DOI:10.13203/j.whugis20150538　　文章编号:1671-8860(2018)01-0017-07

一种基于复合特征的中文地名识别方法

魏　勇[1]　李鸿飞[1,2]　胡丹露[2]　李　响[2]　马雷雷[3]

1　31008 部队,北京, 100091
2　信息工程大学地理空间信息学院,河南 郑州, 450051
3　95291 部队,湖南 衡阳, 421010

摘　要:中文地名识别是命名实体识别的重要研究课题之一,也是提高地理信息系统应用水平的关键。传统的地名识别主要基于词性或地名要素特征,特征类型有限。提出了一种基于复合特征的中文地名识别方法,挖掘中文地名在自然语言中的特点,设计了类型、路径、距离和数量四种句法特征,基于地名要素特征、词性特征、句法特征三种复合特征利用条件随机场模型实现了中文地名的训练和识别。通过实验对比复合特征在中文地名识别方法的效果,结果表明复合特征能够有效提高中文地名识别的准确率和召回率,尤其是对于复杂地名的识别,具有良好的效果。

关键词:地名识别;复合特征;句法分析;条件随机场

中图法分类号:P208　　**文献标志码**:A

自然语言是人类对空间物像认知结果的重要表现形式,从自然语言中获取地理空间信息是地理信息科学的重要研究议题[1]。实现自然语言中地理空间信息的抽取,不仅能够丰富地理空间信息的数据来源,而且可以进一步提高地理空间信息的表达能力和交互能力[2]。地名是人们对特定空间位置的文本标识,是自然语言中重要的地理信息实体。中文文本中的地名识别对于地理信息系统应用、地理信息检索、基于位置的服务等领域都具有重要意义。

地名识别的方法主要分为基于规则的方法和基于统计的方法两种类型。基于规则的方法表达直观、自然,便于人工理解和扩展,但规则编写依赖具体的语言知识和领域知识,规则较为复杂,很难覆盖全部的模式,可移植性也比较差。基于统计的方法不需要过多的语言知识和领域知识,可移植性强,但需要人工标注语料库,并选择合适的统计学习模型及参数。程昌秀[3]、张雪英[4]、谭侃侃等[5]讨论了基于规则的中文地名识别方法,通过设计地址要素库、定义地址匹配规则来实现中文地名的识别;杜萍等[6]提出一种基于本体的中文地名识别方法,引入地名本体识别文本中的县级以上行政区划地;邱莎等[7]提出了使用条件随机场(conditional pandom fields,CRF)在字一级粒度上对中文地名的自动识别方法,通过丰富的特征组合和大规模语料训练,取得了良好的识别效果;唐旭日[8]讨论了中文文本的地名解析流程,提出基于条件随机场和篇章地名关系的地名识别方法、基于局部模糊匹配的地名标准化方法以及基于认知显著度的地理编码方法;Aaron[9]提出了字符级别的中文命名实体识别条件随机场模型,利用单字字符的特点实现人名、地名和组织名的识别;Chen 等[10]利用边界特征和单字特征进行中文命名实体识别,并对句子中的命名实体识别结果进行筛选后处理,取得了较好的识别结果。

语言的分析和理解过程是一个层次化的过程,现代语言学家把这一过程分为词法分析、句法分析和语义分析三个层次[11-12]。现有的地名识别算法,多是使用词性、词缀或词典作为特征,进行规则匹配或统计学习。词性和词典特征属于词法分析的范畴,这种分析只利用了词法这一级别的信息,而没有考虑自然语言中的句法背景。在中文自然语言中,有些候选的地名短语由于语义歧义,仅仅通过词法信息并不能完全判断其是否为

收稿日期:2016-01-10
项目资助:国家自然科学基金青年基金(41401467);四川省应急测绘与防灾减灾工程技术研究中心开放基金(K2015B014)。
第一作者:魏勇,博士,主要从事互联网空间数据获取与信息抽取研究。whuwy@163.com
通讯作者:胡丹露,博士,教授。hudanlu@vip.sina.com

地名,需要结合其在语言中的上下文信息来判断。句法结构是一种常见的上下文信息,对于地名的识别具有重要的意义。本文提出一种基于复合特征的中文地名识别方法,结合地名要素特征、词性特征和句法特征,使用条件随机场来进行中文地名的训练和识别,实验表明,基于复合特征的条件随机场能够有效识别中文地名信息,具有较高的准确率和召回率。

