基于字符串相似度的维吾尔语中汉语借词识别

米成刚[1,2]，杨雅婷[1]，周 喜[1]，李 晓[1]，杨明忠[3]

（1. 中国科学院 新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院大学，北京 100049；
3. 哈密地区电子政务办公室，新疆维吾尔自治区 哈密 839000）

(michenggang@gmail.com, yangyt@ms.xjb.ac.cn, zhouxi@ms.xjb.ac.cn,
xiaoli@ms.xjb.ac.cn, 1092851239@qq.com)

摘要：维汉机器翻译过程中会出现较多的未登录词，这些未登录词一部分属于借词(人名、地名等)。该文提出一种新颖的根据借词与原语言词发音相似这一特性进行维吾尔语中汉语借词识别的方法。该方法对已有语料进行训练，得到面向维吾尔语中汉语借词识别的维吾尔语拉丁化规则；根据以上规则对维吾尔语拉丁化，并对汉语词进行拼音化，将借词发音相似转换为字符串相似这一易量化标准；提出了位置相关的最小编辑距离模型、加权公共子序列模型以及二者的带参数融合模型。实验结果表明，综合考虑字符串全局相似性和局部相似性的带参数融合模型取得了最佳的识别效果。

关键词：借词；未登录词；发音相似度；字符串相似度

Recognition of Chinese Loan Words in Uyghur Based on String Similarity

MI Chenggang[1,2], YANG Yating[1], ZHOU Xi[1], LI Xiao[1], YANG Mingzhong[3]

(1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences , Urumqi 830011,China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. E-government office of Hami, Hami, Xinjiang Uygur Autonomous Region 839000, China)

(michenggang@gmail.com, yangyt@ms.xjb.ac.cn, zhouxi@ms.xjb.ac.cn,
xiaoli@ms.xjb.ac.cn, 1092851239@qq.com)

Abstract: There are many Out-Of-Vocabulary words in Uyghur-Chinese machine translation, a large part of them are loan words (including person names, place names, et.al). This paper presents a novel method that recognition the Chinese loan words in Uyghur according to the feature that one loan word pronounce similar with its original word. This method training the existing corpus first, and getting the Uyghur Latin rules that use to recognize Chinese loan word in Uyghur; this paper Latin the Uyghur words according to the rules, Romanization of Chinese words, these transform the sounds similarity to strings similarity which is easy to quantification; proposed three models: Position-related Minimum Edit Distance model, Weighted Common Subsequence model and the fusion model that fused above two with parameters. The experimental results show that the fusion model considering strings' global similarity and local similarity, so it gets the best recognition results.

Key words: loan words; Out-Of-Vocabulary words; pronunciation similarity; string similarity

---

基金项目：中国科学院战略性先导科技专项(XDA06030400)，中国科学院“西部之光”人才培养计划“西部博士资助项目”(XBBS201216)，中国科学院西部行动计划项目（KGZD-EW-501）

作者简介：米成刚(1986-)，男，博士研究生，研究方向为自然语言处理、机器翻译；杨雅婷(1985-)，女，副研究员，研究方向为多语种信息处理；周喜（1978-），男，副研究员，研究方向为多语种信息处理；李晓（1957-），男，研究员,研究方向为多语种信息处理；杨明忠（1971-），男，研究方向为电子政务。

1 引言

随着时代的快速发展，国与国之间、各民族之间的交流日益频繁。语言作为人们交流的主要工具，发挥着不可替代的作用。由于政治、地域等原因，使用一种语言的人们在交流过程中会用到另外一种语言中的词，经过一定时期，就会形成语言中的借词，也称外来词。如汉语中的“卡拉 OK(**からオケ**)”等词借自日语，“麦克风(Microphone)”、“沙发(Sofa)”等借自英语。

新疆维吾尔自治区地处亚欧大陆中部，同时受东西方文化影响。维吾尔语本身也接受了一些外来词。维吾尔语借词主要来自于汉语、俄语和阿拉伯语，本文针对其中的汉语借词进行识别。

