文章编号：1003-0077(2011)04-0082-06

维吾尔语中汉族人名的识别及翻译

李佳正[1]，刘 凯[1]，麦热哈巴·艾力[1,2]，吕雅娟[1]，刘 群[1]，吐尔根·依布拉音[2]

(1. 中国科学院计算技术研究所 中国科学院智能信息处理重点实验室，北京 100190；
2. 新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046)

摘 要：该文研究了一种维吾尔语中汉族人名的识别和翻译方法。该方法在词典等传统方法的基础上，运用语言模型实现维语中的汉族人名的识别和翻译。针对维语人名的构词和拼写特点，增加了名词词缀识别预处理模块，补充了维语字母到汉语拼音的映射规则，有效提高了人名识别的正确率及召回率。在1 000句含有汉族人名的维语语料上进行测试，汉族人名识别的正确率和召回率分别达到75.2%和91.5%。

关键词：语言模型；名词词缀；拼写规则；人名识别及翻译

中图分类号：TP391 **文献标识码**：A

Recognition and Translation for Chinese Names in Uighur Language

LI Jiazheng[1], LIU Kai[1], Mairehaba·Aili[1,2], LV Yajuan[1], LIU Qun[1], Tuergen·Yibulayin[1,2]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China)

Abstract: Name translation in the minority languages is still in its infancy. This paper presents a method for recognizing and translating Chinese Names in Uighur Language. In addition to using the traditional rule approach, we use Uighur and Chinese language models to recognize and translate Chinese names in Uighur Language. On this basis, we add the appropriate rules and algorithms to solve the problem of names with noun affixes and incomplete rules. This improves the accuracy of translation and the recall rate. We test the translation system with 1 000 random sentences with Chinese names. The results show that the accuracy can reach 75.2% and the recall rate can reach 91.5%.

Key words: language model; noun affixes; spelling rules; recognition and translation of names

1 前言

我国是一个统一的多民族、多语言的国家，除汉族外，少数民族有55个，其中有语言80多种，40多种现行文字，其中22个民族使用28种文字。随着社会的发展，无论是经济还是文化，各民族之间的交流越来越频繁。而语言的差异，作为一个客观现实，在一定程度上制约了民族之间的文化交流、经济发展以及社会进步。与此同时，语言作为搭载民族文化的独特因素，对于民族文化的保护、研究和开发，以及弘扬本民族的传统文化，都起着十分重要的作用。因此，对少数民族语言与汉语之间机器翻译的研究是十分必要的。民族语言翻译不仅对少数民族和民族地区的经济、文化发展起作用，更在维护祖国统一，增强民族团结等方面发挥了不可替代的重要作用。

所谓的命名实体(Named Entity)主要包括实体(组织名、人名、地名)、时间表达式(时间、时间)、数字表达式(货币值、百分数)等。命名实体识别是对文本进行理解的前提工作，属于文本信息处理的基础研究领域，它的研究成果将对后续的一系列工作产生影响。人名识别在命名实体识别中是一个富有

收稿日期：2011-04-29 **定稿日期**：2011-05-21

基金项目：国家自然科学基金重点资助项目(60736014)；国家自然科学基金资助项目(60873167)

作者简介：李佳正(1988—)，女，硕士生，主要研究方向为自然语言处理技术；刘凯(1987—)，男，硕士生，主要研究方向为自然语言处理技术；麦热哈巴·艾力(1973—)，女，博士生，讲师，主要研究方向为自然语言处理和机器翻译。

万方数据

挑战的问题，它在英文中已经得到很好的研究。目前，人名识别的方法主要有基于规则的方法和基于机器学习的方法。孙茂松，宋柔等，采用基于规则的方法识别中国人名[1]；罗智勇，宋柔[2]从 10 万条人名库、2 亿字的真实语料库中将姓名用字分为了 9 类，并总结了 21 条识别规则。但是无论是收集规模巨大的人名库与真实语料库，还是提炼识别规则，都是一个费时费力的工程。随着技术的进步，利用统计方法进行人名识别成为主流。其中 HMM[3] 方法被认为是更容易捕捉局部的语言对象，成为众多研究者的选择，尤其是已用于已有的汉语命名实体识别系统中，如：张华平等[4]结合 Viterbi 算法实现角色的自动标注；吕雅娟[5]采用分解处理策略和动态规划方法识别中外人名和中国地名；Wu Youzheng[6]等提出了基于多特征相融合的汉语命名实体模型。

