第 39 卷 第 6 期
Vol.39 No.6

计 算 机 工 程
Computer Engineering

2013 年 6 月
June 2013

• 人工智能及识别技术 •　文章编号：1000—3428(2013)06—0266—06　文献标识码：A　中图分类号：TP391

基于 HMM 的京剧机构命名实体识别算法

乐 娟[1,2]，赵 玺[3]

(1. 北京理工大学计算机学院，北京 100081；2. 北京戏曲艺术职业学院，北京 100068；
3. 北京联合大学师范学院，北京 100011)

摘 要：针对机构命名实体识别效率低的问题，提出一种基于隐马尔科夫模型(HMM)的京剧机构命名实体识别算法。利用 HMM 模型标注文本切分结果的词性消除歧义，通过 Viterbi 算法计算某种分词结果所对应的可能性最大的词性序列。根据定制的名称识别规则，借助机构前缀词库、后缀词库获得机构名称左右边界，通过自动机算法识别语料中的机构命名实体，并将新词加载到分词词典中。针对京剧领域语料进行开放测试验证，结果表明，该算法的识别正确率可达到 99%。

关键词：开放领域；命名实体识别；隐马尔科夫模型；Viterbi 算法；规则树

Algorithm of Beijing Opera Organization Names Entity Recognition Based on HMM

LE Juan[1,2], ZHAO Xi[3]

(1. College of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;
2. Beijing Vocational Institute of Local Opera and Arts, Beijing 100068, China;
3. Teachers' College, Beijing Union University, Beijing 100011, China)

【Abstract】Aiming at the inefficiency of organization named entity recognition, this paper proposes an algorithm of Beijing opera organization Named Entity Recognition(NER) based on Hidden Markov Model(HMM). It uses HMM to take part-of-speech tagging and solve the problem of disambiguation of the words. The Viterbi algorithm is used to calculate the maximum probability tagging sequence to the sentence segmentation. It defines the rules to recognize the organization names. The left and right boundary of the organization is identified with the help of organization postfix lexicon. The new names in corpus are recognized by automatic algorithm and be loaded into the dictionary. This paper takes the test in open materials, the result shows the recognition accuracy can achieve 99%.

【Key words】open-domain; Named Entity Recognition(NER); Hidden Markov Model(HMM); Viterbi algorithm; rule tree
DOI: 10.3969/j.issn.1000-3428.2013.06.059

1 概述

随着信息搜索技术的不断进步，文本挖掘技术以其自动、高效获取知识的能力被越来越多的研究者所关注，其中，抽取出文本中人名、机构名称等词条的技术，即命名实体识别(Named Entity Recognition, NER)[1]技术，是文本挖掘技术重要研究方向之一。该技术是挖掘海量信息及其语义关系的前提与关键，被广泛应用于信息抽取、问答系统、句法分析、机器翻译等众多领域，显然命名实体的识别正确率是判定 NER 技术优劣的重要指标。

笔者目前正在开展“基于深度规则语义分析的京剧知识问答系统”构建研究，包含对京剧文献等相关信息中机构命名实体识别技术的研究。本文提出一种基于隐马尔科夫模型的京剧机构命名实体识别算法。在总结前人相关研究成果后，针对识别正确率较低这一问题，对传统命名实体识别算法进行改进，为京剧知识问答系统的构建做好前期准备工作。

2 京剧领域机构命名实体识别的特点

中文分词是在海量信息中抽取命名实体的第一步，即将文本以短语为单位进行切分。目前，中文分词系统主要采用基于词典匹配的分词方法，切分时按照词典中登录的词进行文本切分。当切分遇到词典中未登录的词条时会加大词以及词语之间存在的歧义，降低句子整体切分的精确度，利用命名实体识别技术则可以及时识别未登录命名实体并将其加载到分词词典以完善词典，一方面避免人工创

基金项目：北京市优秀人才培养计划基金资助项目(2012D002002000001)；北京市职业院校教师素质提高工程基金资助项目
作者简介：乐 娟(1978－)，女，高级讲师、硕士研究生，主研方向：信息检索，软件工程；赵 玺，实验师、硕士研究生
收稿日期：2012-05-28　**修回日期：**2012-08-20
E-mail：1659038116@qq.com

建词条的低效率，另一方面改善分词性能。在命名实体识别中，机构名称占有一定比例，机构名称是否被正确切分，直接影响切分粒度，以“杜镇杰：北京京剧院一团”举例说明机构名称对切分结果的影响：当分词词典中没有“北京京剧院一团”这个词时，对上句的切分结果如图 1 所示。识别出该机构命名实体并加载入分词词典后的切分结果如图 2 所示。