目前主流的自然语言处理方法是基于统计的方法[1]，其最大的特点就是依赖于大规模语料。受语料规模及语言自身特性影响，在进行有关维吾尔语的自然语言处理(信息检索、语音识别、机器翻译等)研究过程中，会出现较多的未登录词 [2]，而其中的一部分未登录词就属于借词。本文根据借词发音较为相似这一特性，首先参考维吾尔语拉丁化规则，同时考虑维汉两种语言发音差异，将发音相似这一概念转化为字符串相似这一量化标准，同时考虑维吾尔语粘着性这一特点，提出了位置相关的最小距离模型(**Position-related Minimum Edit Distance，PMED**)以及加权的公共子序列模型(**Weighted Common Subsequence，WCS**)。在此基础上，进行两种模型的带参数融合。融合模型同时考虑维吾尔语中汉语借词识别的实际应用及维吾尔语语言特性，因而取得了最佳的识别效果。本文提出的将语音相似度转换为字符串相似度的方法，可以为发音较相似语言之间的机器翻译等研究提供新的思路。

2 相关工作介绍

借词(**Loan words**)，又称外来词。在历史发展的过程中，国家与国家之间，民族与民族之间，总会发生交流，当某种物品的名字在交流一方使用的语言中并不存在，或其中的一方特别强大时，借词就产生了，顾名思义，所谓借词就是一种语言从另一种语言中“借”来的词，通常这种词大部分属于音译词。

目前，国内外对借词的研究大都停留在语言学的范畴。对于英语这一国际化语言，主要面向其中的汉语普通话借词[3]，日语借词等展开研究；日语中的英语借词对日本社会、经济、文化等产生了巨大的影响[4]。通过调查社会上英语外来词的使用情况，研究人员对现代汉语中的英语外来词进行了全面、系统的分析[5][6]。

国内学者在汉维语外来词借入方法的对比[7]，借词对维吾尔语词汇的影响[8]，外来语对维吾尔语行业词的影响，现代维吾尔语中汉语借词[9]以及新疆地区方言借词[10]等方面对维吾尔语中的借词进行研究。针对维汉机器翻译中的具体应用，科研人员对维吾尔语中汉族人名的识别和翻译进行了研究[11]。

文中方法与以上论文中研究方法的区别在于，根据借词与原语言词发音相似这一特征，借鉴统计机器翻译中词对齐的思路获取维吾尔语字符与汉语拼音字母的最佳对齐规则(拉丁化规则)，使用综合考虑实际应用(维吾尔语中汉语借词识别)及维吾尔语语言特性的字符串相似度计算模型，识别出维吾尔语中的汉语借词。

3 面向汉语发音习惯的维吾尔语词拉丁化模型

2000 年，新疆大学等单位联合推广了拉丁维文，可以进行传统维吾尔文字向拉丁字母的转换。然而，由于维吾尔语和汉语两种语言发音上的差异性，导致维吾尔语中的汉语借词

发音与汉语原词发音有一定的差异，如汉语名词“桌子”(拼音:“zhuozi”)在维吾尔语“جوزا”中的发音“joza”(拉丁化)。为了减小这一差异性，本文提出一种面向汉语拼音发音习惯的维吾尔语汉语借词拉丁化模型。借鉴统计机器翻译中词对齐的思路，训练出适用于维吾尔语中汉语借词识别的拉丁化规则。

目前，统计机器翻译[12]领域的词对齐大多以 IBM 的 5 个模型[13]作为理论基础，其主要思路是根据平行语料中的词共现信息，采用统计机器学习方法，得到最佳的词对齐[14][15][16]效果。考虑到本文的实际问题，维吾尔语词中的字符应和拼音化汉语词字母之间顺序对齐，因此，选择实现较容易的 IBM 模型 2。

IBM 模型 2 可用公式(1)表示：

$$\begin{aligned} \mathrm{P}(\mathbf{e}|\mathbf{f}) &= \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t\left(e_j \,|f_{a(j)}\right) \qquad (1) \\ &= \frac{\varepsilon}{(l_f+1)^{l_e}} \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t\left(e_j \,|f_{a(j)}\right) \\ &= \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i) \end{aligned}$$

公式中的$\mathbf{e}$和$\mathbf{f}$分别是指分割后维吾尔语词字符向量以及分割后对应拼音化的汉语词字母向量。ε是归一化常数，l_e是 $\mathbf{e}$ 的长度，l_f是 $\mathbf{f}$ 的长度，$a_{(j)}$表示与维语词中第j个字符对齐的拼音字母在字母向量中的索引。

