文章编号:1000-5641(2008)03-0083-09

信函英文地址的自动识别和翻译

屠 晓[1], 陈国跃[2], 吕 岳[1]
(1. 华东师范大学 计算机科学与技术系,上海 200062;
2. 秋田县立大学 电子与信息系统系,日本秋田 015-0055)

摘要: 根据地址的语言特性,定义和归纳了适用于邮政信函英文地址自动识别和翻译的用语规则.针对由字符识别技术获得的信函地址,提出了一种基于非精确字符串匹配技术的地址翻译方法,自动识别出英文地址并翻译成中文.实验结果验证了该方法的有效性,能较好地减少识别错误带来的影响,提高系统的翻译性能.

关键词: 非精确匹配; 基于规则; 地址识别; 地址翻译; OCR
中图分类号: TP391.2 **文献标识码**: A

Automatic recognition and translation of English address on postal mails

TU Xiao[1], CHEN Guo-yue[2], LÜ Yue[1]
(1. *Department of Computer Science and Technology, East China Normal University, Shanghai* 200062, *China*;
2. *Department of Electronics and Information Systems, Akita Prefectural University, Akita* 015-0055, *Japan*)

Abstract: According to the characteristics of the language used in address, rules different from those of natural languages were defined and applied to automatically recognize and translate English address on postal mails. To deal with the address got from Optical Character Recognition, an English-to-Chinese address translation method based on inexact string matching technology was proposed. The experimental results showed that the present method is capable of reducing OCR errors, and improving the translation performance.

Key words: inexact matching; rule-based; address recognition; address translation; OCR

0 引 言

机器翻译是使用计算机把一种语言自动翻译成另一种语言的一项新技术.从支持多种语言互译的实用化机器翻译系统SYSTRAN到搜索引擎Google的翻译功能,人们已经取

收稿日期:2007-09
基金项目:国家自然科学基金(60475006);教育部新世纪优秀人才支持计划(NCET-05-0430);上海市曙光计划(05SG29)
第一作者:屠晓,女,博士研究生.
通讯作者:吕岳,男,博士,教授,博士生导师,主要研究方向为模式识别、图像处理和自然语言理解等.
E-mail: ylu@cs.ecnu.edu.cn

得了许多成果[1,2],并应用于各个领域.目前机器翻译系统的输入一般是文本格式的文字,极少是以图像识别结果作为处理对象.随着 OCR 技术的发展,有效的图像文字识别技术已经日渐成熟.将机器翻译技术和图像识别技术有机结合起来会有广阔的应用前景,例如邮政信函上英文地址的自动识别和翻译等.

国外寄达中国的信函需由邮政部门的专业批译人员将英文收信人地址翻译成中文,批注在信封上,以便于投递人员送达目的地.随着信函处理量的与日俱增,批译人员不堪重负.因此,开发邮政信函自动地址翻译系统有迫切的需要.该系统采集信函图像,从中分割和识别出收信人地址的英文字符,结合非精确字符匹配技术和地址用语规则自动识别出地址信息,并将其翻译成中文.本文提出的地址识别和翻译方法有机地将 OCR 技术和机器翻译技术结合起来,是面向邮政自动化领域的一项应用研究和探索.

1 地址识别和翻译方法

在采用 OCR 技术的系统中,错识是不可避免的.图 1 是本系统从一封真实信函上采集到的图像,利用图像分析技术,从中分割和定位出收信人地址区域,见图 2.