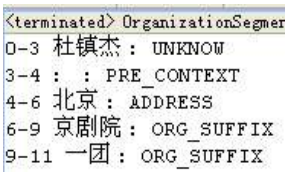

图 1　未识别的切分结果

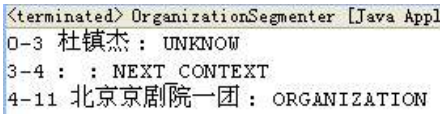

图 2　识别后的切分结果

通过对以上两图进行对比可以发现，机构命名实体的正确识别对句子整体切分正确率、切分速度以及切分性能的影响。以一则京剧演出信息举例说明京剧信息的特点：

长安大戏院演出京剧《西厢记》

演出时间：2012.07.17

演出地点：长安大戏院

主演：崔莺莺--洪岩，红娘--沈涛，张珙--外借，崔夫人--赵娟娟

剧情：唐代时，洛阳青年张珙赴试，途中宿普救寺，遇故相崔珏之女莺莺，一见钟情。驻军孙飞虎兵变，围住寺院，欲夺莺莺。崔夫人情急，声称将莺莺许婚给能退贼兵者。张珙请来友人白马将军杜确打退孙飞虎。崔夫人却悔约，使张与莺莺兄妹相称。张通过婢女红娘传递诗简，跳墙夜会莺莺，莺莺责备张珙非礼。张忧愤成疾，红娘引莺莺来问候，两人订情。崔夫人闻知，怒打红娘，红娘反责其失信。崔夫人无奈，只好约张珙中试后成婚。

在上述演出信息中包含着明显的京剧专业知识及知识间的关系，如演出单位与演员之间的关系、剧目[2]与剧目角色之间的关系、演员与角色之间的关系、剧情的介绍等。除去显示的京剧专业知识以外，其中不同的命名实体之间还存在着丰富的隐式的语义关系，比如：剧目中的角色“崔莺莺”在京剧中属于青衣行当，那么可以通过构建知识库推理得出演员“洪岩”的行当也属于青衣。因此，命名实体的识别是文本挖掘的第一步，识别出命名实体后进一步挖掘信息之间存在的语义关系，构建知识资源库，进而通过推理等技术得出实体关系，从而构建京剧知识问答系统。

3　研究现状

针对机构命名实体的自动识别问题，人们已经尝试过多种方法[3-4]，主要包括规则方法、统计方法以及规则与统计相结合等多种方法，但识别效率均有待提高。文献[5]提出了一种基于规则识别中文组织机构全称的方法，该方法借助机构后缀词库获得其右边界，然后通过规则匹配并借助贝叶斯概率模型加以决策获得其左边界的方法识别机构名称，该方法在开放性测试中的识别准确率达到 83.03%。文献[6]针对金融领域机构命名实体进行识别，在识别策略上综合考虑了机构名的结构特征和文本上下文信息，利用机器统计和人工辅助相结合的方法进行识别，在开放性测试中的准确率为 62.8%。文献[7]提出了一种基于角色标注的机构名称自动识别方法，其基本思想是采取 Viterbi 算法对切分结果进行角色标注，在角色序列的基础上，进行字符串识别，最终实现中文机构名的识别，开放测试的识别准确率达到 88%。

本文在前人研究工作的基础上，对京剧领域机构命名实体进行了分析研究，总结出其识别难点：(1)机构命名实体识别不同于人名识别，汉语人名长度基本可以确定，而机构名称的长度变化范围过大，如何界定机构名称的起始及终止位置是识别难点之一。(2)机构命名实体中常包含部分分词词典中已存在的词条，这对整句的切分会造成歧义。(3)类似“中国青年京剧团一团”的命名实体识别，通常“中国青年京剧团”是总称，“一团”是其下属单位，2 个词之间存在一定的级别关联。针对上述难点，本文利用隐马尔科夫模型(Hidden Markov Model, HMM) [8]对原文的一次分词结果进行词性标注，利用 Viterbi[9]算法消除分词结果存在的歧义，得出分词方案所对应的可能性最大的词性序列。同时在充分分析京剧领域机构命名实体构成特征的基础上，建立了京剧机构命名实体中常见的前缀、后缀、上下文等特征词库，进而定制符合机构名称及剧目名称的识别规则，识别原文中的命名实体并加载到分词词典。

本文与前人所做的工作不同之处在于，识别命名实体前先利用 HMM 模型消除分词结果存在的歧义，然后进行规则匹配。因为 HMM 模型很好地解决了歧义问题，所以机构命名实体的识别率得到大幅提高。同时，本文将机构命名实体的识别与京剧领域进行结合研究，较好地解决了命名实体识别的实用性问题。本文将上述方法应用于识别百度百科、互动百科、京剧门户网站等开放语料中与京剧领域相关的语料进行验证，实验结果表明，机构名称识别正确率可达到 92%，完全可以满足实际需求。

4　算法实现过程

本文应用算法识别京剧领域的机构命名实体和剧目命名实体，其识别过程分为 3 个模块，分别是原文处理模块、词性标注模块和规则匹配模块。原文处理模块接受用户输入的文本并以短语为单位进行切分，生成邻接表；词性标注模块通过 HMM 模型过滤切分结果的词性，消除歧义并

对每个切分结果标注词性；规则匹配模块利用自动机算法对词性标注结果序列进行规则匹配，满足规则的词序列即为所识别出的机构命名实体及剧目命名实体。上述过程中词性标注与规则匹配是机构名称与剧目名称识别算法的 2 个关键点。算法的实现过程如图 3 所示。