拉丁化规则的训练过程如图 1 所示：

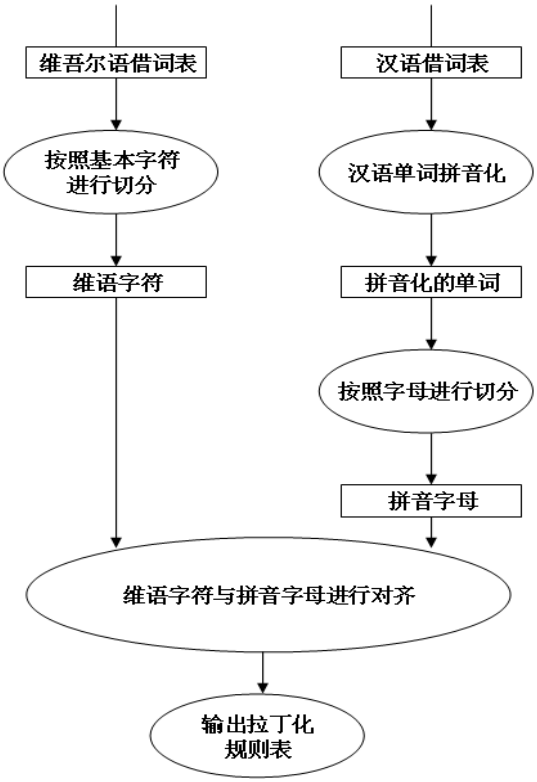

图 1 面向维吾尔语中汉语借词识别的维语拉丁化规则训练

在进行拉丁化训练的过程中，首先对维吾尔语词按照字符进行切分，再对对应汉语词进行拼音化，并按字母切分，然后分别将维语端的字符和汉语端的字母作为向量 e 和向量 f，进行对齐。得到面向维吾尔语中汉语借词识别的拉丁化规则。

4 维吾尔语中汉语借词识别

维吾尔语中汉语借词发现，是从维吾尔语单语语料中查找与汉语词发音相似的维吾尔语词的过程。本文提出的方法，是将语音层面的相似度通过维吾尔语词拉丁化(如本文第 2 章所示)和汉语词拼音化转化为字符串相似度进行计算，以获取最佳的识别效果。

本文选用字符串相似度算法进行计算。现有的字符串相似度算法属于通用的计算方法，不针对具体的应用场景。结合维吾尔语、汉语语言特征及维吾尔语中汉语借词识别这一特殊应用，本文以最小编辑距离算法和最长公共子序列算法为基础，提出了位置相关的最小编辑距离模型(Position-related Minim Edit Distance，**PMED**)和加权的公共子序列模型(Weighted Common Subsequence，**WCS**)以及两种模型的带参数融合模型(**PMED_ WCS**)。

4.1 位置相关的最小编辑距离模型(PMED)

4.1.1 最小编辑距离算法

编辑距离，又称 Levenshtein 距离，是指将一个字符串转换为另一个字符串需要进行字符的增加、删除和交换等操作的次数。最小编辑距离，即是进行字符串转换所需上述三种操作的最少次数。

如公式(2)所示：

(2)

初始化：

$$\mathrm{D}(0,0) = 0$$
$$\mathrm{D}(i,0) = \mathrm{D}(i-1,0) + del[x(i)]; \qquad 1 < i \le N$$
$$\mathrm{D}(0,j) = D(0,j-1) + ins[y(j)]; \qquad 1 < j \le N$$

递归方程：

$$\mathrm{D}(i,j) = \min\begin{cases} \mathrm{D}(i-1,j) + del[x(i)] \\ \mathrm{D}(i,j-1) + ins[y(j)] \\ \mathrm{D}(i-1,j-1) + sub[x(i),y(j)] \end{cases}$$