对于本文涉及到的维吾尔语人名的翻译，衣马木艾山·阿布都力克木在 2010 年提出了基于规则的维吾尔人名汉文机器翻译算法[7]。而有关维语中的汉族名字该如何翻译这个问题，基本没有相关研究工作。

本文提出一种维吾尔语中汉族人名识别及翻译方法。在普通人名的翻译上可以有很大的自主性和灵活性，但对于诸如国家领导人姓名这样特殊的姓名集合，则必须要求精准翻译。因此，有必要建立一部包括国家领导人、艺术家等名人的人名库。与此同时，在进行普通人名翻译的时候，姓名各个单字的词典也是必需的。在识别汉族人名的过程中，我们使用词典和拉丁维语及汉语的语言模型进行识别和翻译。此外，针对维语中人名可以缀接名词性后缀的特点以及拼写特点，我们添加了名词词缀识别预处理模块，补充了维语字母到汉语拼音的映射规则，有效提高了人名识别的正确率及召回率。

2　维吾尔语中汉族人名的特点

由于不同民族的历史、语言等方面的原因，维语人名与汉语人名有着一定的差异。汉语及维语人名都由两部分组成[8]。汉族人名有名字有姓氏，由姓和名两部分组成，有专用的姓；但是维语人名却没有专用的姓，采用父子连名制，用父名作姓，其全名由本名和父名组成。汉族及维族人名形式不同。汉族人名姓在前名在后，即姓＋名，如“张伟”；维吾尔族人名排列次序恰好相反，本名在前，父名在后，即本名＋父名，本名与父名之间用间隔号，如某人本名叫艾尼瓦尔，父名叫萨迪克，则其维语名字即为“艾尼瓦尔·萨迪克”。

基于维语人名组成的特殊性，在实际的翻译系统里，对于维语本土的名字，我们采用词典匹配的方法。但对于庞大的汉族人名来讲，建立完整的字典难度是很大的。而且因为汉语中存在多音字的关系，将维语翻译成汉语的时候，如果仅依赖词典会使翻译结果非常单一，无法满足灵活的需要。因此我们考虑，如果引入人名中每个字之间的统计关系，以及结合上下文的语境来进行人名识别和翻译，将会更加灵活和人性化。

需要注意的是，人名作为一种特殊的名词，可以缀接名词词缀。可以预见，这种情况会给人名识别带来很大难度，而缀接了名词词缀人名的识别也会有很大的不同。本文后面将对这种情况展开详细研究，此处不再累述。

3　维汉字母拼音映射关系

现在中国境内的维吾尔语使用的文字是以阿拉伯字母为基础的老维文（UEY）和拉丁字母为基础的拉丁维文（ULY）。在研究中我们发现，老维文可以无歧义地转换成拉丁维文，因此本文仅对拉丁维文进行处理。现代拉丁维文共有 32 个字母，其中有 8 个元音，24 个辅音。尤为重要的是，维吾尔语是一种拼音式文字。值得注意的是，维语的构成与特征与汉语拼音有着一定的映射联系（见表 1）。

表 1　汉语拼音与维语字母映射表

汉语	b	p	m	f	d	t	n	l	g	k	h	j	q	x	zh	ch	sh
维语	b	p	m	f	d	t	n	l	g	k	x	j	ch	sh	j	ch	x
汉语	r	z	c	s	y	w	a	o	e	i	u	v	ai	ei	ui	ao	ou
维语	r	z	s	s	y	w	a	o	ë	i	u	ü	ay	ëy	uy	aw	ow
汉语	iu	ie	ve	er	an	en	in	un	vn	ang	eng	ing	ong	uo	uan		
维语	yu	yë	yö	ër	en	ën	in	un	ün	ang	ëng	ing	ong	o	üan		

万方数据

通过观察维语语料中的汉族人名，我们发现汉语拼音(组合)到维语字母(组合)的映射并不是完全对应了以上规则。通过查阅资料，我们了解到在维语发音中，根据不同人的不同习惯，同样的发音可以有多种多样的拼写方式。在统计了大量维语汉族人名后，本文总结出了一些规则，共有7条(见表2)。

