分 0.881，名列第一；北京大学受限训练任务（PKc）中综合得分 0.951，名列第一；北京大学非受限训练任务（PKo）中综合得分 0.953，名列第二。在未登录词识别方面，宾州树库受限训练任务（CTBc）中召回率 0.705，名列第一；北京大学受限训练任务（PKc）中召回率 0.724，名列第二；北京大学非受限训练任务（PKo）中召回率 0.743，名列第二。

4　结论

本文系统地总结了中文命名实体的特点，分析了中文命名实体识别上的诸多难点，提出了一种基于层叠隐马尔可夫模型的中文命名实体一体化识别方法。并利用各类命名实体构成角色表及其相关统计信息，对句子中的不同成分进行角色标注，在角色序列的基础上进行字符串匹配，从而识别出中文命名实体。通过对大规模真实语料库的测试，该方法取得了相当好的效果，该方法是行之有效的。最后感谢为我们提供训练和测试语料的北京大学和富士通公司，感谢中国科学院计算技术所 NLP 小组的所有成员。

参考文献：

[1] 季姮，罗振声. 基于反比概率模型和规则的中文姓名自动辨识系统[A]. 自然语言理解与机器翻译[C]. 北京：清华大学出版社，2001. 123-128.
JI H, LUO Z S. Name frequency model and rules based on Chinese name identifying[A]. Natural Language Understanding and Machine Translation[C]. Beijing: Tsinghua Univ Press, 2001. 123-128.

[2] 何燕. 基于单字词转移概率的未登录词识别[A]. 自然语言理解与机器翻译[C]. 北京：清华大学出版社，2001. 141-146.
HE Y. Identification of unlisted words on transitive probability of monosyllabic words[A]. Natural Language Understanding and Machine Translation[C]. Beijing: Tsinghua Univ Press, 2001. 141-146.

[3] 吕雅娟, 赵铁军, 杨沐昀等. 基于分解与动态规划策略的汉语未登录词识别[J].中文信息学报，2001,15(1):28-33.
LV Y J, ZHAO T J, YANG M Y, *et al*. Leveled unknown Chinese words resolution by dynamic programming[J]. Journal of Chinese Information Processing, 2001,15(1):28-33.

[4] 王宁,葛瑞芳, 苑春法等. 中文金融新闻中公司名的识别[J]. 中文信息学报，2002, 16(2):1-6.
WANG N, GE R F, YUAN C F, *et al*. Company name identification in Chinese financial domain[J]. Journal of Chinese Information Processing, 2002,16(2):1-6.

[5] 张艳丽，黄德根等. 统计和规则相结合的中文机构名称识别[A]. 自然语言理解与机器翻译[C]. 北京：清华大学出版社，2001. 233-239.
ZHANG Y L, HUANG D G, *et al*. Identification of Chinese organization names based on statistics and rules[A]. Natural Language Understanding and Machine Translation[C]. Beijing: Tsinghua Univ Press, 2001. 233-239.

[6] 罗智勇, 宋柔. 现代汉语自动分词中专名的一体化、快速识别方法[A]. 国际中文电脑学术会议论文集[C]. 新加坡, 2001. 323-328.
LUO Z Y, SONG R. Integrated and fast recognition of proper noun in modern Chinese word segmentation[A]. 2001 International Conference of Chinese Computing(ICCC'2001)[C]. Singapore, 2001. 323-328.

[7] SUN J, GAO J F, ZHANG L, *et al*. Chinese named entity identification using class-based language model[A]. Proc of the 19th International Conference on Computational Linguistics[C]. Taipei: Morgan Kauffmann Press, 2002.967-973.

[8] 刘群，张华平，俞鸿魁等. 基于层次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.
LIU Q, ZHANG H P, YU H K, *et al*. Chinese lexical analysis using cascaded hidden markov model[J]. Journal of Computer Research and Development, 2004,41(8):1421-1429.

[9] 张华平，刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报，2004,27(1): 85-91.
ZHANG H P, LIU Q. Automatic recognition of Chinese personal name based on role tagging[J]. Chinese Journal of Computers, 2004, 27(1):85-91.

[10] YU H, ZHANG H, LIU Q. Recognition of Chinese organization name based on role tagging[A]. Advances in Computation of Oriental Languages[C]. Beijing: Tsinghua University Press,2003. 79-87

[11] ZHANG H, LIU Q, YU H, *et al*. Chinese named entity recognition using role model[J]. The International Journal of Computational Linguistics and Chinese Language Processing, 2003,8(2): 1-31.

[12] RICHARD S, THOMAS E. The first international Chinese word segmentation bakeoff[A]. Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo: Sapporo Press, 2003. 133-143.