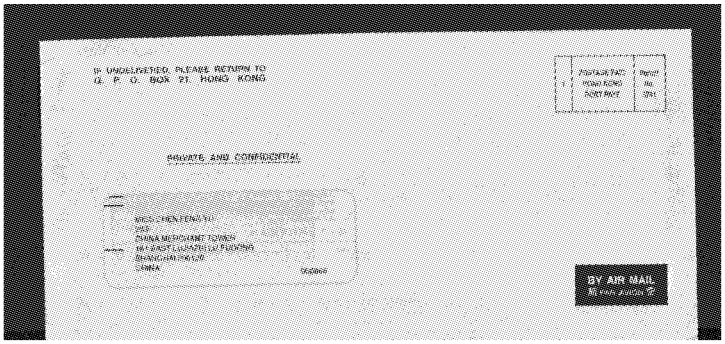

图 1 采集到的一幅真实信函图像

Fig. 1 An envelop image captured from a real letter

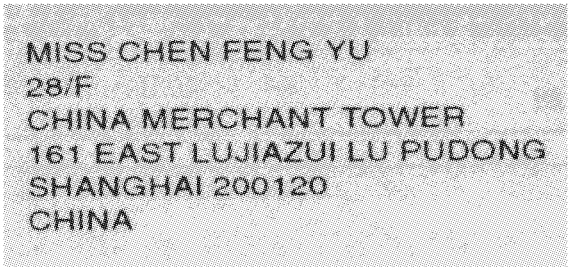

图 2 定位得到的收信人地址区域

Fig. 2 The mail address area

对图 2 所示的图像进行字符识别,结果如下:

MISS□CHEN□FENG□YU
28/F
CHINA□MERCHANT□TOWER
161□EAST□LUJIAZU1□LU□PUDONG

SHANGHAI□2000120
CHINA

其中‘□’表示空格. 显然可以看出“LUJIAZUI”的“I”被错识成“1”. 类似的误识降低了后续翻译的正确性. 本文针对这个问题提出了一种基于非精确字符串匹配技术的地址翻译方法，以降低字符识别错误带来的影响.

1.1　基于非精确匹配的字符串相似性度量

设 A 为标准字符串，有 m 个字符组成，用 $a_1, a_2, \cdots, a_m$ 表示；B 为 OCR 识别结果，由 n 个字符串组成，用 $b_1, b_2, \cdots, b_n$ 表示. 如何计算 B 与 A 之间的相似度是地址翻译的关键. 本文采用基于动态规划[3]的非精确字符串匹配方法来计算 B 与 A 之间的相似度，用一个 $(m+1)\times(n+1)$ 的矩阵 $\boldsymbol{V}$ 记录比较结果.

矩阵 $\boldsymbol{V}$ 初始化：$\boldsymbol{V}(i,j)=0, 0\leqslant i\leqslant, 0\leqslant j\leqslant m$.

$\boldsymbol{V}(i,j)$ 的值按如下循环计算：

$$\text{For}\quad 1\leqslant i\leqslant m, 1\leqslant j\leqslant n,$$

$$\boldsymbol{V}(i,j)=\max\begin{cases}\boldsymbol{V}(i-1,j-1)+\sigma(a_i,b_j),\\ \boldsymbol{V}(i-1,j),\\ \boldsymbol{V}(i,j-1).\end{cases}$$

其中 $\sigma(a_i,b_j)$ 为字符 a_i 与字符 b_j 之间的相似性度量，定义为

$$\sigma(a_i,b_j)=\begin{cases}2 & a_i=b_j,\\ -2 & a_i\neq b_j.\end{cases}$$

定义 B 与 A 的相似度为

$$\mathrm{Sim}(A,B)=\frac{\boldsymbol{V}(m,n)}{\tilde{\boldsymbol{V}}_A}.$$

其中 $\boldsymbol{V}_A=m\times\sigma(a_i,a_i)=2m$，是 A 与其本身的相似度. 取一定的阈值 θ，当 $\mathrm{Sim}(A,B)$ 大于 θ 时，则认为两个字符串是一致的.

1.2　地址识别

地址识别采用了机器翻译技术，即对由 OCR 得到的结果进行地址分析和理解，提取出地址的相关信息，以便对照信函地址的数据库记录，将信函的英文地址自动翻译成中文. 目前大多数实用的机器翻译系统的基本方法仍然是基于语言规则的[1,2]，即使是基于语料库的机器翻译系统也离不开语言规则. 不论是基于统计的机器翻译[4,5]还是基于实例的机器翻译[6]都是如此. 然而使用英文书写的中国地址夹杂了大量具有一音多形性的中文拼音，同一地址又有多种表达方式. 本文针对邮政地址翻译这一特定应用，并以寄达上海地区的国际信函为研究对象，定义和归纳了相应的规则，用于地址的分析和理解.

1.2.1　词性标定

地址的词性标定与自然语言中的词性标定本质上是类似的，即确定每个词的词性. 不同的是，地址中的语言具有其特殊性，如地址中通常包括 Road, Street, Room 和 Building 等有利于地址识别的关键词. 因此需要定义不同于自然语言的词性标注集、词典以及词法规则.