4.1.2 位置相关的最小编辑距离模型 PMED

最小编辑距离算法可以全局地考虑两个字符串的相似度。针对本文中的问题，由于维吾尔语自身的语言特征及其构词方式(通过在词干后附加若干词缀构成新词)，维吾尔语中的汉语借词词尾可能包括词缀，这就使得在使用编辑距离算法计算相似度时可能在词尾进行多次删除操作，导致编辑距离过大，影响最终的识别效果。PMED 在继承最小编辑距离算法全局性这一优点的同时，关注拉丁化维吾尔语词与拼音化汉语词计算编辑距离时删除操作的位置，若有连续的删除操作发生在拉丁化维吾尔语词的词尾，则计算编辑距离时减去在词尾连续删除操作的次数，最终相似度得分取其与最小编辑距离两者中较小值。如公式(3)所示：

$$\mathrm{Sim}_{PMED} = \begin{cases} MED_{PMED}(u_i, c_j); & \text{no continue delete occur} \\ \min\{MED_{PMED}(u_i, c_j), ED_{PMED}(u_i, c_j) - times_{ECD(u_i)}\}; & \text{continue delete} \end{cases} \quad (3)$$

上式中$ED_{PMED}(u_i, c_j)$是拉丁化维吾尔语词u_i与拼音化汉语词c_j的编辑距离，$MED_{PMED}(u_i, c_j)$是最小编辑距离，$times_ECD(u_i)$是指计算维语词u_i和汉语词c_j编辑距离时，维语词u_i结尾连续删除操作的次数。

4.2 加权的公共子序列模型(WCS)

4.2.1 最长公共子序列

最长公共子序列，英文缩写为 LCS（Longest Common Subsequence）。其定义是，一个序列 S ，如果分别是两个或多个已知序列的子序列，并且是所有符合此条件序列中最长的，则 S 称为已知序列的最长公共子序列。

其核心算法可用公式(4)表示：

(4)

$$\mathrm{L[i,j]} = \begin{cases} 0 & ,i = 0 \ or \ j = 0 \\ L[i-1, j-1] + 1 & ,i > 0, j > 0, a_i = b_j \\ \max\{L[i, j-1], L[i-1, j]\} & ,i > 0, j > 0, a_i \neq b_j \end{cases}$$

4.2.2 加权的公共子序列模型 WCS

本文提出的识别模型，主要是根据借词与原词发音相似这一特征，进行维吾尔语中汉语借词的识别。由于两种语言发音的差异性，造成拉丁化的维吾尔语词与拼音化的汉语词之间不能做到完全对应，因此，基于最长公共子序列算法，本文提出了加权的公共子序列模型(WCS)[17]。此模型不仅考虑最长公共子序列，而是考虑所有的公共子序列，并为不同长度的公共子序列赋予不同的权值。对所有的公共子序列与权值乘积求和，以和最大者对应维语词为借词。此方法最大的特点是量化了“字符串连续相似”这一因素。如公式(5)所示：

$$\begin{aligned} Sim_{WCS} &= WeightedCommStrings(u_i, c_j) \\ &= NUM_1 * LEN_1 + NUM_2 * LEN_2 + \cdots + NUM_n * LEN_n \\ &= \sum_n NUM_i * LEN_i \end{aligned} \quad (5)$$

u和c分别是拉丁化维语词及拼音化汉语词，NUM_i是长度为i的公共子序列数目，LEN_i为可能的子序列长度。考虑到公共子序列长度的不可预测性，我们将最长公共子序列长度设置为两个字符串中较短字符串长度。为了使得较长的公共子序列获得较高的得分，本文在计算相似度时，求公共子序列长度与子序列数目的乘积，并将结果求和。

4.3 融合两种模型的相似度计算(PMED + WCS)

基于最小编辑距离算法重点考量的是字符串之间进行互相转换(将字符串 A 转换为字符串 B)时的最小代价，不能反映“连续子序列相似”这一事实，其改进算法 PMED 也存在这一问题；基于公共子串算法从局部相似出发，一定程度上解决了最小编辑距离算法存在的问题，然而，此算法及其改进算法 WCS 却有全局性不强的缺点。因此，结合两种模型的优点，构成最终的相似度计算模型$SIM_{PMED+WCS}$。如公式(6)所示：

$$SIM_{PMED+WCS} = \alpha Sim_{WCS} + \beta(-Sim_{PMED}) \quad (6)$$

由于Sim_{PMED}使用的是基于编辑距离的相似度计算方法，计算结果越小两个字符串越

相似，因此，Sim_{PMED}中使用了$(-Sim_{PMED})$。另外，针对本文中的具体应用，两种模型所占的比率有所不同；为了获取最佳的识别效果，分别在每个模型前附加参数α、β。参数通过 EM(Expectation Maximization)模型[18]进行训练。