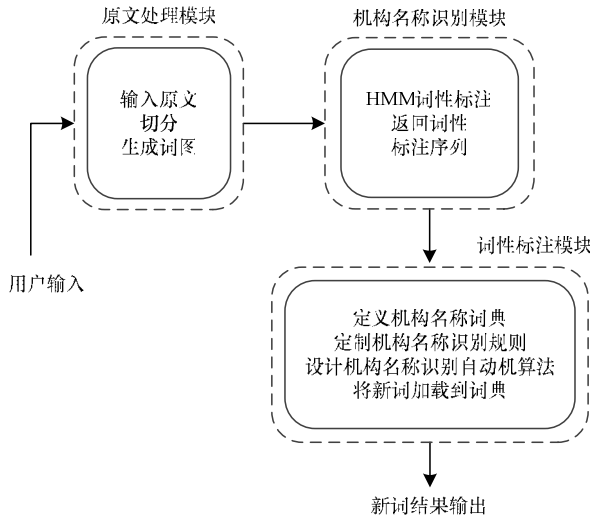

图 3 机构名称识别过程

5 词性标注

由于中文歧义，每个分词结果可能存在多个词性，词性标注(Part-of-Speech tagging)的关键问题就是消除这种歧义，为每个切分结果在一定的上下文中确定一个适当的词性，将隐马尔科夫模型应用于词性标注可以很好地消除歧义。HMM 模型[10]是一种用参数表示的用于描述随机过程统计特性的概率模型，包括隐状态和显状态，利用 HMM 模型能够完成从可以观测到的显状态序列计算出可能性最大的隐状态序列。具体到词性标注问题，显状态是文本的切分结果——单词 W，隐状态是每个分词需要标注的词性 C。HMM 模型要解决的问题就是利用 Viterbi[11]算法在初始概率、转移概率以及发射概率已知的情况下，计算出给定的某种分词结果所对应的可能性最大的词性序列。Viterbi 算法源于动态规划的思想，运行 Viterbi 算法前已经构建了初始概率表，词性之间的转移概率表以及每个词作为某个词性出现的发射概率表。其求解方法由 2 个过程组成：前向累积概率计算过程和反向回溯过程。文本的每个分词结果对应一个词，前向累积过程按阶段完成每个节点的概率计算，其中每个词对应一个求解阶段，计算得出每个词的词性。回溯过程则完成隐状态的提取。下面以“北京市百花京剧团”的计算过程举例说明。

首先对“北京市百花京剧团”进行切分，根据本文所使用的词性标注集，切分结果以及每个分词的词性如下：北京市(ADDRESS)/百花(Unknow, N)/京剧团(Org_Suffix)，其中，“百花”一词具有 2 个词性，需要消歧。“北京市百花京剧团”的发射概率示意图如图 4 所示。

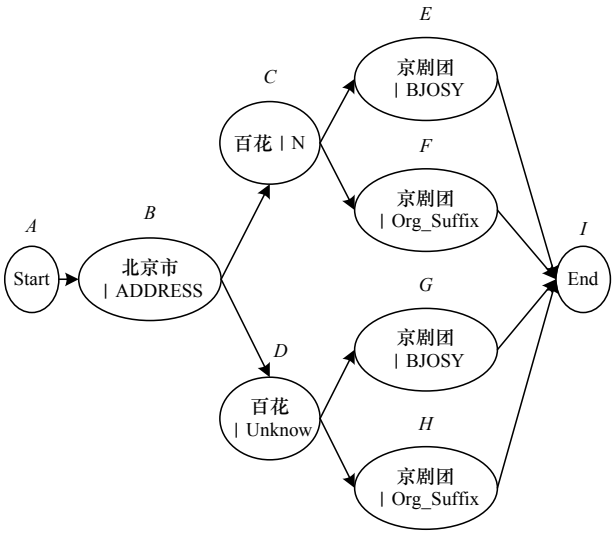

图 4 “北京市百花京剧团”的发射概率示意图

Viterbi 算法的前向累积过程[12]从节点“Start”开始计算每个节点的概率，当前节点概率的计算公式如下：

$$Best(Nodei)=Best(Nodei-1)+P(\text{词性}|\text{词性})+P(\text{词}|\text{词性})$$

其中，$Best(Nodei)$表示当前节点；$Best(Nodei-1)$表示上一个阶段的节点概率；$P(\text{词性}|\text{词性})$表示上一个阶段分词的词性到当前节点分词词性的转移概率；$P(\text{词}|\text{词性})$表示当前节点的词作为某个词性的发射概率。