1 条件随机场与中文句法分析

1.1 条件随机场模型

中文地名识别可以看作是序列标注问题。地名是多个词语按照一定的顺序排列组合而成,地名实体识别就是从这些词语序列中标注出正确名称的组合。序列标注问题的有效解决方法是条件随机场模型,Lafferty 等[13]提出一种判别式概率无向图学习的条件概率模型,它结合了在最大熵模型和隐马尔科夫模型优点,能够用于序列数据的标注和切分。

对于给定的观测序列 $X=\{x_1,x_2,\ldots,x_n\}$,条件随机场将其对应的状态序列 $Y=\{y_1,y_2,\ldots,y_n\}$ 的条件概率定义为:

$$P(Y|X)=\frac{1}{Z(X)}\exp\Big(\sum_{i,j}l_j t_j(y_{i-1},y_i,x,i)+\sum_{i,k}m_k s_k(y_i,x,i)\Big) \quad (1)$$

式中,$Z(X)$ 为归一化因子,使得所有状态序列的概率和为 1;$t_j(y_{i-1},y_i,x,i)$ 为观测序列 $i-1$ 和 i 标记处的状态转移函数;$s_k(y_i,x,i)$ 是观测序列 i 标记处的状态特征函数;l_j 和 m_k 是相应特征函数的权值,通过训练估计得到。在建立 $P(Y|X)$ 的概率模型后,状态序列标记 Y 的求解就可以转化为求解 $P(Y|X)$ 最大化时的 Y^*:

$$Y^*=\mathrm{argmax}_Y P(Y|X) \quad (2)$$

条件随机场使用的概率图模型能够有效表达长距离的、相互依赖的特征,且所有特征可以进行全局归一化,进而求得全局最优解。

1.2 中文句法分析

句法分析根据给定的语法,自动推导出句子的语法结构,确定句子所包含的句法单位以及这些句法单位之间的关系,它将句子从线性的词语序列转换为结构化的句法树,从而可以捕捉到句子内部词语之间的修饰或搭配关系。句法分析的主要任务是消除句子在句法结构上的歧义,为句子的正确理解提供语法基础。

目前存在短语结构句法分析、依存关系句法分析两种主流的句法分析方法[14]。短语结构分析的目的是正确划分出句子中的短语结构以及这些短语之间的层次结构关系。依存关系分析则通过研究语言单位内成分之间的依存关系揭示其句法结构,它认为句子中的核心动词是全句支配其它成分的中心,其本身不受其它任何成分的支配,所有受支配成分都以某种依存关系从属于支配者。依存句法分析可以反映出句子各成分之间的修饰关系,能够获得长距离的搭配信息,且与句子成分的具体位置无关。

2 中文地名识别的特征选择

2.1 中文地名特点分析

条件随机场是一种有监督的机器学习方法,通过学习标注集中的特征数据来构建预测模型,特征的选择直接影响着条件随机场模型的性能,因此需要充分挖掘命名实体上下文的相关信息,并有效地将信息融合起来。传统的地名实体识别使用词语、词性或地名要素作为主要特征,而没有考虑到自然语言的上下文环境。

在中文自然语言中,地名一般为名词性短语,由名词、数词或量词组成。地名在句子中主要用作主语、宾语或状语。以地名“天安门广场”为例,其主语、宾语、状语和定语的用法示例如下。

(1)主语:天安门广场是世界上最大的城市中心广场。

(2)宾语:北京的标志性建筑是天安门广场。

(3)状语:中国人民抗日战争暨世界反法西斯战争胜利70周年大会于2015年9月3日上午在天安门广场举行。

(4)定语:天安门广场上五星红旗迎风飘扬。

句法分析是自然语言识别的核心任务之一,其输入为经过分词后的句子,输出为短语结构或依存关系结果。常见的中文句法分析工具有哈尔滨工业大学 LTP、Stanford Parser、Berkeley Parser 等。本文选择 Stanford Parser 作为中文自然语言句法分析的工具,Stanford Parser 是斯坦福大学自然语言研究小组开发的语法解析工具,使用词汇化的概率上下文无关算法对自然语言进行分析,能够进行词性标注、短语结构分析、依存关系分析,并提供了多种语言模型数据用于多语种文本的处理。以“天安门广场是世界上最大的城市中心广场”为例,使用中文分词工具 ICTCLAS 对其进行分词处理,再使用 Stanford Parser 对其