1.2.1.1　词性标注集

根据国外寄达信函的地址特点，本文将词的类别归纳为 17 种主词性，个别主词性又细分为若干副词性，参见表 1.

表 1 词性标注集

Tab. 1 Tagset

词性			描述及例子
市(City-keyword)		拼音(Ck-c)	如 Shi(市)
		单词(Ck-e)	如 City
市级地名(CityRegion)			如 Shanghai(上海)等
区(District-keyword)		拼音(Dk-c)	如 Qu(区)
		单词(Dk-e)	如 District
区级地名(DistRegion)			如 Pudong(浦东),Xuhui(徐汇)等
邮政编码(Postcode)			6 位连续数字,如 200062 等
路名关键词(Road-keyword)		拼音(Rk-c)	如 Lu(路),Dadao(大道),Gonglu(公路)等
		单词(Rk-e)	如 Road,Street,Boulevard,Avenue
方位词 (Oriental)			如 East,West,South,North,Middle,Central 等
小区/园区名关键词 (Area-keyword)		拼音(Ak-c)	如 cun(村)等
		单词(Ak-e)	如 Park,Zone 等
号码关键词 (No-Keyword)	一级号码关键词 (No-keyword1)	拼音(Nk1-c)	如 Nong(弄)等
		单词(Nk1-e)	如 Lane 等
	二级号码关键词 (No-keyword2)	拼音(Nk2-c)	如 Hao(号),Lou(楼),Danyuan(单元)等
		单词(Nk2-e)	如 No,Building,Block 等
	三级号码关键词 (No-keyword3)	拼音(Nk3-c)	如 Shi(室)
		单词(Nk3-e)	如 Room,Flat,Apartment,Suite,Unit 等
楼层关键词(Floor-keyword)			如 Floor,F 等
大楼关键词 (Building-keyword)		拼音(Bk-c)	如 Dalou(大楼),Dasha(大厦)等
		单词(Bk-e)	如 Building,Mansion,Hotal 等
公司关键词(Company-keyword)			如 Company,Limited,Corporation 等
称谓关键词(Title-keyword)			如 Miss,Mrs, Mr,Prof 等
号码(Number)		数字(Digital)	数字串。如 36,1025
		单字母(Single)	用于区分同一区域的楼房或者房间,如 A 楼,B 座等
		序数词(Order)	如“3rd”,“11st”等
符号(punctuation)			标点符号,如“,”,“.” “-”,“/”等
逻辑词(Logic-keyword)			如“&”,“And”,“or”等
字母串(Chars)			除去上述之外的字母串,如“HUANGPI”等

1.2.1.2 词典

本系统定义的词典以词条为基本单位,标注每个词的词性以及用于判断词性的规则,其具体格式如下:〈词〉〈匹配阈值〉〈词性〉〈规则序号〉.

其中,〈匹配阈值〉是采用非精确匹配算法判断待定词与词条中的词是否一致的参数. 计算待定词和词典中所有词的相似度,假设最高相似度为 Sim_{max},其对应的词条中〈匹配阈值〉的值为 θ,如果 $Sim_{max} \geqslant \theta$,则按照该词条的后两项〈词性〉和〈规则序号〉确定词性. 本文提出的地址识别是建立在关键词确定的基础上,因此确定关键词的匹配阈值设置为1.0,而其它设置为0.8.〈词性〉表示词的类别,可以是多项,用逗号分开. 本文将词分为单性词(只有一种词性的词)和多性词(具有多种词性的词).〈规则序号〉的值是词法规则的序号,为可选项,单性词的词条无〈规则序号〉这一项,多性词根据〈规则序号〉所指的词法规则来确定其词性. 对

于词典中不存在的词标为“字母串”.

1.2.1.3　词法规则

由表1可以看出大部分由英文单词构成的关键词的词性是唯一的,而由拼音构成的关键词则具有多种词性,因此用于消除歧义的词法规则主要是针对拼音的.词法规则根据前后相邻的词的词性以及在文本行中的位置确定当前词的词性,也可称为夹逼规则[7].本文将词法规则形式化地表示为

〈序号〉〈词〉〈[条件1],[条件1’],词性1〉〈[条件2],[条件2’],词性2〉…〈[条件n],[条件n’],词性n〉〈词类0〉.