前向累积计算节点概率如表 1 所示。

表 1 前向累积计算节点概率

节点	文本	计算
A	Start	$Best(A)=1$
B	北京市	$Best(B)=Best(A)+\log P(\text{Address}\|\text{Start})+\log P(\text{北京市}\|\text{Address})=1+(-11.013\ 1)+2.78=-7.232$
C	百花	$Best(C)=Best(B)+\log P(N\|\text{Address})+\log P(\text{百花}\|N)=(-7.232)+(-4.077\ 5)+(-10.990\ 3)=-22.300\ 4$
D	百花	$Best(D)=Best(B)+\log P(\text{Unknow}\|\text{Address})+\log P(\text{百花}\|\text{Unknow})=(-7.232)+(-4.077\ 5)+2.78=-8.529\ 4$
E	京剧团	$Best(E)=Best(C)+\log P(\text{BJOSY}\|N)+\log P(\text{京剧团}\|\text{BJOSY})=(-22.300\ 4)+(-11.683\ 4)+(-3.912\ 6)=-37.896\ 4$
F	京剧团	$Best(F)=Best(C)+\log P(\text{Org\_Suffix}\|N)+\log P(\text{京剧团}\|\text{Org\_Suffix})=(-22.300\ 4)+(-12.376\ 5)+2.78=-31.896\ 3$
G	京剧团	$Best(G)=Best(D)+\log P(\text{BJOSY}\|\text{Unknow})+\log P(\text{京剧团}\|\text{BJOSY})=(-8.529\ 4)+(-4.077\ 5)+(-3.912\ 6)=-16.519\ 6$
H	京剧团	$Best(H)=Best(D)+\log P(\text{Org\_Suffix}\|\text{Unknow})+\log P(\text{京剧团}\|\text{Org\_Suffix})=(-8.529\ 4)+(-4.077\ 5)+2.780\ 6=-9.826\ 4$
I	End	$Best(I)=\text{Max}(Best(E), Best(F), Best(G), Best(H))=Best(H)=-9.826\ 4$

前向累积过程完成后开始执行回溯过程，并发现最佳隐状态(红线连接的节点)，所以猜测词性输出：北京市/ADDRESS 百花/Unknow 京剧团/Org_Suffix。图 5 中虚线箭头标示出的线路即为标注所得的词性序列。

词性标注模块通过 HMM 模型解决歧义问题，并通过对显状态的计算回溯得出可能性最大的隐状态序列，即根据分词结果推测计算出词性序列。标注模块将词性序列传给规则匹配自动算法，开始进行规则匹配。

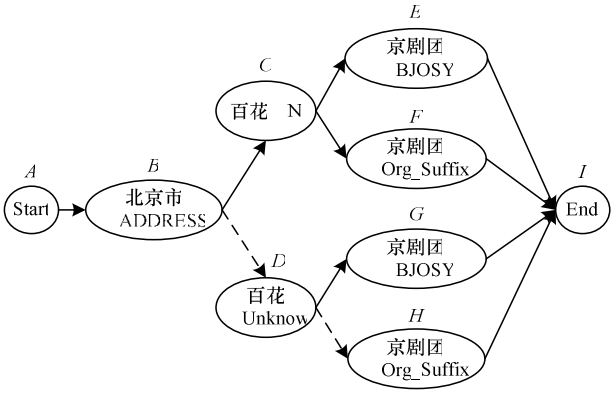

图 5　最佳隐状态示意图

6　规则匹配

实现规则匹配自动算法需要形成识别机构名称的词典，包括前缀词典(定位机构名称左边界)、后缀词典(定位机构名称右边界)、上下文词典、地址词典、人物名称词典和部分关键词词典。本文主要针对京剧院团机构命名实体进行识别测试，因此，上述词典中包含的词涵盖了京剧组织名称中常见的关键字，比如“北京京剧院”中的地名“北京”、“梅兰芳大剧院”中的人物名称“梅兰芳”等。通过对语料进行分析可以发现机构名称前后常会出现一些标志性词语及标点符号，比如:“长安大戏院(点击查看剧院情况)”，其中，左括号是“长安大戏院”的下文，“杜镇杰:北京京剧院一团领衔主演”，其中，冒号是“北京京剧院一团”的上文。因此，总结机构名称前后经常出现的上下文，包括全角中文标点符号及英文标点符号等，将其划分为上文、下文 2 种类型并载入上下文词典。对于词典中不包含的词进行合并，定义为未知类型词。本文中机构名称识别专用词典中一共设置以下类型:

(1)START//开始节点;

(2)END//结束节点;

(3)PRE_CONTEXT//上文;

(4)NEXT_CONTEXT//下文;

(5)BJO_PERSON//人名，机构名称中经常出现的人物名称;

(6)ADDRESS//地名，机构名称中经常出现的地域性词汇;

(7)BJO_KEYWORD//行业关键词，机构名称中经常出现的与行业相关的关键词;

(8)ORG_SUFFIX//机构后缀;

(9)ORGANIZATION//机构名;

(10)UNKNOW//未知词。

完成词典的设计后，根据机构名称的组成特点定制机构名称识别规则，比如“北京百花大剧院”经过 HMM 模型消歧后得到的词性序列为“Address+Unknow+Org_suffix”，如果规则表中有以上序列的规则，那么就将“北京百花大剧院”识别为一个机构名称。本文中定义了 17 条机构名称识别规则，如表 2 所示。