进行中文句法分析,得到的短语结构和依存关系分析结果如图 1 所示,左侧为短语结构分析结果,右侧为依存句法分析结果。

```
(ROOT
  (IP
    (NP (NN 天安门) (NN 广场))
    (VP (VC 是))
      (NP
        (CP
          (IP
            (VP
              (LCP
                (NP (NN 世界))
                (LC 上))
              (ADVP(AD 最))
              (VP (VA 大))))
          (DEC 的))
        (NP (NN 城市) (NN 中心) (NN 广场))))))
```

```
nn (广场-2，天安门-1)
nsubj (广场-11，广场-2)
cop (广场-11，是-3)
loc (大-7，世界-4)
case (世界-4，上-5)
advmod (大-7，最-6)
relcl (广场-11，大-7)
mark (大-7，的-8)
nn (广场-11，城市-9)
nn (广场-11，中心-10)
root (ROOT-0，广场-11)
```

图 1　基于 Stanford Parser 的句法分析结果

Fig.1　Syntactic Analyze Result Based on Stanford Parser

通过句法分析可以看出,“天安门广场”为一个名词性短语(NP),它由两个名词构成(NN 天安门和 NN 广场),广场-2 是句子的名词性主语(nsubj),其宾语为广场-11,整句为“主-系-宾”结构,系动词为“是”(cop)。因此通过句法分析可以正确分析出句子的成分及成分之间的关系,这种关系与词语的具体含义和位置无关,是结构化的语言上下文信息。

为了更有效地提高中文地名识别效果,本文设计了地名要素、词性和句法 3 大类特征,分别描述了中文地名在自然语言中的词缀、词性和句法信息。句法特征包括类型特征、路径特征、距离特征和数量特征四种。

2.2　地名要素特征

中文地名通常是由多个要素构成,每个地名要素为地名实体中的一个独立部分,例如“郑州市二七区陇海中路 66 号”由 4 个地名要素构成:“郑州市”、“二七区”、“陇海中路”、“66 号”。这 4 个地名要素表达了不同等级的信息。中文自然语言中常见的地名要素特征如表 1 所示。在条件随机场中,地名要素特征用来标示当前词语与地名要素之间的关系:如果当前词语中包含地名要素,则地名要素特征为要素符号;如果不包含,则地名要素特征为空。地名要素特征标记为“GP”。

2.3　词性特征

词性是指语词在其语法功能分类中具有的属性,词性特征表达了词语在语法结构中的功能,是一种重要的命名实体识别特征。

本文将词性特征标记为“POS”,使用中文分词工具 ICTCLAS 进行分词和词性标注,词性特征使用北京大学《现代汉语语料库加工规范》[15]。

表 1　地名要素特征表

Tab.1　Gazetteer Feature Table

要素类别		要素符号	举例说明
行政区界	省级	RD1	省、直辖市、自治区、特别行政区
	市级	RD2	市、地区、盟、自治州
	县区级	RD3	县、旗、区
	乡镇级	RD4	乡、镇、街道办
	村级	RD5	村、庄、屯、里
道路		LR	路、大道、道、大街、街、巷、胡同、条、里
住宅小区		PA	里、区、园、坊、居、寓、苑
标志建筑		PH	大厦、广场、饭店、中心、大楼、楼、场、广场、馆、酒店、局
门牌号		PD	号、#

2.4　句法特征

句法特征从句法层次上描述了句子元素之间的关系,句法特征在命名实体识别[16-17]、语句分类[18-20]、关系抽取[21-25]、机器翻译[26]、自动问答[27-28]等领域中得到了广泛的应用。

目前机器学习领域中对于句法特征的使用主要包括类型特征、路径特征、距离特征和数量特征四种。类型特征是指词语在句法结构中的短语类型或依存关系类型,它是句法特征的核心要素,从句法的角度描述了词语在句子中的角色特点;路径是语法树中当前词语节点到根节点的遍历路径,描述了词语与句子核心词之间的层级结构或依存关系,能够表示词语在语句结构中的位置以及与其他元素之间的关系;句法距离特征描述词语与句子核心词之间的距离关系,表达了词语与核心词的位置关系及对于语句结构的重要性:在依存关系的维度上,通常词语距离核心词越远,在句子中的作用也就越弱;句法数量特征指在一段