其中[条件]是词性、行首和行尾的组合,即 词性1|词性2|…|行首|行尾,“|”表示逻辑或.若当前词前一个词的词性符合条件1,且后一个词的词性符合条件1’,则当前词的词性为词性1.如果前面条件都不满足,则该词的词性为词性0.例如

〈01〉〈shi〉〈市级地名,市-拼音〉〈数字,符号|行尾,三级号码关键词-拼音〉〈字母串〉.

其含义为:若当前词“shi”的前一个词是“市级地名”,则该词的词性为“市-拼音”;若当前词“shi”的前一个词是“数字”,并且后一个词是“符号”或者是文本行的行尾,则“shi”的词性为“三级号码关键词-拼音”;若都不成立,则“shi”的词性为“字母串”.

1.2.2　地址信息项的生成

信函收信人地址一般包含收信人姓名、房间号、楼号、弄、路名、区、市、国家、邮政编码、写字楼名称和居民小区名称等,将这些称为地址信息项.根据标定出的词性,以关键词为起点,采用有线状态自动机识别出各个地址信息项.

以路名为例.路名是表达方式较多的一类信息项,也是地址的重要信息,因此是地址识别的难点.常见的路名有以下6种形式:

(A) 字母串…字母串 路名关键词,如Century Boulevard(世纪大道);

(B) 字母串…字母串 方位词 路名关键词,如Zhong Shan North Road(中山北路);

(C) 字母串…字母串 数字 路名关键词,如Rui Jin 1 Lu (瑞金一路);

(D) 方位词 字母串…字母串 路名关键词,如West Nan Jing Road(南京西路);

(E) 字母串…字母串 路名关键词 方位词,如Guang Yuan Road West (广元西路);

(F) 字母串…字母串 No数字 路名关键词 方位词, 如Zhong Shan No 2 Road South.(中山南二路).

将一个完整的路名分为三个部分:前缀、路名关键词和后缀,其中后缀可缺省.以“Guang Yuan Road West”为例,“Guan Yuan”为前缀,“Road”为路名关键词,“West”为后缀.采用有限状态自动机的方法,对路名前后缀进行确定.图3(a)(b)分别给出了用于确定前缀和后缀的有限状态自动机DFA1和DFA2.将词性标定后的地址以路名关键词为界,分为两个部分,分别作为DFA1和DFA2的输入串.DFA1可以接受的最大子串为路名的前缀;DFA2可接受的最大子串为路名后缀.合并前缀、路名关键词和后缀可等到完整的路名.其他地址信息项则根据其对应的状态自动机得到.

1.3　地址翻译

信函数据库的每个地址由多个地址信息项组成,且每个地址信息项有其对应的中文表述.将地址识别得到的地址信息项与数据库中的记录进行相似性比较,若得到的最高相似度满足一定判断条件,则对应记录中的中文表述就是信函地址的翻译结果.

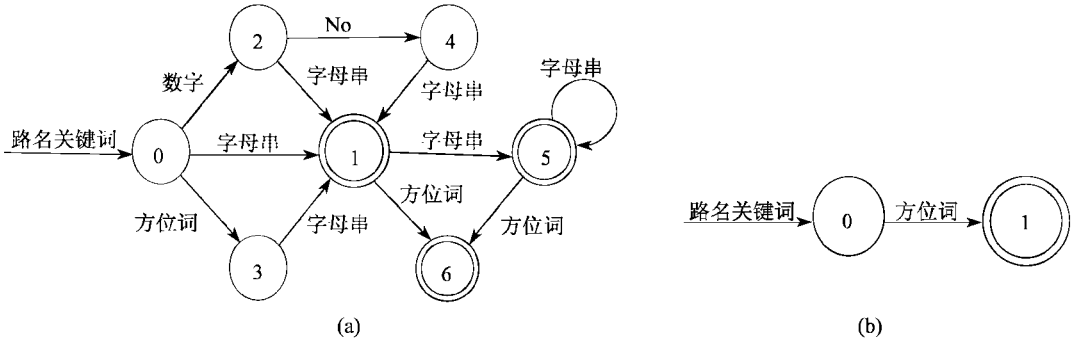

图 3 (a) 确定路名前缀的有限状态自动机 DFA1;(b) 确定后缀的有限状态自动机 DFA2

Fig. 3 (a) DFA1 for the prefix of a road name; (b) DFA2 for the suffix of a road name