4.4 举例

以维吾尔语中汉语借词“جوزاڭنى”(“桌子”的第三人称单数)的识别为例，对本文提出的 PMED 和 WCS 模型做进一步的说明。PMED + WCS 模型是两种模型的带参数融合。

1) 维吾尔语词拉丁化

传统拉丁化：“jozangni”

修正拉丁化：“zhozangni”，根据维吾尔语拉丁化的修正规则(5.2.1 节)，将“ج”转写为“zh”

2) 汉语词拼音化

汉语词拼音化：“zhuozi”

3) 相似度计算

PMED 模型

如公式(3)所示

Sim_{PMED}(“zhozangni”, “zhuozi”) (7)

= min { MED_{PMED}(“zhozangni”, “zhuozi”), ED_{PMED}(“zhozangni”, “zhuozi”) – $times_{ECD(\text{“zhozangni”})}$}

= min {5, 2}

= 2

公式(7)中，$MED_{PMED}(u_i, c_j)$计算u_i与c_j的最小编辑距离，$ED_{PMED}(u_i, c_j)$计算u_i与c_j编辑距离，会在维吾尔语词尾发生连续删除操作，$times_{ECD(u_i)}$为词尾连续删除操作的次数。维语词“جوزاڭنى”词尾含第三人称单数后缀。

WCS 模型

如公式(5)所示

Sim_{WCS}(“zhozangni”, “zhuozi”) $= \sum_n NUM_i * LEN_i$ (8)

= 1*1 + 2*2

= 5

公式(8)中，NUM_i是长度为i的公共子序列数目，LEN_i为可能的子序列长度。

5 实验设计与数据分析

5.1 实验语料

本文实验所用语料包括：一、维吾尔语拉丁化修正规则训练语料，主要是人名、地名等维汉对应的双语词(共 1000 词对)；二、维吾尔语汉语借词识别语料，主要是借词识别测试语料(共 50000 句，平均每句含 20 个维吾尔语单词)及其参考测试结果语料(共 5000 词对)，测试语料来自新闻领域。

5.2 实验过程

以下分维吾尔语拉丁化规则修正和维吾尔语中汉语借词识别两个阶段进行实验。

5.2.1 维吾尔语拉丁化规则修正

为了减小拉丁化后维吾尔语词与拼音化后汉语词的差异，根据现有的语料，借鉴统计机器翻译中的词对齐方法，获取适合本文实际应用的拉丁化规则。

首先，对维汉词对中的维吾尔语词按字符进行切分；对汉语词进行拼音化，并对汉语拼音按照字母进行切分，获得维吾尔语字符向量、汉语拼音字母向量对齐语料；

其次，将拉丁化规则的获取问题转换为维吾尔语字符与汉语拼音字母对齐问题；对齐采用统计机器翻译中广泛使用的词对齐工具 GIZA++进行。综合考虑此处面临问题及其运行效率，使用其中的 IBM 模型 2 即可；

最后，根据对齐结果，获取修正的拉丁化规则。如表 1 所示(其中粗斜体的拉丁字母对应的即是进行修正后的拉丁化规则，为了简明起见，将字符“ئ”对空)：

ا	ه	ب	پ	ت	ج	چ	خ	د	ر	ز	ژ	س	ش	غ	ف
a	***an***	b	p	t	***zh***	ch	h	d	r	z	zh	s	x	gh	f
ق	ك	گ	ڭ	ل	م	ن	ھ	و	ۇ	ۆ	ۋ	ۇ	ي	ى	ې
k	k	g	ng	l	m	n	h	o	u	ue	v	w	***e***	i	y

表 1 面向维吾尔语中汉语借词识别的维语拉丁化规则

5.2.2 维吾尔语中汉语借词识别

为了验证各个模型的有效性，分别在位置相关的最小编辑距离模型(**PMED**)，加权的公共子序列模型(**WCS**)以及带参数融合模型(**PMED + WCS**)三种模型上进行实验。实验结果用准确率 P (Precision)、召回率 R(Recall)以及 F1 值来表示。