路径中某类元素的出现的次数,它描述了在句法结构中的特定部分元素的重要性程度。

本文基于短语结构和依存关系句法分析结果构造类型、路径、距离和数量4类句法特征,见表2。由于长句中的整句句法路径较长,容易造成数据稀疏现象,特征数量过多,标注语料内容有限,有效的特征无法集中,影响条件随机场的识别精度。为解决这个问题,本文借鉴自然语言处理的n-gram词元思想,构造了3-gram句法路径,即计算当前词语到第三层父节点之间的路径,并使用距离特征来描述词语与核心词之间的句法位置关系,以降低使用整句路径时的数据稀疏性。

表2 句法特征集

Tab.2 Syntactic Feature Set

分类	标记	特征名称	说明
类型特征	TP	短语类型	当前词语位于的短语结构类型
	TD	依存类型	当前词语与其支配词之间的依存类型
路径特征	PP	短语结构路径	当前词语到第三层父节点之间的路径
	PD	依存关系路径	当前词语到第三层父节点之间的路径
距离特征	DP	语法树距离	当前词语在语法树中的层级深度
	DD	依存距离	当前词语与核心词之间的依存关系数量
数量特征	NP	名词短语数量	从当前词语到语法树根部的名词短语数量

3 基于复合特征的中文地名识别实验

基于条件随机场的中文地名识别系统主要由5个模块组成:①数据处理;②特征生成;③特征选择;④参数训练;⑤地名识别,其中特征模板包括地名要素、词性和句法3类特征。系统整体结构如图2所示。

3.1 不同特征组合的中文地名识别对比

为了评价顾及句法特征条件随机场的中文地名识别性能,本文使用CRF++作为条件随机场工具,设计了基于条件随机场模型的中文地名识别试验,分别测试了使用复合特征中不同特征组合时的识别情况。

传统的地名识别方法只考虑了词法特征,在标注语料库时只需有地名短语即可,可以从地名黄页等数据源收集语料[29]。复合特征包括了地名要素、词性和句法特征,描述中文语句中的各类信息,因此需要收集包含地名短语的整条句子内

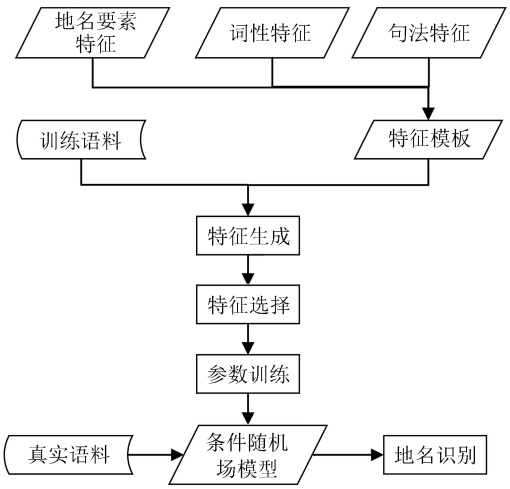

图2 基于复合特征的中文地名识别流程

Fig.2 Process of Chinese Place Name Recognition Based on Composite Features

容,并使用中文分词和句法分析工具进行处理。本文从互联网新闻中收集整理了1 282条包含中文地名的句子,使用ICTCLAS作为中文分词工具,使用Stanford Parser作为句法分析器,构造词语的短语结构和依存关系中的4类句法特征,并对地名进行标注。随机选择800条语句作为训练语料库,剩余的482条作为测试语料库,设置上下文窗口为{-5,+5},分别测试了添加不同特征后的中文地名识别结果。测试结果采用自然语言处理领域的三大评测指标,即准确率(P)、召回率(R)和综合值(F):

$$P=\frac{\text{正确识别的实体个数}}{\text{识别出的实体总数}}\times 100\%$$

$$R=\frac{\text{正确识别的实体个数}}{\text{文档中的实体总数}}\times 100\%$$

$$F=\frac{2\times P\times R}{P+R}\times 100\%$$

不同特征组合条件随机场中文地名识别的准确率、召回率和综合值见表3。

由表3可知:(1)在地名要素特征和词性特征的基础上增加句法特征后,中文地名识别效果的准确率和召回率都有明显的提升,说明句法特征能够有效提高中文地名识别的效果;(2)4类句法特征中路径特征对于中文地名识别效果的提高最为突出,增加了路径特征后,同基准测试中相比,准确率提高了5.98%,召回率提高了6.53%;(3)由于各类句法特征所表达的对象特点不同,不同特征组合会对中文地名识别的效果产生不同的影响,本次实验中"GP+POS+TP+TD+PP+PD+DP+DD"特征组合的准确率最高,为93.03%;"GP+POS+TP+TD+PP+PD+DP"特征组合

