CRF 与规则相结合的维吾尔文地名识别研究*

买合木提·买买提[1,2]，卡哈尔江·阿比的热西提[1,2]，艾山·吾买尔[1,2]，吐尔根·依布拉音[1,2]，王路路[1,2]

（1.新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；2.新疆多语种信息技术重点实验室，新疆 乌鲁木齐 830046）

摘要：通过维吾尔文地名的分析研究，提出了一种基于条件随机场和规则的维吾尔文地名识别方法。根据维吾尔文地名黏着性、音译等特点，针对维吾尔文地名识别任务，在词汇和词性特征基础之上，引入音节、词向量获取的相似单词、常用地名词典、地名特征词、地名词缀等特征进行实验，结果表明这些特征对识别性能有较大的影响。通过对错误识别结果分析，提出了基于规则的后处理，进一步提高了识别性能，准确率达到 94.68%，召回率达到 89.52%，F 值达到 92.03%。

关键词：命名实体；维吾尔文；地名；条件随机场；词向量

中图分类号：TP 391.2　　**文献标识码：**A

Recognition of Uyghur Location Names Based on Conditional Random Fields and Rules

Maihemuti Maimaiti[1,2], Kahaerjiang Abiderexiti[1,2], Aishan Wumaier[1,2], Tuergen Yibulayin[1,2], Wang Lu-lu[1,2]

(1. School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China; 2. Xinjiang Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China)

Abstract: By the analysis of Uyghur location names, a conditional random fields and rule based Uyghur location name recognition method was proposed. Except word and part of speech features, morphological and contextual features such as syllable, word embedding based similar words, common gazetteer, characteristic word, the common affixes of location names are used according to agglutination and transliteration characteristics of Uyghur location names. The results show that these features have a greater impact on the recognition performance. Error recognition results are analyzed, and a rule-based post-processing method, which effectively improved the recognition performance, was put forward. The precision, recall and F-score of the system reached 94.68%, 89.52% and 92.03% respectively.

Key words: named entity; Uyghur; location name; CRFs; Word Embedding

0 引言

随着互联网技术的迅速发展，各类信息聚增，网上每天都有海量信息在生成，存储和传播，人类面临前所未有的信息膨胀。如何从海量信息中快速寻找，并抽取所需信息是当今信息处理领域面临的一个重要问题，其中命名实体识别是信息抽取的重要部分。因为命名实体识别的性能，对句法分析、语义分析、关系抽取等具有极其重要的影响。

命名实体(Named Entity，NE)是文本信息中的基本单位，是文本中的固有名称、缩写及其他唯一标识，是正确理解文本的基础[1]。狭义上可把命名实体分为人名、 地名、组织名等。广义上命名实体包括时间表达式，数值表达式等，在不同的应用领域，还可以根据具体的需要定义其他类型的命名实体，例如，在某个具体应用中，可能需要把住址、电子信箱、电话号码、会议名称等作为命名实体。

目前命名实体识别方法分为三种：基于规则的方法[2],基于统计的方法[3]以及基于神经网络的方法[4]。基于规则的命名实体识别的基本思路是人工编写上下文敏感的产生式，使用普通的 NE 数据库，都将不同的权值赋给不同的规则以便在产生规则冲突时可以选择具有最大权值的规则。基于统计的方法将专名识别看作一

*收稿日期：　　定稿日期：

基金项目：国家自然科学基金（61462083，61262060，61331011，61463048），国家重点基础研究发展计划（973）资助项目（2014CB340506），新疆多语种信息技术实验室开放课题（2016D03023）。

作者简介：买合木提·买买提（1980-），男，博士研究生，主要研究方向：自然语言处理及机器翻译；†通讯作者：艾山·吾买尔（1982-),男,副教授,硕士生导师,主要研究方向：自然语言处理及机器翻译.hasan1479@xju.edu.cn

般模式识别中分类问题的一个特例，利用字标注的方法来进行命名实体识别。其基本步骤包括:特征选择、机器学习、标注、后处理。基于深度学习的方法通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示，需要比较大的数据规模。

目前维吾尔文命名实体识别研究处于起步阶段，国内有关维吾尔文命名实体识别主要集中在人名识别[5-8]，时间表达式识别与抽取[9]，基于规则的机构名[10]、地名识别[11]方面。未有见到使用统计的方法研究维吾尔文地名识别有关的报道。

维吾尔文地名识别具有独特的词法、语言特点，所以直接套用英语和汉语的方法并不合适。目前，还没有公开的维吾尔语地名标注语料，因此，本文通过人工标注建立了 1.3 余万句子的维吾尔文地名标注语料库。在深入分析维吾尔文地名语法和语义特征的基础上，鉴于条件随机场在序列标注任务中的优异表现，首次使用条件随机场模型实现了维吾尔文地名自动识别方法。在特征模板的设计上，我们使用词、音节、词性标注、分布式向量表示[12]等不同特征，分析了它们对地名识别的影响。实验结果表明，我们的方法在测试数据上的 F 值达到 92.03%。