表 2　机构名称识别规则

序号	规划	示例
1	Address+unknow+Org_Suffix	北京百花大剧院
2	Address+bjo_keyword+Org_Suffix	上海京剧艺术团
3	Address+bjo_keyword+ bjo_keyword+Org_Suffix	北京京剧艺术学院
4	Address+Org_Suffix	北京京剧院
5	Address+Address+ bjo_keyword+Org_Suffix	首都朝阳国粹艺术网
6	Address+Address+Org_Suffix	中国上党戏曲网
7	Pre_Context+unknown+Org_Suffix	长安大戏院
8	Next_Context+unknown+Org_Suffix	长安大戏院
9	Pre_Context+unknown+ bjo_keyword+Org_Suffix	玲珑戏曲艺术网
10	Next_Context+unknown+ bjo_keyword+Org_Suffix	玲珑戏曲艺术网
11	Start+unknown+ bjo_keyword+Org_Suffix	华夏戏曲网
12	Start+bjo_keyword+ unknown+Org_Suffix	京剧戏考网
13	bjo_keyword+Pre_Context+ Address+Next_Context+unknow	科班有北京的富连成
14	bjo_keyword+Org_Suffix	京剧戏考网
15	Bjo_Person+bjo_keyword+Org_Suffix	李少春京剧艺术网
16	BJO_Person+Org_Suffix	梅兰芳大剧院
17	Unknown+bjo_keyword+Org_Suffix	张克京剧艺术网

将上述规则以树结构存储，从根结点出发，一个分支表示一条识别规则，双圈结点表示规则的终止结点。以上 17 条识别规则所形成的规则树如图 6 所示。

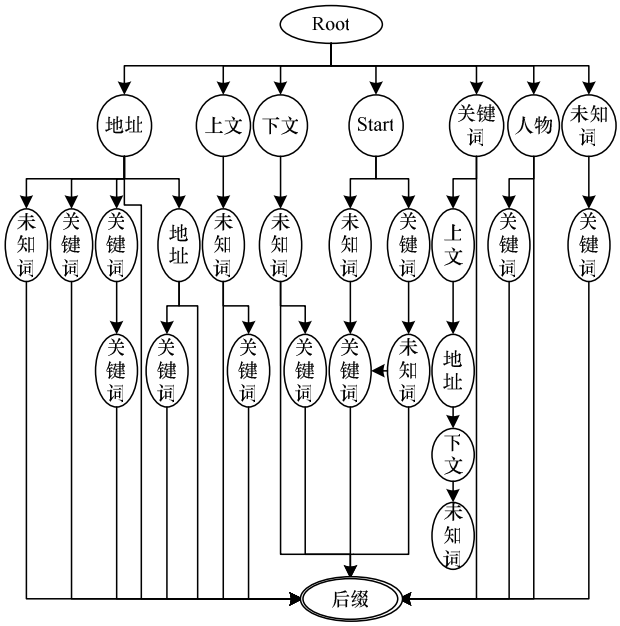

图 6　机构名称识别规则树

以识别“天津市青年京剧团是全国京剧重点院团”中的机构名称“天津市青年京剧团”为例说明机构名称识别器的具体工作过程。算法首先对原文进行一次切分，切分结果如表 3 所示。

表 3 “天津市青年京剧团”的一次切分结果

开始-结束	文本	词性
0-3	天津市	ADDRESS
3-5	青年	BJO_KEYWORD
5-8	京剧团	ORG_SUFFIX

HMM 模型对以上切分结果消歧并标注词性，并将标注后的结果传给规则匹配自动算法进行匹配。通过观察表 3 可以发现，“天津市”、“青年”、“京剧团”3 个词条文本所具有的词性序列恰好符合下图规则树中虚线标记出的一条路径，如图 7 所示，图中的虚线路径表示与识别规则相匹配的一条路径。因此，识别器将符合识别规则的文本组合“天津市青年京剧团”作为机构名称加载到分词词典。

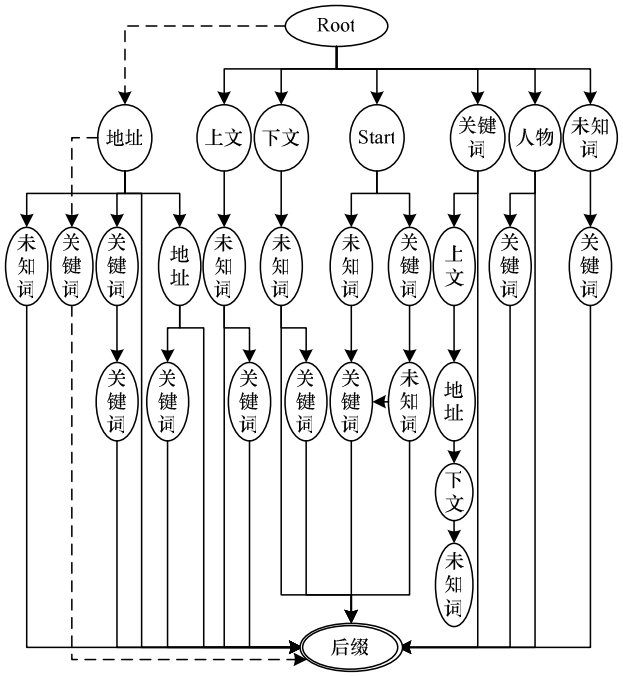

图 7 符合机构名称识别规则的路径

程序对以下原文进行测试：

原文：11 家获选的京剧院团为：中国京剧院、北京京剧院、天津京剧院、上海京剧院、湖北省京剧院、天津市青年京剧团、山东省京剧院、云南省京剧院、沈阳京剧院、黑龙江省京剧院、江苏省京剧院。