表 3　中文地名识别实验结果/%

Tab.3　Result of Chinese Place Name Recognition Experiment/%

特征组合	准确率	召回率	综合值
GP+POS	87.05	84.45	85.73
GP+POS+TP	88.07	85.71	86.87
GP+POS+TP+TD	88.13	85.65	86.87
GP+POS+TP+TD+PP	91.08	90.56	90.82
GP+POS+TP+TD+PP+PD	91.17	90.37	90.76
GP+POS+TP+TD+PP+PD+DP	92.99	90.98	91.97
GP+POS+TP+TD+PP+PD+DP+DD	93.03	90.88	91.94
GP+POS+TP+TD+PP+PD+DP+DD+NP	92.52	91.02	91.76

的召回率和综合值最高,为 90.98%、91.91%。因此在进行基于条件随机场的命名实体识别时,应充分扩展、挖掘不同类型的命名实体特征,并分析、验证各类特征之间的关系,选取组合效果最好的特征模板。

3.2　基于复合特征的简单地名与复杂地名识别对比

中文地名中既有简单的单词地名,又有多个词语组成的复杂地名,在地名识别中,简单地名的识别较为简单,复杂地名的识别是影响地名识别精度的关键问题。为了测试句法特征对于复杂地名识别的效果,本文将标注的 1 282 条地名数据进行分类整理,共得到简单地名 721 条,复杂地名 561 条,分别对两类语料库进行地名识别测试,识别结果如表 4 所示。

表 4　简单与复杂地名识别实验结果/%

Tab.4　Result of Simple and Complex Place Name Recognition Experiment/%

特征组合	简单地名			复杂地名		
	准确率	召回率	综合值	准确率	召回率	综合值
GP+POS	95.31	92.76	94.02	76.43	73.78	75.08
GP+POS+TP	95.55	94.58	95.06	78.45	74.32	76.33
GP+POS+TP+TD	95.57	94.55	95.06	78.57	74.21	76.33
GP+POS+TP+TD+PP	96.24	95.12	95.68	84.45	84.71	84.58
GP+POS+TP+TD+PP+PD	96.29	95.07	95.68	84.58	84.32	84.45
GP+POS+TP+TD+PP+PD+DP	97.44	95.82	96.62	87.28	84.75	86.00
GP+POS+TP+TD+PP+PD+DP+DD	97.46	95.58	96.51	87.34	84.83	86.07
GP+POS+TP+TD+PP+PD+DP+DD+NP	97.37	95.43	96.39	86.29	85.35	85.82

通过对简单和复杂地名识别的对比测试可知,简单地名的识别所需特征较少,利用地名要素特征+词性特征即可实现较高的精度,增加其他特征后识别精度有所提高提高但并不显著,准确率最高提高了 2.15%,召回率最高提高了3.06%;对于复杂地名,简单的地名要素和词性特征组合效果较差,使用复合特征能够很好地提高识别效果,实验表明,增加句法特征后,准确率最高提高了 9.91%,召回率最高提高了 11.57%。

4　结　语

针对当前中文地名识别特征类型单一的问题,本文提出了基于复合特征的中文地名识别方法,设计了类型、路径、距离和数量四类句法,使用地名要素特征、词性特征、句法特征进行条件随机场的训练和识别。实验表明,复合特征能够有效提高基于条件随机场的中文地名识别效果,对于复杂地名识别的效果提高显著。

句法特征能够很好地表征完整语句中的地名信息,但在网络文本尤其是互联网新媒体数据中还存在很多不规范或不完整的句子。对于这类非常规的语句,应该结合上下文或语义关系来进行判断。本文的下一步工作是将语义关系特征引入到条件随机场中,实现基于词法、句法、语义的三级复合特征的中文地名识别。

参　考　文　献

[1] Zhang Xueying, Zhang Chunju, Lv Guonian. Design and Analysis of a Classification Scheme of Geographical Named Entities[J]. *Journal of Geo-information Science*, 2010(02): 2 220-2 227(张雪英,张春菊,闾国年. 地理命名实体分类体系的设计与应用分析[J]. 地球信息科学学报,2010(02): 2 220-2 227)