表2 补充的维语字母到汉语拼音的映射规则

维语字母(组合)	x	ay	u	e	üan	u ë	ong
汉语拼音(组合)	sh	ey	ü	ë	üen	ö	ung

4 维吾尔语中汉族人名的识别和翻译方法

本部分我们将详细介绍维吾尔语中汉族人名的识别和翻译方法，包括基于语言模型的汉族人名识别和翻译、维吾尔语名词词缀两个方面。

4.1 基于语言模型的汉族人名识别和翻译

统计语言模型(Statistical language model)通过大量对文本文件的统计，提取不同字、词之间先后发生的统计关系。目前主要采用的是n元语法模型(N-gram model)，这种模型构建简单、直接。本文主要借助SRILM工具包来进行语言模型的创建。SRILM是一个建立和使用统计语言模型的开源工具包，在Cygwin的平台上能实现训练、预测、计算的一系列操作。利用SRILM，我们可以方便地创建和运用多种基于N-gram的统计语言模型。

本文搭建了两个语言模型。分别用于维文中汉族人名的识别和翻译。在识别方面，汉族人名的构成与维语普通词的构成是有着一定差别的，这在统计信息上可以予以体现。利用这种不同，本文搭建拉丁维语语言模型来识别出维语中的汉族人名，由于维语中的汉族人名一般占用两个维语单词，所以采用维语二元语言模型。当识别出的汉族人名是词典中的人名时，对其翻译只需查找词典即可。而翻译的难点就在于那些不在词典中的人名该如何翻译。显然，为不断出现新的人名而建立丰富、全面的对照词库是不现实的。在汉语里，每个字在人名中出现的概率是不一样的，甚至于有些字的组合出现的概率也是不同的。譬如，“志洋”二字在名中出现的概率就要大于其他“zhi yang”组合的概率。由于汉族人名长度绝大多数为2和3，所以本文用汉语字符的三元语言模型选择最符合汉族人习惯的中文人名。下面我们详细介绍如何识别及翻译维语中的汉族人名。

识别的主要任务是要识别出文本中出现的拉丁维语中的汉族人名。由前文知，维语中的汉族人名在书写上与汉语人名有相同的规则，均为“姓[空格]名”，即姓名之间用空格隔开。但在实际情况中，我们发现由于书写习惯的不同，在用维语书写汉族人名时，有可能写成“姓名”的形式，即姓和名没有用空格间隔开。这就要求我们在识别过程中要能区分并正确识别这两种正常的拼写形式。识别时，我们先去查询输入的单词是否为无空格间隔开的人名，若不是则去查询是否为姓氏，若为姓氏，则初步判定当前输入词和下一输入词为人名，这时我们用拉丁维语的二元语言模型来判断这两个词的组合概率是否在阈值控制的范围内，以此来判断输入的两个词是否是真正可翻译的人名。具体识别步骤见图1。

翻译的主要任务是对识别为人名的两个单词翻译成中文人名。在翻译的过程中要考虑一下三种情况：(1)“姓 名”为人名库词典中存在的词条；(2)“名”为单字；(3)“名”为双字。

翻译时，我们在姓氏词典里查询输入的第一个词，再对第二个输入词进行分析，判断其是单字还是两个字，若是两个字则对其进行拆分。这样取出每一个单字后，我们用汉语的三元语言模型对每个单字的组合求概率，选择概率最大的组合为最佳翻译。具体翻译步骤见图2。

4.2 维吾尔语名词词缀

维吾尔语是一种形态变化很复杂的语言，其中名词是一种形态变化复杂的词类。维吾尔语属于阿尔泰语系突厥语族，黏着型语言。黏着语语言是一种有时态变化的语言类型，通过在单词的词尾粘贴不同的词缀来实现语法功能。维吾尔语中的名词词缀共有49个。在本文所搭建的人名翻译系统中，人名作为一类特殊的名词，其后也会缀接名词词缀。因此，在翻译过程中，需要识别出词缀才能截取出我们需要的人名，后续的翻译等工作才能顺利进行。

在图1中，相邻两个单词a、b作为输入，判断a是否为无空格间隔的人名，若非此种情况，则判断a是否为姓氏，若非姓氏，则判断“a 空格 b”是否为人名，若是人名则用拉丁维语二元语言模型计算a、b组合的概率，若小于固定阈值则识别成功，其余情况均视为失败。