识别结果：[中国京剧院，北京京剧院，天津京剧院，上海京剧院，湖北省京剧院，天津市青年京剧团，山东省京剧院，云南省京剧院，沈阳京剧院，黑龙江省京剧院，江苏省京剧院]。

本文通过上述方法定制规则对网页中的京剧剧目[8]命名实体同时进行识别，效果良好，以下文举例：

原文：代表剧目有《伍子胥》、《杨家将》、《失空斩》、《击鼓骂曹》、《洪羊洞》、《法场换子》、《红鬃烈马》和《精忠报国》等。

识别结果：[伍子胥，杨家将，失空斩，击鼓骂曹，洪羊洞，法场换子，红鬃烈马，精忠报国]。

7 实验结果及分析

7.1 测试集及评测指标

下面给出实验过程中所采取的测试集和评测指标以及实验结果分析。机构命名实体识别评测所用的测试集一般有 2 种：只含有专名句子的测试集和完全真实的语料库。只含有专名的句子测试集会忽略真实语言环境中不含特定专名的句子，测试结果可能包含不是机构命名实体的短语，即不是机构命名实体的短语会被错误地识别出来。因此，采用真实的语料库进行测试更具有实际的分析意义，能体现算法的性能及效率。在不做任何筛选的大规模真实语料库上进行测试，评测结果更符合真实的语言环境。测试时根据测试集和训练集的不同关系，测试可以分为封闭测试与开放测试 2 种。封闭测试通常使用实验时所用的训练子集进行测试，这样的测试结果往往比较理想，但是真正将算法用于非训练子集时，实验结果偏差会比较大。因此，本文的测试采用开放测试，所有测试用语料均采用来自互动百科、百度百科以及维基百科中与京剧领域相关的开放语料，所有测试用语料均未经过任何处理，直接用于算法测试。实验时直接对语料进行分词、词性标注，在得到词性标注序列的基础上进行机构命名实体识别。针对机构命名实体识别，采取 3 个常用的评测指标，即准确率(P)、召回率(R)以及综合指标 F 值(F)。其定义如下：

P=正确识别出的名称数/识别出的名称数×100%

R=正确识别出的名称数/实际名称数×100%

$$F=\frac{R\times P\times(1+\beta^2)}{R+P\times\beta^2}$$

其中，β 是准确率 P 和召回率 R 之间的权衡因子，本文实验 β 取 1。

7.2 结果分析

对识别算法进行测试时，分别进行 2 种测试。第 1 种将算法用于机构命名实体识别。测试时利用网络爬虫抓取百度百科等开放领域语料中与京剧领域相关的语料直接测试算法。因为设计算法识别规则时所采用的前缀、后缀及上下文集均与京剧领域相关，所以算法识别的准确率、召回率与 F 值均理想，完全满足实际需求。识别算法完成识别任务后即将所识别的新词加载入分词词典，及时补录词典以提高分词的速度与精度。机构名称测试结果见表 4。

表 4 机构名称测试结果

指标	结果
语料库大小	1 031 KB
实际名称数	358
识别出的名称数	369
正确识别的名称数	338
准确率	91.6%
召回率	94.4%
F 值	93%

第 2 种测试将前缀、后缀及上下文集稍做改动，针对京剧剧目命名实体的上下文环境进行调整，进行京剧剧目的识别。京剧剧目的表示有其显著特点，通常京剧剧目前后会有明显的书名号，演员名称等有助于识别规则进行匹配，因此识别的准确率、召回率及 F 值均理想，完全可以满足实际需求。识别算法完成识别任务后即将所识别的剧目名称以指定的词性加载入分词词典，及时补录词典以提高分词的速度与精度。剧目名称测试结果如表 5 所示。

表 5　剧目名称测试结果

指标	结果
语料库大小	987 KB
实际名称数	583
识别出的名称数	579
正确识别的名称数	571
准确率	98.6%
召回率	97.9%
F 值	99.5%

本文将上述算法识别出的命名实体结果加载到京剧行业词典，通过不断完善的京剧行业词典对京剧相关语料进行分词，以下述文本举例说明：

“北京京剧院是经由国家文化部评定的全国京剧重点院团，是北京市文化局主管的差额事业单位。成立于 1979 年，其前身是梅、尚、程、荀四大名旦各自领导的流派剧团汇合而成的北京市京剧团和以马连良、张君秋、谭富英、裘盛戎、赵燕侠为领衔主演的北京京剧团。全院阵容强大，名角荟萃，流派纷呈，剧目丰富。代表剧目有《画龙点睛》、《杨家将》、《失空斩》、《击鼓骂曹》、《洪羊洞》、《法场换子》、《红鬃烈马》和《精忠报国》等。”