[2] Li Yusen, Zhang Xueying, Yuan Zhengwu. Study on Geographical Entity Recognition in GIS[J]. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, 2008, 20(6): 719-726(李玉森,张雪英,袁正午. 面向 GIS 的地理命名实体识别研究[J]. 重庆邮电大学学报(自然科学版),2008, 20(6): 719-726)

[3] Cheng Changxiu, Yu Bin. A Rule-Based Segmenting

and Matching Method for Fuzzy Chinese Addresses[J]. *Geography and Geo-Information Science*, 2011(03): 26-29(程昌秀,于滨. 一种基于规则的模糊中文地址分词匹配方法[J]. 地理与地理信息科学,2011(03): 26-29)

[4] Zhang Xueying, Lv Guonian, Li Boqiu. Rule-based Approach to Semantic Resolution of Chinese Addresses[J]. *Journal of Geo-information Science*, 2010(01): 9-16(张雪英,闾国年,李伯秋. 基于规则的中文地址要素解析方法[J]. 地球信息科学学报, 2010(01): 9-16)

[5] Tan Kankan. Rule-based Chinese Address Segmentation and Matching Methods[D]. Qingdao: Shandong University of Science and Technology, 2011(谭侃侃. 基于规则的中文地址分词与匹配方法[D]. 青岛: 山东科技大学, 2011)

[6] Du Ping, Liu Yong. Recognition of Chinese Place Names Based on Ontology[J]. *Journal of Northwest Normal University (Natural Science)*, 2011, 47(6): 87-93(杜萍,刘勇. 基于本体的中文地名识别[J]. 西北师范大学学报(自然科学版), 2011, 47(6): 87-93)

[7] Qiu Sha, A. Yuan, Wang Fuyan, et al. Study on Automatic Recognition of Chinese Location Names Based on Statistical Method[J]. *Computer Technology and Development*, 2011, 21(11): 35-38(邱莎,阿圆,王付艳,等. 基于统计的中文地名自动识别研究[J]. 计算机技术与发展, 2011, 21(11): 35-38)

[8] Tang Xuri, Chen Xiaohe, Zhang Xueying. Research on Toponym Resolution in Chinese Text[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(08): 930-935(唐旭日,陈小荷,张雪英. 中文文本的地名解析方法研究[J]. 武汉大学学报·信息科学版, 2010, 35(08): 930-935)

[9] Aaron L F H, Derek F W, Lidia S C. Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics[M]. LP&IIS2013, Warsaw: Springer, 2013

[10] Chen Wenliang, Zhang Yujie, Hitoshi Isahara. Chinese Named Entity Recognition with Conditional Random Fields[C]. The Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 2006

[11] Gao Lingling. A Study on Chinese Syntax Analysis Based on Dependency Grammer [D]. Qingdao: Ocean University of China, 2009(高玲玲. 基于依存语法的汉语句法分析研究[D]. 青岛: 中国海洋大学, 2009)

[12] Yin Dechun. Chinese Syntactic Parsing Based on Linguistic Entity Relationship Model[D]. BeiJing: Beijing Institute of Technology, 2014(尹德春. 基于语言实体关系模型的汉语句法分析[D]. 北京: 北京理工大学, 2014)

[13] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. The 18th ICML, San Francisco, USA, 2001

[14] Li Zhenghua. Research on Key Technologies of Chinese Dependency Parsing[D]. Harbin: Harbin Institute of Technology, 2013(李正华. 汉语依存句法分析关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2013)

[15] Yu Shiwen, Duan Huiming, Zhu Xuefeng, et al. The Basic Processing of Contemporary Chinese Corpus at Peking University[J]. *Journal of Chinese Information Processing*, 2002(05): 49-64(俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002(05): 49-64)

[16] Benajiba Y, Zitouni I, Diab M, et al. Arabic Named Entity Recognition: Using Features Extracted from Noisy Data[C]. The ACL 2010 Conference Short Papers, Uppsala, Sweden, 2010

[17] Laokulrat N, Miwa M, Tsuruoka Y, et al. Uttime: Temporal Relation Classification Using Deep Syntactic Features[C]. Second Joint Conference on Lexical and Computational Semantics, Atlanta, USA, 2013

[18] Arisoy E, Saraclar M, Roark B, et al. Syntactic and Sub-lexical Features for Turkish Discriminative Language Models[C]. Acoustics Speech and Signal Processing, Dallas, USA, 2010