1 维吾尔文地名特点及识别难点

维吾尔文地名像中文、英文地名一样，具有数量庞大，音译地名较多，地名用词比较自由，地名词长没有限制，多词地名结尾经常有地名特征词出现，不同词性的词（比如形容词、人名、实物名、方位词、连词等）经常出现在多词地名的首词或中间词位置等共同特点，如表 1：

表 1 维吾尔文地名与其他语言地名的共有特点

中文	特点类型	维吾尔文地名举例
北京	单词地名	بېيجىڭ
海西蒙古族藏族哈萨克族自治州	多词地名	خەيشى موڭغۇل - زاڭزۇ - قازاق ئاپتونوم ئوبلاستى
北京市	特征词（市、地区、江河..）	بېيجىڭ **شەھىرى**
和田地区		خوتەن **ۋىلايىتى**
黄河		خۇاڭخې **دەرياسى**
英吾斯唐乡	形容词	**يېڭى** ئۆستەڭ كەنتى
红海	形容词	**قىزىل** دېڭىز
东南亚	方位词	**شەرقى جەنۇبىي** ئاسىيا
阿合买提江路	人名	**ئەخمەتجان** يولى
圣多美和普林西比民主共和国	连词	سان تومې **ۋە** پرىنسىپې دېموكراتىك **جۇمھۇرىيىتى**

除共有特点之外，维吾尔文地名也有以下比较独特特点也是维吾尔文地名识别所面临的挑战：

1） 维吾尔语的黏着性、元音弱化等特性导致数据稀疏：地名可以连接名词格词缀，维吾尔语名词格词缀有 24 个，这样一个地名可能会出现 24 中形态变化，降低词汇重复率，导致统计模型中的未登录词问题。如表 2：

表 2 维吾尔语地名黏着性示例（单词地名）

维吾尔语地名	**汉语译文**
شىنجاڭ	新疆（词干形式）
شىنجاڭنىڭ	新疆的
شىنجاڭدىكى	在新疆的
شىنجاڭدىن	从新疆
شىنجاڭغا	（向，给）新疆
شىنجاڭلىق	新疆人
شىنجاڭدا	在新疆

值得注意的是，大多数情况下，多词地名的最后一个词（或特征词）才会连接附加成分，该地名中的其余词语一般不会连接附加成分。如表 3 所示：

表 3 维吾尔语地名黏着性示例（多词地名）

维吾尔语地名	汉语译文
غالبىيەت يولىنىڭ	胜利路的
ئۈرۈمچى شەھىرىدىكى	在乌鲁木齐市的
شىنجاڭ ئۇيغۇر ئاپتونوم رايونىمىزدا	在新疆维吾尔自治区
ئاسىيا قىتئەسىگە	（向，往）亚洲

维吾尔语地名中的部分单词地名，多词地名中的部分特征词在连接附加成分时将会出现元音弱化现象，使数据进一步稀疏。比如：«ئۆلكە،شەھەر،ۋىلايەت،بازار، يېزا»（省，市，地区，镇，农村）等常见地名特征词连接后缀时，经常出现元音弱化。如表 4：

表 4 维吾尔语地名元音弱化示例

维吾尔语地名	汉语译文
ئامېرىكا(ئامېرىكىنىڭ،ئامېرىكىدىكى،ئامېرىكىنى)	单词地名美国，加附加成分后«ا»弱化为«ى»
تۇرپان شەھەر(تۇرپان شەھەرلىك،تۇرپان شەھىرى،تۇرپان شەھىرىدىكى)	吐鲁番市，加附加成分后«ە»弱化为«ى» ，有时还不弱化

2）由于地名词长一般没有严格的限制，单词地名又可以作为多词地名的一部分出现，多词地名又可以根据文本中的上下文描述需要，忽略其中的中间词或特征词或者只用特证词代替整个地名来使用其简称，再加上黏着性，使得数据更加稀疏，导致识别更加困难：如下例子：

ئىلى قازاق ئاپتونوم ئوبلاستى (伊犁哈萨克自治州) شىنجاڭ ئۇيغۇر ئاپتونوم رايونىنىڭ (新疆维吾尔自治区) غەربىي شىمال قىسمىغا جايلاشقان بولۇپ ، ئاپتونوم رايونىمىزدىكى (我区或自治区) قازاق مىللىتى توپلىشىپ ئولتۇراقلاشقان ئوبلاست. ئىلى ئوبلاستىنىڭ (伊犁州) مەركىزى غۇلجا شەھىرى. ئىلى (伊犁) خەلقى خۇشچاقچاق، مېھماندوست، ئافكۆڭۈل خەلق.