在图2中，相邻单词a、b作为输入，若b不为单

万方数据

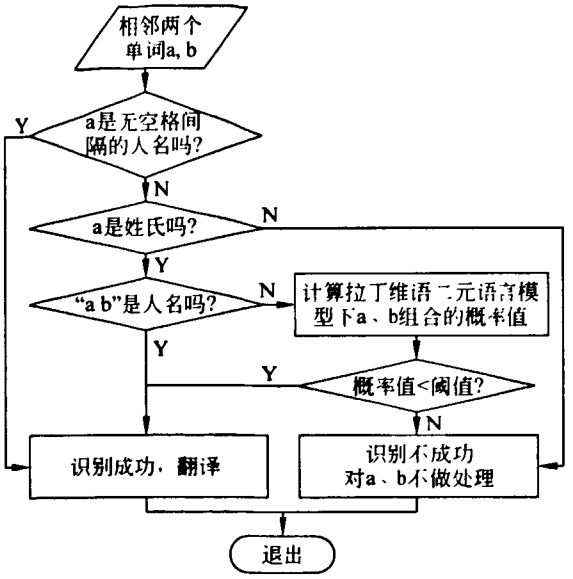

图 1　识别主要流程

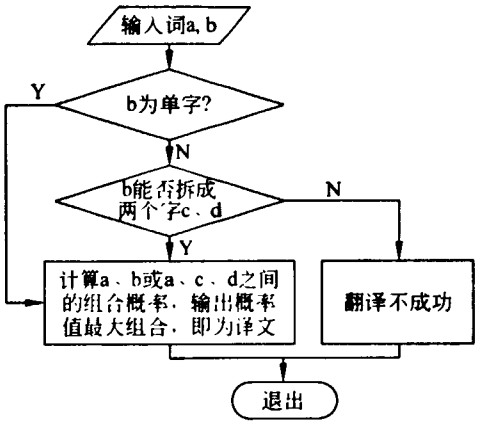

图 2　翻译主要流程

字，且可拆分成两个字 c、d，且用汉语三元语言模型计算姓和名各单字的组合概率，选取最大值组合为译文；否则视为不可译。

本文将对词缀识别的改进算法添加在对翻译文件的预处理阶段，即在人名的识别翻译前先对词缀进行过滤。我们考虑输入词有以下几种情况：(1)普通词，即非人名的词；(2)为无间隔空开的人名库中的姓名；(3)为姓氏；(4)为名；(5)为缀接词缀的无空格间隔的人名库中的姓名；(6)为缀接词缀的人名中的名字部分，其中名字可以为单字名，亦可以为双字名。识别词缀时，若是前四种种情况我们则不对输入词进行处理，若是后两者即缀接了词缀的人名的情况，我们用反向最大匹配去识别词缀，识别出词缀后，为保证切割掉词缀的部分可以正常翻译，需要对切割掉词缀的部分进行单字或双字的词典匹配，若可以匹配成功，则表明缀接了词缀的人名识别成功。其主要流程如图 3 所示。