[19] Hancke J, Vajjala S, Meurers D. Readability Classification for German Using Lexical, Syntactic, and Morphological Features[C]. 24th International Conference on Computational Linguistics, Mumbai, India, 2012

[20] Bykh S, Meurers D. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization[C]. 25th International Conference on Computational Linguistics, Dublin, Ireland, 2014

[21] Dai Min, Wang Rongyang, Li Shoushan, et al. Opinion Target Extraction with Syntactic Features[J]. *Journal of Chinese Information Processing*, 2014, 28(04): 92-97(戴敏,王荣洋,李寿山,等. 基于句法特征的评价对象抽取方法研究[J]. 中文信息学报, 2014, 28(04): 92-97)

[22] Guo Xiyue, He Tingting, Hu Xiaohua, et al. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features[J]. *Journal of Chinese Information Processing*, 2014, 28(6): 183-189(郭喜跃,何婷婷,胡小华,等. 基于句法语义特征的中

文实体关系抽取[J]. 中文信息学报,2014, 28(6): 183-189)

[23] Xu Bing, Zhao Tiejun, Wang Shanyu, et al. Extraction of Opinion Targets Based on Shallow Parsing Features[J]. *Acta Automatica Sinica*, 2011, 37(10): 1 241-1 247(徐冰,赵铁军,王山雨,等. 基于浅层句法特征的评价对象抽取研究[J]. 自动化学报,2011, 37(10): 1 241-1 247)

[24] Mukherjee S, Tiwari A, Gupta M, et al. Shallow Discourse Parsing with Syntactic and (a Few) Semantic Features[C]. The Nineteenth Conference on Computational Natural Language Learning, Beijing, China, 2015

[25] Johansson R, Moschitti A. Syntactic and Semantic Structure for Opinion Expression Detection[C]. The Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden, 2010

[26] Stein D, Peitz S, Vilar D, et al. A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation[C]. Conference of the Association for Machine Translation, Denver, USA, 2010

[27] Loni B, Van T G, Wiggers P, et al. Question Classification by Weighted Combination of Lexical, Syntactic and Semantic Features[M]. Text, Speech and Dialogue, Pilsen, Czech: Springer, 2011

[28] Grundström J, Nugues P. Using Syntactic Features in Answer Reranking[C]. AAAI 2014 Workshop on Cognitive Computing for Augmented Human Intelligence, Québec, Canada, 2014

[29] Jiang Wenming, Zhang Xueying, Li Boqiu. CRFs-based Approach to Recognition of Chinese Address Element[J]. *Computer Engineering and Application*, 2010, 46(13): 129-131(蒋文明,张雪英,李伯秋. 基于条件随机场的中文地址要素识别方法[J]. 计算机工程与应用,2010, 46(13): 129-131)

A Method of Chinese Place Name Recognition Based on Composite Features

WEI Yong[1]　*LI Hongfei*[1,2]　*HU Danlu*[2]　*LI Xiang*[2]　*MA Leilei*[3]

1　Troops 31008, Beijing 100091, China

2　Institute of Geographical Spatial Information, Information Engineering University, Zhengzhou 450001, China

3　Troops 95291, Hengyang 421010, China

Abstract: Chinese place name recognition is a research topic in named entity recognition, and a key to improve the application level of the geographic information systems in China. The traditional place name recognition method is based on the element characteristics of a place name and the part of speech of words, and employs limited features. This paper proposes a method of Chinese place name recognition method using syntactic features, and mines the syntactic characteristics of place names in natural language. The design employs four syntactic features, class, path, distance, and number, in conditional random fields (CRF) to train and recognize Chinese place names based on place name elements, position of speech (POS) and syntactic features. Comparative experiments with composite features and traditional features for Chinese place name show that with the help of the three composite features, Chinese place name recognition accuracy and recall rate can be improved effectively and with good results for complex place names.

Key words: place name recognition; composite features; syntactic parsing; conditional random field

First author: WEI Yong, PhD, specializes in web geospatial data acquisition and information extraction. E-mail: whuwy@163.com

Corresponding author: HU Danlu, PhD, professor. E-mail: hudanlu@vip.sina.com

Foundation support: The National Natural Science Foundation of China, No. 41401467; the Open Research Fund by Sichuan Engineering Research Center for Emergency Mapping & Disaster Reduction, No. K2015B014.