3）新涌现的地名大多数以音译为主，相对维吾尔文中的自然地名，这类外来词地名不受到严格的维吾尔语拼写规则的限制，部分中间词和特征词也有这种情况出现，从而往往会导致以下两种情况：一是字母连接不同寻常，音节特殊；二是经常出现拼写错误。如表 5：

表5 维吾尔语外来音译地名示例

维吾尔语地名	汉语译文
سىچۇەن،ۋېنېسۇئېلا	四川, 委内瑞拉（两个元音一起出现）
جۇڭگو(جوڭگو،جۆڭگۆ، جۇڭگۇ)	中国（汉语音译，经常被错误拼写）
ئامېرىكا(ئامىرىكا،ئامرىكا،ئامىركا)	美国（英语音译，经常被错误拼写）
ئاپتونوم(ئاپتۇنۇم،ئاپتونۇم،ئاپتۇنوم،ئاپتنۇم،ئاپيونوم،ئاپتونم،ئاپتنوم)	自治（中间词，经常被错误拼写）

4）部分单词地名具有共同的特征词缀。如表 6：

表 6 维吾尔文共同词缀地名示例

特征词缀	维吾尔文地名举例
ىيە	ياپونىيە، كورىيە، ئەنگلىيە، روسىيە، گېرمانىيە، ئاۋسترالىيە، فرانسىيە... 译文：（日本，韩国，英国，俄罗斯，德国，法国，澳大利亚）
ىستان	قازاقستان، تاجىكستان، قىرغىزستان، ئۆزبېكستان، تۈركمەنستان، پاكستان... 译文：（哈萨克斯坦，塔吉克斯坦，吉尔吉斯斯坦，乌兹别克斯坦，土库曼斯坦，巴基斯坦）
باغ،تاغ	چىلانباغ، نەزەرباغ ، شامالباغ، سايباغ، تەڭرىتاغ، مايتاغ...

	译文（其兰巴格，乃扎尔巴格，夏马勒巴格，（萨，沙）依巴格，天山，独山子）
شەن،سەن	چىەنشەن ،ۋۇلياڭشەن،چىلىەنشەن ، پىڭدىڭسەن ، يۇياسەن ، خۇاڭسەن... 译文（千山，祁连山，无量山，平顶山，月牙山，黄山）
خې،جياك	خۇاڭخې،بەيياڭخې،سەنتۇنخې ، چاڭجياك،سۇڭخۇاجياك... 译文（黄河，白杨河，三屯河，长江，松花江）

2 基于条件随机场的维吾尔语地名识别

2.1 条件随机场模型

条件随机场模型是在给定输入节点条件下计算输出节点的条件概率的无向图模型，定义$\mathrm{W}=w_1w_2\cdots w_n$为给定的 n 个输入节点的值，比如一个句子。定义 O 为有限状态机的状态，$\mathrm{O}=o_1o_2\cdots o_N$为一个长度为 N 的输出节点的值。对于一个带有参数$\Theta=\theta_1\theta_2\cdots\theta_k$的线性链，将给定的序列 W 得到的状态序列的条件概率定义为：

$$P_\Theta(\mathrm{O}|\mathrm{W})=\frac{1}{Z_W}\{\sum_{n=1}^{N}\sum_k f_k(O_{n-1},O_n,\mathrm{W},n)\} \tag{1}$$

$$Z_\mathrm{W}=\sum_O exp\{\sum_{n=1}^{N}\sum_k f_k(O_{n-1},O_n,\mathrm{W},n)\} \tag{2}$$

其中Z_W是归一化参数，它使得给定输入的所有可能状态序列的概率之和为 1。$f_k(O_{n-1},O_n,\mathrm{W},\mathrm{n})$是对于整个观察序列 W，标记位于 N 和 N-1 的特征函数，特征函数可以是 0,1 值，也可以是任意实数。$\Theta=\theta_1\theta_2\cdots\theta_k$是特征函数对应的权重。对于 W 来说，要做的是搜索概率最大的$\mathrm{O}^*=\arg\max P(\mathrm{W}|\mathrm{O})$。

2.2 分布式向量表示

在自然语言处理过程（NLP）中，首先需要考虑的问题是如何对自然语言进行建模，使得计算机可以处理自燃语言。目前常用的表示方法主要有两种：“one-hot”表示和分布式向量表示(word embedding)。分布式向量表示是 Mikolov 在 2013 年提出[12]。

目前 word2vec 已经在众多 NLP 领域得到应用。如词性标注[13]、命名实体识别[14]、聚类和情感分析[15]等。Word2vec 内置两种训练模型，CBOW 模型和 Skip-gram 模型。CBOW 模型是根据输入当前词的上下文来预测当前词，如图 1 所示。Skip-gram 模型是根据输入的当前词去预测当前词的上下文，如图 2 所示。