F1 计算方法如公式(9)所示：

$$\mathrm{F1} = 2 \times \frac{P \times R}{P + R} \tag{9}$$

实验 1　位置相关的最小编辑距离模型(PMED)

首先，分别进行汉语词的拼音化和维吾尔语词的拉丁化(使用 5.2.1 中得到的修正的维吾尔语拉丁化规则)，根据 4.1.2 中的方法，从位置相关的最小编辑距离模型(PMED)中得到各个维吾尔语词-汉语词对应得分，取最小项作为最终结果。为了进行对比，此处也在最小编辑距离算法模型(MED)上进行了实验。结果如表 2 所示：

	准确率(P)	召回率(R)	F1 值
MED	62.35%	71.09%	66.43%
PMED	64.73%	75.68%	69.78%

表 2 位置相关的最小编辑距离模型 PMED 和最小编辑距离模型 MED 识别结果

实验 2　加权的公共子序列模型(WCS)

进行汉语词的拼音化和维吾尔语词的拉丁化(使用修正的维吾尔语拉丁化规则进行)，根据 4.2.2 中的 WCS 模型(加权的公共子序列相似度计算模型)，计算出各词对的相应得分，取得分最高的维吾尔语词为识别出的借词。为了与 WCS 模型进行对比，此处也在最长公共子序列模型(CS)上进行了实验。结果如表 3 所示：

	准确率(P)	召回率(R)	F1 值
CS	63.06%	73.12%	67.72%
WCS	65.90%	74.34%	69.87%

表 3 加权的公共子序列模型 WCS 和公共子序列模型 CS 识别结果

实验 3: 融合两种模型的维吾尔语中汉语借词识别

融合模型(**PMED + WCS**)是对两个模型(**PMED 和 WCS**) 进行带参数融合。首先对模型进行训练，确定两参数的最优值。使用拉丁化的维吾尔语汉语借词与拼音化的汉语词进行训练。根据 EM 算法的步骤，首先对参数进行初始化，再重复执行 E 步和 M 步，直到 F1 值收敛。F1 值最高时对应的 α 和 β 作为参数的取值，其中的训练语料使用 10000 词测试语料，F1 值评价采用对应的 200 词参考语料。参考实验 1 中的方法进行借词识别实验。为了显示带参数模型的有效性，同时使用无参数模型(PMED+WCS_P1)进行实验。结果如表 4 所示：

	准确率(P)	召回率(R)	F1 值
PMED+WCS_P1	65.57%	75.31%	70.10%
PMED+WCS	66.32%	77.28%	71.38%

表 4 带参数的融合模型和未带参数融合模型识别结果

5.3 实验数据分析

对实验数据进行分析，可以得出：

实验 1 使用位置相关的最小编辑距离模型求取拉丁化后维吾尔语词与拼音化汉语词的字符串相似度，最小值对应汉语词为识别出的借词。与最小编辑距离算法相比，位置相关的最小编辑距离模型考虑到了维吾尔语的构词方式(词干加若干词缀)，在计算编辑距离的同时，监测进行连续删除操作的位置，若发生在维语词尾，则对编辑距离计算结果进行修正。PMED 模型兼顾字符串相似全局性以及维吾尔语语言特点，因此，与最小编辑距离算法相比，PMED 模型取得了较高的识别准确率，如表 2 所示。

实验 2 根据提出的加权公共子序列模型，不仅考虑到了字符串的局部相似性，而且对所有的公共子序列根据其长度赋予不同的权值。相比于传统的最长公共子序列算法，加权的公共子序列模型(WCS)更好地反映了拉丁化维语词与拼音化汉语词的相似性，因而对借词的识别准确率较高，如表 3 所示。

实验 3 中的带参数融合模型(PMED+WCS)结合了 PMED 和 WCS 的优点。从维吾尔语中汉语借词识别这一具体任务出发，考察维吾尔语构词特点以及维汉两种语言发音差异，综合字符串的全局相似性与局部相似性，并使用 EM 算法，分别赋予两种模型(PMED 和 WCS)不同参数，更好地反映了具体语料中不同模型对最终识别结果的影响。实验结果表明，与上述两种模型相比，PMED+WCS 模型取得了最佳的借词识别效果，如表 4 所示。