利用本文的分词系统对上述文本进行切分，其中“北京京剧院、北京市文化局、北京市京剧团、北京京剧团”为机构命名实体识别结果，“马连良、张君秋、谭富英、裘盛戎、赵燕侠”为人名命名实体识别结果，“画龙点睛、失空斩、击鼓骂曹、洪羊洞、法场换子、红鬃烈马、精忠报国”为剧目名称实体识别结果，具体切分结果如下：

北京京剧院/BJOJG 是/V 经由/P 国家/N 文化部/N 评定/VN 的/UJ 全国/N 京剧/N 重点/D 院团/BJOSY，/W 是/V 北京市文化局/BJOJG 主管/VN 的/UJ 差额/N 事业/N 单位/N。/W 成立/V 于/P 1979/m 年/Q，/W 其/P 前身/F 是/V 梅/NG，/W 尚/VG、/W 程/NRF、/W 荀/NRF 四大名旦/BJOSY 各自/UNKNOW 领导/N 的/UJ 流派/N 剧团/N 汇合/V 而/C 成/V 的/UJ 北京市京剧团/BJOJG 和/C 以/P 马连良/BJOYY、/W 张君秋/BJOYY、/W 谭富英/BJOJS、/W 裘盛戎/BJOYY、/W 赵燕侠/BJOYY 为/V 领衔/VD 主演/V 的/UJ 北京京剧团/BJOJG。/W 全院/N 阵容/N 强大/A，/W 名角/N 荟萃/V，/W 流派/N 纷呈/V，/W 剧目/N 丰富/A。/W 代表/N 剧目/N 有/V《/W 画龙点睛/BJOJM》/W、/W《/W 杨家将/N》/W、/W《/W 失空斩/BJOJM》/W、/W《/W 击鼓骂曹/BJOJM》/W、/W《/W 洪羊洞/BJOJM》/W、/W《/W 法场换子/BJOJM》/W、/W《/W 红鬃烈马/BJOJM》/W 和/C《/W 精忠报国/BJOJM》/W 等/U。/W

本文分别对未使用命名实体识别的切分算法以及使用命名实体识别算法进行测试，分别得到切分正确率，如表 6 所示。

表 6　切分正确率比较　　(%)

切分算法	切分正确率
未用命名实体识别算法	92
使用命名实体识别算法	99

通过对以上切分正确率进行对比分析可以发现，本文所提出的分词系统在处理京剧领域语料时的分词正确率以及切分粒度均优于目前主流中文分词系统。本文所使用的分词系统在引入人名命名实体识别、机构命名实体识别及京剧剧目名称命名实体识别算法后，分词的正确率达到 99%，完全满足系统的实际需求。

8　结束语

本文通过分析机构命名实体的构成特色及其相关统计信息，提出了一种基于 HMM 模型的自动识别算法，识别机构命名实体及京剧剧目命名实体。本文将自然语言处理技术与人工智能技术应用于京剧领域展开研究，在互联网的海量开放语料中通过以上技术获得有实际帮助的内容，建立起行业词典库，这种结合特定领域研究人工自然语言处理技术的应用很好地解决了实用性问题，不是范围广效率低的重复开发，而是针对某一具体领域高效率地利用海量资料解决实际问题。通过对大规模完全真实语料库的开放测试，该方法取得较好的效果，能够满足实际的需求。下一步的研究工作是分析人名命名实体、机构命名实体以及其他信息之间存在的潜在语义关系，通过开放领域的语料抽取实体关系，构建具有推理能力的京剧专业知识库，最终完成基于自然语言处理技术的京剧知识问答系统的构建。

参考文献

[1] Wang Dan, Fan Xinghua. Named Entity Recognition for Short Text[J]. Journal of Computer Applications, 2009, 29(1): 143-145.

[2] 曾白融. 京剧剧目辞典[M]. 北京: 中国戏剧出版社, 1989.

[3] 孙　镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010, 26(6): 42-47.

[4] Chinchor N A. Overview of MUC-7/MET-2[C]//Proc. of the 7th Message Understanding Conference. Fairfax, Virginia: [s. n.], 1998.

[5] 沈嘉懿, 李　芳, 徐飞玉. 中文组织机构名称与简称的识别[J]. 中文信息学报, 2007, 21(6): 17-20.