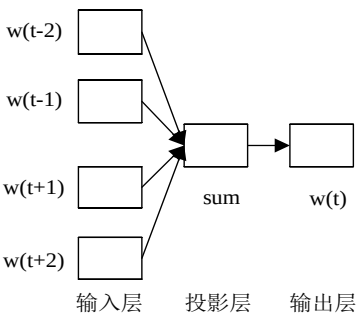

图 1 CBOW 模型结构

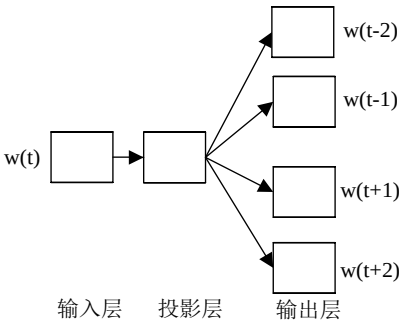

图 2 Skip-gram 模型结构

2.3 特征描述

CRFs 模型的性能取决于特征，因此我们根据命名实体识别领域常用特征，增加了 Word Embedding 特征，我们以容易获得为主要特征选择原则，所选择的特征如下：

1） **单词特征（F_w）。**即 tokens 特征，我们分别选择当前 token 及其前后的一个 token（F_{w1}），两个 token（F_{w2}）和三个 token（F_{w3}）来分析了 tokens 的不同窗口大小对地名识别的影响，并确认了最佳窗口大小为 5（F_{w2}）。

2） **音节特征（F_s）。**从第一章中描述的维吾尔语地名特点可知，大多数地名词语的最后一个或两个音节对识别该地名的影响比较大，因此我们从单词后缀部分摘取一个或两个音节作为特征。当单词音节数小于等于 2 时，取一个音节；当单词音节数大于 2 时，取两个音节。单词特征确认后，我们又分别选择当前 token 及其前后的一个 token 的音节（F_{s1}），两个 token 的音节（F_{s2}）和三个 token 的音节（F_{s3}）来分析了音节的不同窗口大小对地名识别的影响，并确认了最佳窗口大小为 3（F_{s1}）。

3） **词性特征（F_p）。**单词特征确认后，我们又分别选择当前 token 及其前后的一个 token 的词性（F_{p1}）、两个 token 的词性（F_{p2}）和三个 token 的词性（F_{p3}）来分析了词性的不同窗口大小对地名识别的影响，并确认了最佳窗口大小为 3（F_{p1}）。

4） **Word Embedding 特征（F_{ew}）。**我们在词向量特征的选择上，分别选择了与当前 token 最相似的一个词（F_{ew1}）、两个词（F_{ew2}）、三个词（F_{ew3}）、四个词（F_{ew4}）和五个词（F_{ew5}）来进行实验分析。选择相似词的时候，分别通过 CBOW 和 Skip-gram 获取的词加在一起，根据相似度进行排序，然后提取了相似度最高的前 5 个词。Word Embedding 特征的窗口大小设置为 3，即只考虑当前词的前后词。

5） **词典特征（F_{dic}）：**根据地名特点，我们主要建立了三种词典，分别是：常用地名词典（F_{loc}）、地名特征词词典（F_{fw}）以及地名共同词缀词典（F_{sfx}）。

常用地名词典（F_{loc}）：该词典收录了世界各国家及其主要城市名、中国省份及主要城市等，共 3013 个。当前词在常用地名词典内则其特征值为 D-Loc，否则为 D-No。

地名特征词特征（F_{fw}）：由于双词或多词地名结尾经常带有地名特征词，如“شەھرى”、“يولى”、“دەرياسى”等，因此它们是地名非常重要的结构特征。本文对《新疆地名大词典》中的多词地名进行分析，收录这些常见的地名尾词作为地名特征词表，共 121 个，其示例见表 3。根据维吾尔语语法和对这些特征词分析发现，维吾尔文特征词作为地名一部分出现时往往采用第三人称单数形式，因此构造特征词表时我们也考虑了此特点。特征生成时，如果当前词在特征词表里，则对应的特征值为 F-Y，否则为 F-N。另外，考虑到当前特征词在真实语料中可能附加其他附加成分，如“شەھرىنىڭ”、“يولىغا”、“دەرياسىدىكى”等，我们在特征词词表匹配时选择了模糊匹配方法，从而保证更多的特征词能够正确匹配。

地名词缀特征（F_{sfx}）：从表 6 可知，维吾尔语有部分地名具有共同词缀，因此我们收集了经常作为地名词缀的字符序列作为地名词缀特征表，共 29 个，其示例见表 6。特征生成时，如果当前词词缀在词缀特征表里，则对应特征值为 SFX-Y，否则为 SFX-N。