6 结束语

本文根据维吾尔语中汉语借词与原汉语词发音相似这一特点，将语音相似度转换为字符串之间相似度进行维吾尔语中汉语借词的识别。对现有的维吾尔语词借词-汉语语料进行处

理，对维吾尔语词进行字符切分，对对应汉语词进行拼音化，借鉴词对齐方法，训练出适合汉语拼音发音的维吾尔语拉丁化规则；根据字符串相似度这一量化标准，分别将测试语料中维吾尔语词进行拉丁化(修正的拉丁化规则)，汉语词拼音化，使用文中提出的三个模型 PMED、WCS 和 PMED + WCS 进行实验。结果显示，综合考虑字符串全局相似性、局部相似性以及维吾尔语语言特性等因素的 **PMED + WCS** 模型获得了较高的识别准确率。文中采用的方法为发音较相似语言对之间的翻译提供一种研究思路；本文的实验结果可以作为维汉机器翻译的辅助知识源；另外，可以应用本文提出的方法根据汉语中发现的新词进行相应维吾尔语文本中新词的发现。然而，汉语中存在多音字的情况，会影响到借词识别结果，从而影响到最终的应用(如维汉机器翻译)，后续将针对这一问题展开研究。

参考文献

[1] Chris Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing [M],Cambridge: MIT Press, 1999

[2] Chung-Chi Huang and Ho-Ching Yen and Ping-Che Yang, et al. Using Sublexical Translations to Handle the OOV Problem in Machine Translation [J]. ACM Transactions on Asian Language Information Processing, 2011, 10(3): 16

[3] LAUREN ASIA HALL-LEW. ENGLISH LOANWORDS IN MANDARIN CHINESE [D]. ARIZONA: THE UNIVERSITY OF ARIZONA, 2002

[4] GILLIAN KAY. English loanwords in Japanese [J]. World Englishes, 1995, 14(1): 67-76

[5] 潘子助. 试谈汉语中的英语借词[J]. 湖北函授大学学报,2011,24(7):110-111

[6] Kui Zhu.On Chinese-English Language Contact through Loanwords[J]. English Language and Literature Studies,2011,1(2):100-105

[7] 陈燕，陈平. 汉维语外来词借入方法对比研究[J]. 喀什师范学院学报，2011,32(2):51-55

[8] 郑燕. 借词对维吾尔语词汇的影响[J]. 湖北第二师范学院学报, 2011,28(1):37-39

[9] 陈世明. 维吾尔语汉语借词新探[J]. 西北民族研究，2007,1：5

[10] 周磊. 乌鲁木齐方言借词研究[J]. 方言，2004,4: 347-355

[11] 李佳正，刘凯，麦热哈巴·艾力，吕雅娟，刘群，吐尔根·依布拉音. 维吾尔语中汉族人名的识别及翻译[J]. 中文信息学报，2011，25(4)：82-87

[12] Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical Phrase-Based Translation[C]// Proceeding NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language. Edmonton, Canada: ACL, 2003:48-54

[13] Peter F.Brown, Stephen A.Della Pietra, Vincent J. Della Pietra, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation [J]. Computational Linguistics, 1993, 19(2): 263-311

[14] Yang Liu, Qun Liu, Shouxun Lin. Log-linear Models for Word Alignment[C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor: ACL, 2005: 459-466

[15] Chris Dyer, Jonathan Clark, Alon Lavle, et al. Unsupervised Word Alignment with Arbitrary Features[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon: ACL, 2011: 409-419

[16] Robert C.MOORE. Improving IBM Word-Alignment Model1[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain: ACL,

2004:519-526
[17] 阿依克孜·卡德尔，开沙尔·卡德尔，吐尔根·依布拉音. 面向自然语言信息处理的维吾尔语名词形态分析研究[J]. 中文信息学报, 2006, 20(3): 43-48
[18] Mehryar Mohri, Fernando Pereira, Michael Riley. Weighted Automata in Text and Speech Processing[C]// 12th European Conference on Artificial Intelligence. Budapest: John Wiley & Sons, Ltd, 1996: 5

作者联系方式： 米成刚　新疆维吾尔自治区乌鲁木齐市北京南路40-1号中国科学院新疆理化技术研究所615室　830011 15899155629　michenggang@gmail.com