作者简介：

俞鸿魁（1978-），男，浙江镇海人，北京信息科技大学硕士、工程师，主要研究方向为自然语言处理。

当分词系统在人名识别 HMM 的基础上进一步引入地名识别 HMM 时，我们从表 6 中的数据发现，与只进行人名识别相比，各类人名的准确率 *P* 和 *F*−1 值都有不同程度的提高，分词系统的切分正确率也进一步提高到 97.98%。

表 6　　+LOC 地名识别性能指标

+LOC	*P*	*R*	*F*−1
中国人名	91.26	94.61	92.90
音译人名	86.03	92.60	89.19
日本人名	84.24	91.94	87.92
Σ人名	90.29	94.25	92.23
中国地名	76.90	96.23	85.48
音译地名	85.30	96.19	90.42
Σ地名	78.24	96.22	86.30

表 7　　+ORG 机构名识别性能指标

+ORG	*P*	*R*	*F*−1
中国人名	91.84	94.60	93.20
音译人名	86.81	92.60	89.61
日本人名	85.50	91.94	88.60
Σ人名	90.92	94.24	92.55
中国地名	93.38	95.62	94.49
音译地名	93.59	95.93	94.74
Σ地名	93.42	95.67	94.53
机构名	87.70	85.35	86.51

如表 7 所示，随着命名实体识别层叠隐马尔可夫模型的最高层——机构名识别 HMM 的加入，各类命名实体识别的性能又有不同幅度的提升，尤以地名识别最为明显，这时分词系统的切分正确达到了最高值 98.37%。

从以上测试我们可以看出，随着各层 HMM 的不断加入，不仅能极大地提升本层隐马尔可夫模型的命名实体识别的性能，而且还能进一步提高低层 HMM 识别的性能。

为了验证层叠隐马尔可夫模型的命名实体识别技术的实用性，采用《人民日报》1998 年 6 月的语料为测试语料，进行开放测试，测试结果如表 8 所示。

可以看出，在如此大规模语料的测试下，除了日本人名因为在训练语料中所占的比例太小而导致性能下降较大外，其它各类命名实体识别的性能仍旧很高。

表 8　　开放测试命名实体识别性能指标

OPEN	*P*	*R*	*F*−1
中国人名	81.27	89.11	85.01
音译人名	78.71	86.73	82.52
日本人名	49.23	60.38	54.24
Σ人名	80.47	88.40	84.25
中国地名	82.83	89.73	86.14
音译地名	85.60	80.70	83.08
Σ地名	83.25	88.19	85.65
机构名	74.56	73.02	73.78

3.2　第一届国际分词大赛中的未登录词识别

为了比较和评价不同方法和系统的性能，第四十一届国际计算语言联合会(41st annual meeting of the association for computational linguistics, 41th ACL)下设的汉语特别兴趣研究组(the ACL special interest group on Chinese language processing, SIGHAN；www.sighan.org) 于 2003 年 4 月 22 日至 25 日举办了第一届国际汉语分词评测大赛(first international Chinese word segmentation bakeoff)[12]。报名参赛的分别是来自于大陆、台湾、美国等 6 个国家和地区，包括微软研究院和北京大家计算语言所在内的共计 19 家研究机构，最终有 12 家提交结果。

大赛采取大规模语料库测试，以分词切分的 *F*−1 值为最终评价标准，语料库和标准分别来自北京大学（简体版）、宾州树库（简体版）、香港城市大学（繁体版），台湾“中央院”（繁体版）。每家标准分两个任务(track)：受限训练任务（close track）和非受限训练任务（open track）。

表 9　ICTCLAS 在第一届国际分词大赛中的测试结果

Track	CTBc	CTBo	PKc	PKo
参赛单位数	6	7	10	8
测试集规模 (bytes)	125,248	125,248	56,254	56,254
测试集分词数	39,922	39,922	17,194	17,194
Recall (Rank)	0.886 (2^{nd})	0.887 (4^{th})	0.962 (1^{st})	0.963 (1^{st})
Precision (Rank)	0.875 (1^{st})	0.876 (4^{th})	0.940 (3^{rd})	0.943 (2^{nd})
F−1 (Rank)	0.881 (1^{st})	0.881 (4^{th})	0.951 (1^{st})	0.953 (2^{nd})
OOV rate	0.181	0.181	0.069	0.069
OOV Recall (Rank)	0.705 (1^{st})	0.707 (5^{th})	0.724 (2^{nd})	0.743 (2^{nd})
IV Recall(Rank)	0.927 (5^{th})	0.927 (5^{th})	0.979 (2^{nd})	0.980 (1^{st})

ICTCLAS 在第一届国际分词大赛中的测试结果如表 9 所示。其中，OOV（out of vocabulary）是指未

张华平（1978-），男，江西波阳人，博士，中国科学院计算技术研究所助理研究员，主要研究方向为计算语言学、中文信息处理与信息抽取。

吕学强（1970-），男，山东鱼台人，博士，北京信息科技大学副教授，主要研究方向为自然语言处理和信息检索。

刘群（1966-），男，江西萍乡人，博士，中国科学院计算技术研究所研究员，主要研究方向为机器翻译和自然语言处理。

施水才（1966-），男，江苏溧阳人，硕士，北京信息科技大学教授，主要研究方向为中文信息处理。

（上接第 86 页）

作者简介：

时金桥（1978-），男，黑龙江哈尔滨人，哈尔滨工业大学博士生，主要研究方向为网络与信息安全、网络匿名通信技术。

程晓明（1970-），男，吉林德惠人，硕士，国家计算机网络应急技术处理协调中心工程师，主要研究方向为网络与信息安全。