实验的所有数据均使用新疆多语种信息技术实验室自然语言处理组维吾尔语自然语言处理工具包（网络服务）[†]进行分词、分音节、词性标注处理。使用 400 万句单语语料进行词向量训练，工具使用 Txt2Vec[‡]，分别基于 Skip-gram 模型和 CBOW 模型训练 tokens 的 Embedding。本实验设置窗口的大小为 5，Embedding 的维度为 200。

数据通过处理之后，每行有 12 列，包括词、音节、词性标注、相似词、地名词典特征、地名特征词词典特征、地名词缀特征以及地名标注符号等。其中地名标注符号列的标记有 3 种（使用 IOB2[16]标记），分别是：O：非地名标记，B-Location：地名首词标记，I-Location：地名非首词标记。经过这种标记标注后，一个单独的 B-Location 标记表示单词地名。B-Location+I-Location+ [I-Location]…标记表示双词或多词地名。

2.4 基于规则的后处理

[†] http://202.201.255.248:8088/xjuapi/uyghurtext/

[‡] https://github.com/zhongkaifu/Txt2Vec

基于规则的方法可以有效的弥补机器学习不能表达语言的确定性的缺点，因此为了进一步提高识别性能，我们通过分析 CRF 对真实语料错误识别的示例，归纳了修正规则。下面是部分规则的描述：

规则 1：当前词是地名或在常用地名词典中，如果它前面的一个或两个词是（شەرقي،شىمالي،غەربي،جەنۇبي،ئوتتۇرا）等词中的任意一个词，则这两个或三个词是一个多词地名，如果它后面的词又在地名特征词词典中，则这三个或四个词是一个多词地名。比如：شەرقي ئاسىيا،جەنۇبي شىنجاك。

规则 2：当前词是（شەرقي،شىمالي،غەربي،جەنۇبي،ئوتتۇرا）等词中的任意一个词，它前面的词是地名或在常用地名词典中，它后面的词在地名特征词词典中，则这三个词是一个多词地名。比如：قەشقەر جەنۇبي يولى 。

规则 3：当前词被标注为地名，并其前后有“‹”符号，则该词后面的“‹”符号后的词识别为一个单词地名。

规则 4：当前词被标注为地名并且该词在地名词典中出现，则检查该词是否在地名歧义词典中，如果是那么识别该词为不是地名，否则识别为一个地名。比如：قاتارلىق،ئالىملار،ئاراللارنىك。

3 实验

3.1 实验数据集及评价方法

由于目前尚没有公开的地名标注数据，本文手工建立了一个维吾尔文地名标注语料库。我们根据汉语命名实体从新疆多语种信息技术实验室自然语言处理组汉维新闻对齐语料中随机抽取了 1.5 万条句子进行人工标注后，过滤掉其中没有地名或句子质量不高的句子，挑选了 13385 条句子。

数据集的主要情况如表 7 所示：

表 7 维吾尔地名标注语料库

数据集	句子数	Tokens	地名数	不重复地名
整个数据集	13385	428619	41008	20218
训练集	10704	342609	32885	17180
测试集	2681	86010	8123	5424

本文将采用准确率（P），召回率（R）和 F 值等三个指标来评价实验的性能，公式如下（3），（4），（5）。

$$准确率(P)=\frac{正确识别的地名个数}{识别出的所有地名个数}\times 100\% \tag{3}$$

$$召回率(R)=\frac{正确识别的地名个数}{样本中所有地名个数}\times 100\% \tag{4}$$

$$F值=\frac{2\times P\times R}{P+R}\times 100\% \tag{5}$$

3.2 实验结果与分析

3.2.1 单词窗口大小对识别性能的影响

首先验证不同单词窗口大小对测试结果的影响，选取最优的单词窗口进行下一步的实验。在选取窗口大小时，分别选择窗口大小为 3（F_{w1}）、5（F_{w2}）、7（F_{w3}），进行了不同窗口大小实验，如表 8 所示。可以看到，F 值并未随着窗口大小的增加而呈现上升趋势，单词窗口大小为 5（F_{w2}）的时候性能最好，F 值为 81.86%。除此之外，该实验又证明，单词特征对地名识别非常重要，只根据单词特征也可以达到比较好的识别效率。

表 8 单词窗口大小对识别性能的影响

窗口大小	准确率/%	召回率/%	*F*值/%
F_{w1}	95.07	70.92	81.24
F_{w2}	**94.71**	**72.08**	**81.86**
F_{w3}	94.86	70.64	80.98

单词窗口大小确定后（F_{w2}），我们接着考察特征频率对识别性能的影响如表 9 所示。实验结果表明，随着特征频率的增加准确率逐步下降，反而召回率和 F 值比较明显的提高，当特征频率大于等于 2 的时候系统性能最佳。