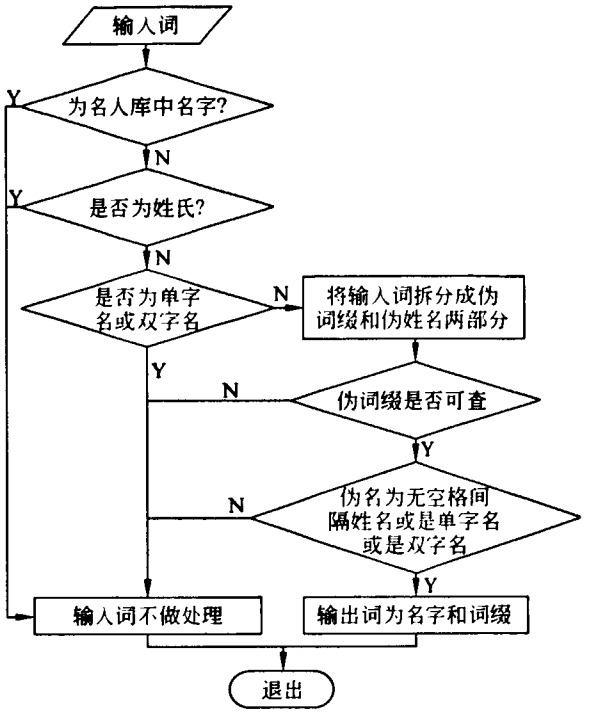

图 3　词缀识别处理主要流程

在图 3 中，依次判断输入词是否为名人库中的名字、是否为姓氏以及是否为单字或双字名，若皆非，则用反向最大匹配识别输入词的词缀，若词缀是可识别的，继续判断去掉词缀的部分是否为无空格间隔姓名、单字名或双字名中的一种，若是其中一种，则词缀识别成功，并将名字和词缀分别输出；其余情况视为识别不成功。

5　实验与分析

本文逐步实现了此前阐述的有关人名识别和翻译的功能，并对维语中汉族人名翻译进行了测试。

5.1　实验数据

本文所用的实验数据来自于新疆大学信息科学与工程学院的学者标注的 12 万维吾尔语语料以及搜狗官方网站提供的人名语料①，并在此基础上根据我们需求做了一定处理。

5.1.1　人名词典

我们从搜狗官方网站下载了国家领导人，名人以及常见人名细胞词库②，其中常见人名共 120 620

① http://pinyin.sogou.com/

② http://pinyin.sogou.com/dict/

万方数据

个词条。本文选取了中国历代国家领导人的姓名，名人的姓名作为名人人名库(共 3 720 词条)。在制作名人库的时候，充分结合了中国历史的特点，利用百家姓对所有人名进行了过滤，以保证所有的人名都是合乎中国文化及特点。在此基础之上，根据建立的汉字与维语拼音的映射，将名人库的姓名翻译成拉丁维语。与此同时，为了后面对人名进行翻译的时候有词典可查，分别生成姓以及名的各单字的维汉词典。

5.1.2 语言模型

新疆大学信息学院学者建立了一个规模为 119 737 句的维语语料库。其中，含有人名的句子有 5 874 句；不含人名的句子有 113 863 句。

本文使用了其中不含人名的 113 863 句来训练拉丁维语二元语言模型，通过此模型来判断输入词是以下哪种情况：(1)维语普通词与普通词；(2)维语普通词与汉族人名的姓氏；(3)汉族人名姓氏与名字；(4)汉族人名的名字与维语普通词。通过实验观察数据得知，以上情况中，绝大部分汉族人名姓氏与名字相邻的概率小于固定阈值。

本文同时使用了名人库(共 3 720 词条)以及常见人名(共 120 620 词条)进行分词，来搭建汉语的三元语言模型，建立汉族人名使用的单字之间的统计关系。

5.1.3 测试数据

本文对维语里中文人名翻译系统进行了测试。所用的测试数据来自于 12 万维语语料中含有汉族人名的句子，共 5 874 句，随机抽取其中 1 000 句进行测试。

5.2 实验结果

我们首先按照最初设计的识别及翻译流程，即不使用人们常用的错误的拼写规则及词缀识别，搭建了维语中汉族人名的翻译系统。在测试数据上对系统进行测试，并统计了识别的正确率及召回率。

接下来，我们根据前面总结的汉语拼音(组合)到维语字母(组合)的映射补充规则来重新建立汉字与维语拼写的映射，并创建词典，包括名人库的人名，姓氏及名字的词典。在此基础之上，用 SRILM 工具包重训拉丁维语二元语言模型以及汉语三元语言模型。用同样的测试数据进行测试，并统计结果。