(下转第 286 页)

由图 4 可知：(1)从处理效果上分析，传统水平集方法存在欠分割现象。该方法只利用了图像边界的梯度信息，因此，在迭代过程中其零水平集曲线容易停留在梯度值较大的非目标区域，使得最终演化结果偏离海马真实边界，海马轮廓得不到准确分割；改进方法能够很好地分割海马，克服了传统方法处理结果存在的问题，所利用的图像梯度和全局方差信息可帮助零水平集曲线向内越过灰度相对种子点波动较大的背景(如脑脊液、中脑及脑白质)区域，从而停留在灰度波动较小的区域梯度极值处。(2)从时间上分析，为更好地比较 2 种方法运算的时间，两者均根据其零水平集曲线的停止演化来确定水平集的最小迭代次数。经实验可知，传统方法中水平集最小迭代次数为 200 次，迭代次数较多，对序列图像进行处理时较为费时；而改进方法在传统方法基础上增加一个驱动水平集演化的动力，可使零水平集曲线相对快速地演化，经实验可知，改进方法仅需 50 次即可完成水平集演化，迭代次数少，因此，在运算时间上优于前者。

5 结束语

本文提出一种基于改进水平集的人脑海马图像分割方法。针对水平集初始轮廓的定义，用包含自适应区域生长全部目标在内的凸多边形作为水平集演化的初始轮廓，减少了传统初始轮廓定义方法中所产生的一系列处理偏差。除利用图像梯度外，该方法还考虑了灰度波动对分割结果的影响，使得零水平集曲线能够越过背景区域的梯度局部极值处，最终到达海马的真实边界。实验结果表明，该方法的分割精度明显高于传统方法，且其演化速度也优于传统方法，对序列图像进行处理时可节约较多时间，实现简单，是一种有效的人脑海马轮廓分割方法。除海马外，本文方法对其他器官组织的轮廓提取也具有一定的借鉴价值。

由于该方法仅从海马二维图像上研究海马轮廓的分割，并未从海马三维序列上考虑，因此今后的研究重点将根据海马三维序列上下层之间的关系，进一步提高海马二维轮廓分割的准确性。

参考文献

[1] 刘智华, 钱学华, 周庭永, 等. 人脑海马结构的形态测量[J]. 中国临床解剖学杂志, 2011, 29(5): 532-535.

[2] 孟 薇, 叶德荣. 人脑磁共振图像中海马结构的计算机辅助分割新方法[J]. 首都医科大学学报, 2008, 29(3): 332-335.

[3] 鲁 向, 芦 勤, 罗述谦. 基于能量驱动的分水岭算法在 MRI 海马图像分割中的应用[J]. 计算机辅助设计与图形学学报, 2008, 20(6): 748-752.

[4] 张继武, 张道兵, 史舒娟, 等. 基于水平集方法的数字胸片图像分割[J]. 中国图象图形学报, 2004, 12(12): 1459-1465.

[5] 龚永义, 罗笑南, 黄 辉, 等. 基于单水平集的多目标轮廓提取[J]. 计算机学报, 2007, 30(1): 120-128.

[6] Chan T, Vese L. Active Contours Without Edges[J]. IEEE Transactions on Image Processing, 2001, 10(2): 266-277.

[7] 杨 勇, 马志明, 徐 春. LCV 模型在医学图像分割中的应用[J]. 计算机工程, 2010, 36(10): 184-186.

[8] Li Chunming, Kao Chiu-Yen, Gore J C, et al. Implicit Active Contours Driven by Local Binary Fitting Energy[C]//Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 2007.

[9] Li Chunming, Kao Chiu-Yen, Gore J C, et al. Minimization of Region-scalable Fitting Energy for Image Segmentation[J]. IEEE Transactions on Image Processing, 2008, 17(10): 1940-1949.

[10] 何传江, 李 梦, 詹 毅. 用于图像分割的自适应距离保持水平集演化[J]. 软件学报, 2008, 19(12): 3161-3169.

[11] 付英杰, 张 剑, 邹 翎, 等. 基于改进水平集和区域生长的轮廓提取方法[J]. 计算机应用研究, 2012, 29(7): 2770-2772.

[12] 夏 晶, 孙继根. 基于区域生长的前视红外图像分割方法[J]. 激光与红外, 2011, 41(1): 107-111.

编辑 刘 冰

(上接第 271 页)

[6] 王 宁, 葛瑞芳, 苑春法. 中文金融新闻中公司名称的识别[J]. 中文信息学报, 2002, 16(2): 1-6.

[7] 俞鸿魁, 张华平, 刘 群. 基于角色标注的中文机构名识别[C]//Proc. of the 20th International Conference on Computer Processing of Oriental Languages. Shenyang, China: [s. n.], 2003.

[8] 梁吉光, 田俊华, 姜 杰. 基于改进 HMM 的文本信息抽取模型[J]. 计算机工程, 2011, 37(20): 178-179.

[9] Gao Wen, Fang Gaolin, Zhao Debin. A Chinese Sign Language Recognition System Based on SOFM/SRN/ HMM[J]. Pattern Recognition, 2004, 37(12): 2389-2402.

[10] Boyle R. HMM Turorial[EB/OL]. (2012-01-01). http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html.

[11] Fleming C. A Tutorial on Convolutional Coding with Viterbi Decoding, Spectrum Applications[EB/OL]. (2011-06-21). http://home.netcom.com/~chip.f/viterbi/tutorial.html.

[12] 罗 刚. 自然语言处理技术与实现[M]. 北京: 电子工业出版社, 2009.

编辑 顾姣健