表 9 特征频率对识别性能的影响

特征频率	准确率/%	召回率/%	*F*值/%
>=1	94.71	72.08	81.86
>=2	**93.02**	**76.61**	**84.02**
>=3	92.94	75.05	83.04
>=4	92.12	75.17	82.79

3.2.2 不同特征及其窗口大小对识别性能的影响

在 3.2.1 中确定单词特征为 F_{w2} 并特征频率大于等于 2 时性能最佳，因此下面在此实验基础上，分别加入音节特征（F_s），词性特征（F_p），word embedding 特征（F_{ew}）以及词典特征（F_{dic}）进行不同特征在不同窗口下的实验。

音节及其窗口大小对识别性能的影响如表 10 所示。从中可以看出，一方面，音节窗口的增加并未导致 F 值的上升，反而逐步下降，当音节窗口大小为 3（F_{s1}）时，系统的 F 值达到了最佳值 88.16%，其原因可能为当前词的最后音节或后缀受到前一个词的影响或它影响下一个词的音节或后缀；另一方面，无论音节窗口大小取为何值，加入音节特征后系统的性能比仅考虑单词特征时要好，F 值提高了 4.14%，这说明音节特征对地名识别具有重要的影响。

表 10 音节及其窗口大小对识别性能的影响

实验	准确率/%	召回率/%	*F*值/%
F_{w2}+ F_{s1}	**93.97**	**83.02**	**88.16**
F_{w2}+ F_{s2}	93.94	82.06	87.60
F_{w2}+ F_{s3}	93.88	80.97	86.95

词性及其窗口大小对识别性能的影响如表 11 所示。从中可以看出，一方面，词性窗口的增加并未导致 F 值的上升，反而在下降，在三种不同窗口下加入词性后的系统性能差别不是很明显，其原因可能为受到了词性标注准确率的影响；另一方面，无论词性窗口大小取为何值，加入词性特征后系统的性能比仅考虑单词特征时要好，当词性窗口大小为 3（F_{p1}）时 F 值最好，提高了 2.25%，这说明词性特征对地名识别也有重要的影响。

表 11 词性及其窗口大小对识别性能的影响

实验	准确率/%	召回率/%	*F*值/%
F_{w2}+ F_{p1}	**93.19**	**80.32**	**86.27**
F_{w2}+ F_{p2}	92.44	80.80	86.22
F_{w2}+ F_{p3}	92.68	79.93	85.84

通过词向量获取的与当前词最相似的前五个词分别加入到特征后的实验结果如表 12 所示。从实验结果可以看出，随着相似词数量的增加，F 值也逐步提高，其原因可能为训练词向量语料规模不够，从而对一些低频词的表示不太准确。无论选取的相似词数量多或者少，加入词表示特征后系统的性能比仅考虑单词特征时要好，最好时，F 值提高了 4.2%，这说明通过使用词表示方法来获取相似词作为特征引入可以提高识别性能。

表 12 基于 word embedding 的相似词对识别性能的影响

实验	准确率/%	召回率/%	*F*值/%
F_{w2}+ F_{ew1}	93.23	80.41	86.35
F_{w2}+ F_{ew2}	93.31	82.05	87.32
F_{w2}+ F_{ew3}	93.11	83.06	87.80

F_{w2}+ F_{ew4}	92.67	83.77	88.00
F_{w2}+ F_{ew5}	**92.76**	**84.09**	**88.22**

总结以上几个实验可以看出，音节、词性以及 word embedding 都对维吾尔语地名识别性能的提高起比较重要的作用，其中 word embedding 和音节的影响最大，F 值分别提高了 4.2%和 4.14%，其次为词性，F 值提高了 2.25%。当 word embedding 的相似词数量大于等于 2 的时候，音节特征和 word embedding 特征的影响不分上下，差别比较小，word embedding 特征的召回率比音节特征稍高，音节特征的准确率均高于 word embedding 特征，因此在没有音节切分工具或词性标注系统的情况下，可以考虑使用词表示来构造特征是可行的。当单词特征选取最佳的 F_{w2}（窗口大小为 5）时，词性和音节可以选取 F_{s1} 和 F_{p1}（窗口大小均为 3），基于 word embedding 的相似词取 F_{ew3}(3 个)及以上来达到最好的系统性能。

3.2.3 不同词典对识别性能的影响

不同词典及其组合对识别性能的影响如表 13 所示。实验结果表明，常用地名词典、特征词词典以及地名词缀词典都有助于提高系统的识别性能，其中常用地名词典的影响最大，其次为地名特征词词典，最后为地名词缀词典。地名词缀词典性能相对低的主要原因是因为，很多地名在文中往往有附加成分连接，此时无法有效提取地名共同词缀特征。三种词典相组合时系统性能最佳，F 值比仅考虑单词特征时提高了 4.52%。