最后，我们尝试根据前文提出的名词词缀识别的算法再次改进系统，期望能够在正确率提高的基础上，进一步提高召回率。

三次测试的结果对比如图 4 所示。

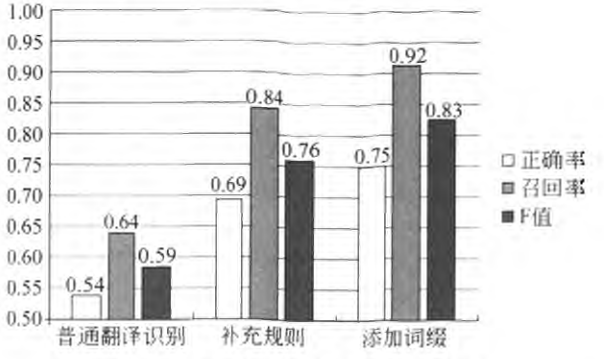

图 4 三次人名识别测试结果对比图

测试结果表明，在第一个人名识别及翻译系统中，仅有超过一半的人名翻译了。通过对比译文，总结原因有两点：(1)由于维吾尔族人发音不同的原因，有些维语字母对汉语拼音的映射并不是一对一的，譬如："yong"会拼写成"yung"；(2)人名是名词的一种特殊形式，而在维语里面有名词词缀，这些词缀在拼接到人名后时，会使人名无法识别和翻译，如"wang bangjün"在缀接了词缀"ning"后变成"wang bangjünning"，在识别过程中，第二个输入词就变成了"bangjünning"，由于词缀的出现，系统无法将其拆分成两个单字，拆分不成功将导致系统无法将其识别为人名。

从图 4 中可以看出，人名识别及翻译系统在补充规则后，识别的正确率提高了 15.5%，达到了 69.4%，召回率则有 20.1%的提高，达到 84.3%；在添加了词缀分析后，正确率在前者的基础上又有了 5.8%的提高，召回率有 7.2%的提高，分别达到 75.2% 和 91.5%。结果表明，人名识别及翻译系统最初使用的规则是不够全面的，补充了总结的汉语拼音(组合)到维语字母(组合)的映射规则后，识别及翻译效果得到了较大提升。而添加了词缀识别预处理模块后，系统识别的正确率达到了 75.2%，召回率更高达 91.5%。

另外，通过对比译文，我们发现，在对测试语料对应的中文译文进行词法分析的时候，词性标注有误，使人名翻译的译文存在噪声，这是系统识别召回率无法提升的主要原因。例如，"谷歌"的拉丁维语拼写为"gugël"，中文分词时将"谷歌"识别为人名，我们随机抽取含有"谷歌"的句子作为测试句，而"gugël"是无法识别和翻译为汉族人名的。对于正确率，由于在拉丁维语中，有些单词是与中文的姓氏拼写相同的，例如"si"、"ni"等，而这些拉丁维语单词

万方数据

出现的频率是比较高的，在语言模型中拥有较高的概率值，因此在通过拉丁维语二元语言模型计算与前后词的组合概率时，与这些单词拼写相同的姓氏也会被当做普通词而不被识别，这是造成正确率不够高的主要原因。

4 总结及下一步工作

本文针对维吾尔语中汉族人名的识别和翻译方法进行了一定的研究和探索。设计了通过汉字与维语拼音的映射规则来构造汉族人名的维语词典，使用语言模型来计算输入译文是否为名字及名字如何翻译。并采用添加规则和对人名词缀进行识别的方法来提高系统的翻译精度。实验结果表明了上述方法的可行性和有效性。

为了进一步提高维汉人名翻译系统翻译的质量，还需要收集和整理更多有关维汉人名翻译的资料，尤其是使用更为精准的测试语料。另外，在进行识别时，我们可以调整阈值寻找更加合理的参数值以识别出更多的汉族人名，进一步提高系统识别的正确率。

参考文献

[1] 宋柔，朱宏. 基于语料库和规则库的人名识别法[C]//陈力为. 计算语言研究与应用. 北京：北京语言学院出版社，1993.

[2] 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[C]//Ji Dong-Hong，国际中文电脑学术会议，新加坡，2001：323-328.

[3] 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.

[4] Zhang Huaping，Liu Qun，Yu Hongkui，et al. Chinese named entity recognition using role model[J]. The International Journal of Computational Linguistics and Chinese Language Processing，2003，8(2)：29-60.

[5] 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.

[6] Wu Youzheng，Zhao Jun，Xu Bo，et al. Chinese named entity recognition based on multiple feature[C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing(HLT/EMNLP)，Vancouver，2005：427-434.

[7] 衣马木艾山·阿布都力克木，吐尔地·托合提，艾斯卡尔·艾木都拉. 基于规则的维吾尔人名汉文机器翻译算法研究[J]. 计算机应用与软件，2010：86-87.

[8] 张秀玲. 汉维语人名文化异同之比较[J]. 新疆大学学报，2009，37(6)：136-139.

万方数据