表 13 不同词典及其组合对识别性能的影响

实验	准确率/%	召回率/%	*F* 值/%
F_{w2}+ F_{loc}	93.56	80.86	86.75
F_{w2}+ F_{fw}	92.67	80.43	86.11
F_{w2}+ F_{sfx}	92.91	79.10	85.45
F_{w2}+ F_{fw}+ F_{loc}	93.47	83.57	88.24
F_{w2}+ F_{dic}	**94.00**	**83.69**	**88.54**

3.2.4 不同特征组合对识别性能的影响

在 3.2.2 节实验的基础上，我们继续组合不同特征进行实验。鉴于 F_{w2}+ F_{s1} 的 F 值最好，我们在此组合基础上分别增加了词性特征和 word embedding 特征，实验结果如表 14 所示。从实验结果可知，分别增加这两种特征后，F 值均有提高，增加 word embedding 的性能优于增加词性特征。增加词性特征时，随着词性特征窗口的变大，系统性能反而下降；增加 word embedding 特征时，随着相似词数量的增加，系统性能并没有一直上升，甚至出现下降现象,但是总体来看呈上升趋势（特别是召回率），因此继续增加相似词数量可能有助于提高系统性能。

表 14 三种特征组合对比实验

实验	准确率/%	召回率/%	*F* 值/%
F_{w2}+ F_{s1}+ F_{p1}	93.53	84.19	88.62
F_{w2}+ F_{s1}+ F_{p2}	92.72	84.21	88.26
F_{w2}+ F_{s1}+ F_{p3}	92.55	83.76	87.94
F_{w2}+ F_{s1}+ F_{ew1}	94.12	84.55	89.08
F_{w2}+ F_{s1}+ F_{ew2}	94.12	85.08	89.37
F_{w2}+ F_{s1}+ F_{ew3}	94.16	85.50	89.62
F_{w2}+ F_{s1}+ F_{ew4}	93.89	85.52	89.51
F_{w2}+ F_{s1}+ F_{ew5}	**93.98**	**85.73**	**89.67**

接下来我们在上一步实验的基础上，对它的特征模板进行了扩充，即使用了一些混合特征，分别增加了当前词及其词性、当前词及其音节以及当前词及其前后词的音节和词性（F_{p12}）等 unigram 特征并与 word embedding 特征组合进行实验，结果如表 15 所示。四种特征组合后，随着 word embedding 的相似词数量的增加，F 值并没有一直呈上升趋势，反而性能差别非常小，由此可以看出四种特征组合进行实验时，相似词数

量对系统性能影响很小，这可能是因为音节特征和词性特征已经覆盖了大多数语言特征。四种特征组合实验中，当选择 $F_{w2}+ F_{s1}+ F_{p12}+ F_{ew5}$ 特征组合时，F 值达到了 90.15%，比上一步实验提高了 0.48%。

表 15 四种特征组合对比实验

实验	准确率/%	召回率/%	*F* 值/%
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew1}$	94.06	86.24	89.98
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew2}$	93.77	86.56	90.02
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew3}$	93.74	86.51	89.98
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew4}$	93.59	86.86	90.10
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew5}$	**93.88**	**86.70**	**90.15**

为了考察词典特征 F_{dic} 对识别性能的影响，我们对以上各类实验中性能最好的特征组合上增加词典特征分别进行了对比实验，结果如表 16 所示。实验结果表明，加入词典特征后，系统性能得到非常显著的提高，特别是召回率比准确率提高的比较明显。

表 16 词典特征对识别性能的影响

实验	准确率/%	召回率/%	*F* 值/%
$F_{w2}+ F_{s1}+ F_{dic}$	94.15	87.41	90.65
$F_{w2}+ F_{p1}+ F_{dic}$	93.58	86.18	89.73
$F_{w2}+F_{s1}+ F_{p12}+ F_{dic}$	94.09	88.22	91.06
$F_{w2}+F_{s1}+ F_{ew5}+ F_{dic}$	94.31	88.01	91.05
$F_{w2}+F_{s1}+ F_{p12}+ F_{ew5}+ F_{dic}$	**94.22**	**89.27**	**91.67**

3.2.5 基于规则的后处理对识别性能的影响

从上一节实验结果可知，基于 CRF 模型的维吾尔语地名识别可以得到非常不错的识别结果，但是结果中还存在一些漏掉的、识别错误的地名，因此我们对这些情况使用规则进行了修剪，从而进一步提高了系统性能，实验结果如表 17 所示。

表 17 基于规则的后处理后的实验结果

实验	准确率/%	召回率/%	F 值/%
$F_{w2}+ F_{s1}+ F_{dic}$	94.62	87.71	91.04
$F_{w2}+ F_{p1}+ F_{dic}$	94.02	86.45	90..07
$F_{w2}+ F_{s1}+ F_{p12}+ F_{dic}$	94.53	88.48	91.40
$F_{w2}+ F_{s1}+ F_{ew5}+ F_{dic}$	94.75	88.29	91.41
$F_{w2}+ F_{s1}+ F_{p12}+ F_{ew5}+ F_{dic}$	**94.68**	**89.52**	**92.03**

4 结论

本文利用地名人工标注语料和大规模单语语料数据，采用 CRFs 结合规则的方法对维吾尔文地名识别进行了研究，通过选取单词、音节、词性、基于词向量的相似词、词典等不同的特征及其组合，在不同的窗口大小下，对维吾尔文地名识别进行了实验。实验结果表明，单词特征、词性特征、音节特征对地名识别具有重要影响；同时引入词表示特征后，也可以达到比较高的系统性能，可以将比较难的维吾尔语词性特征由词表示特征来代替，从而可以减少识别工作的复杂度；再通过引入词典特征（常用地名词典、地名特征词词典、地名特征后缀词典等）和基于规则的后处理，有效的提高了识别性能。这说明本文所提出的不同特征窗口大小的选择、特征的组合方式、词表示特征的应用方法、词典特征以及基于规则的后处理对维吾尔语等资源匮乏、自然语言处理水平较低的语言具有一定的意义。

除了取得比较好的结果以外，本文还有一些局限性，例如：有些地名连接附加成分后会出现元音弱化现象，从而无法提取词典特征，容易出现漏识。在以后的研究中需要进一步改进。因此，下一步工作我们将本

文的成果与人工标注相结合，尝试采用深度学习方法对维吾尔文地名识别进行进一步研究；此外，我们将地名识别和其他命名实体任务，比如人名识别，机构名识别等，相结合进行多类命名实体识别研究。

参考文献：

[1] NADEAU D, SEKINE S. A survey of named entity recognition and classification [J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[2] MIKHEEV A, MOENS M, GROVER C. Named entity recognition without gazetteers[C]//proceedings of the Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1999, 1-8.

[3] 黄德根, 岳广玲, 杨元生. 基于统计的中文地名识别 [J]. 中文信息学报, 2003, (02): 36-41.

[4] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//proceedings of the arXiv preprint arXiv:160301360, 2016.

[5] 加日拉·买买提热衣木, 吐尔根·依布拉音, 艾山·吾买尔. 基于统计和规则混合策略的维吾尔人名识别研究[J]. 新疆大学学报(自然科学版), 2014, 31(03): 319-324.

[6] 艾斯卡尔·肉孜, 宗成庆, 姑丽加玛丽·麦麦提艾力,等. 基于条件随机场的维吾尔人名识别方法[J]. 清华大学学报(自然科学版), 2013(6):873-877.

[7] 热合木·马合木提, 于斯音·于苏普, 张家俊, 等. 基于模糊匹配与音字转换的维吾尔语人名识别[J]. 清华大学学报(自然科学版), 2017, (02): 188-196.

[8] 李佳正, 刘凯, 麦热哈巴·艾力, 等. 维吾尔语中汉族人名的识别及翻译[J]. 中文信息学报, 2011, 25(04): 82-87.

[9] 阿依古丽·哈力克, 艾山·吾买尔, 吐尔根·伊布拉音, 等. 汉维时间数字和量词的识别与翻译研究[J]. 中文信息学报, 2016, (06): 190-200.

[10] 麦合甫热提, 米日姑·肉孜, 麦热哈巴·艾力, 等. 基于语法语义知识的维吾尔文机构名识别[J].计算机工程与设计, 2014, 35(08): 2944-2948.

[11] 木合塔尔·艾尔肯, 艾斯卡尔·艾木都拉, 地里木拉提·吐尔逊. 基于规则的维吾尔地名识别[J]. 通信技术, 2013(7):103-105.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//proceedings of the Advances in neural information processing systems, 2013, 3111-3119.

[13] SANTOS C D, ZADROZNY B. Learning character-level representations for part-of-speech tagging[C]//proceedings of the Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, 1818-1826.

[14] DEMIR H, ÖZGÜR A. Improving named entity recognition for morphologically rich languages using word embeddings[C]//proceedings of the Machine Learning and Applications (ICMLA), 2014 13th International Conference on, IEEE, 2014, 117-122.

[15] TANG D, WEI F, YANG N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]//proceedings of the ACL (1), 2014, 1555-1565.

[16] Tjong K S E F, Buchholz S. Introduction to the CoNLL-2000 shared task: chunking[C]//The Workshop on Learning Language in Logic and the, Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2000:127-132.

作者联系方式：买合木提·买买提，新疆乌鲁木齐市天山区胜利路 666 号新疆大学信息科学与工程学院，830046，13579863941，mahmutjan@xju.edu.cn