将地址识别得到的地址记为 AddrX (Sec_1, Sec_2, …, Sec_9),其中 Sec_j ($1 \leqslant j \leqslant 9$)分别表示地址信息项中邮政编码、市、区、路、小区园区、大楼、号码串、公司和收信人姓名的内容;数据库中某一地址记录记为 DB_k ($Item_1$, $Item_2$, …, $Item_9$, $CItem_1$, $CItem_2$, …, $CItem_9$),其中 $Item_j$ ($1 \leqslant j \leqslant 9$)分别表示地址信息项中邮政编码、市、区、路、小区园区、大楼、号码串、公司和收信人姓名的内容,$CItem_j$ ($1 \leqslant j \leqslant 9$)是对应于 $Item_j$ 的中文表述,$1 \leqslant k \leqslant N$,$N$ 为数据库中的记录总数.

本文基于地址信息项的相似度提出了两种地址相似性的度量方法.

(1) 平均值法

信函地址 AddrX 和数据库中某记录 DB_k 的相似度定义为

$$\varphi_k(\mathrm{AddrX}, \mathrm{DB}_k) = \frac{\sum_{j-1}^{9} \mathrm{Sim} = (\mathrm{Sec}_j, \mathrm{Item}_j)}{m}.$$

其中 m 表示 AddrX 中地址信息项不为空的个数,$\mathrm{Sim}(\mathrm{Sec}_j, \mathrm{Item}_j)$是 Sec_j 和 $Item_j$ 两个字符串之间的相似度. 假设 $\varphi_i = \underset{1 \leqslant k \leqslant N}{\mathrm{Max}} \varphi_k(\mathrm{AddrX}, \mathrm{DB}_k)$. 若 $\varphi_i \geqslant \lambda$ 则认为 AddrX 与 DB_i 相匹配,DB_i ($CItem_1$, $CItem_2$, …, $CItem_9$)为 AddrX 的中文表述,其中 λ 是相似度阈值,取值在 0 ~1 之间.

(2) 改进的平均值法

该方法是对平均值法的改进,目的是减少因为识别原因造成的个别地址信息项的不正确对整个平均取值的影响. 将信函地址 AddrX 和数据库中某记录 DB_k 的相似度定义为

$$\phi_i(\mathrm{AddrX}, \mathrm{DB}_k) = \sum_{j-1}^{9} \mathrm{Sim}(\mathrm{Sec}_j, \mathrm{Item}_j).$$

假设 $\phi_i = \underset{1 \leqslant k \leqslant N}{\mathrm{Max}} \phi_k(\mathrm{AddrX}, \mathrm{DB}_k)$. 改进的平均值法的算法为:

(a) 判断 $\phi_i \geqslant m \times (\lambda + \Delta)$是否成立,($m$ 是 AddrX 中地址信息项不为空的个数,Δ 初始化为 0),如果成立,则认为 AddrX 与 DB_i 是相匹配的. 否则,进入下一步;

(b) 找出 $\mathrm{Sim}(\mathrm{Sec}_1, \mathrm{Item}_1)$, …, $\mathrm{Sim}(\mathrm{Sec}_9, \mathrm{Item}_9)$中非零的最小值 $\mathrm{Sim}(\mathrm{Sec}_h, \mathrm{Item}_h)$,记为 minSim_h;

(c) 若 $\mathrm{minSim}_h \geqslant \mu$,($\mu$ 为控制参数),则认为 AddrX 与 DB_i 是不匹配的. 否则,进入下一步;

(d) 令 $\phi_i = \phi_i - \mathrm{minSim}_h$,将 $\mathrm{Sim}(\mathrm{Sec}_h, \mathrm{Item}_h)$置为 0,$m$ 减少 1,Δ 增加 ν(ν 为控制参

数)，转到(a).

若 AddrX 与 DB_i 相匹配，则 $DB_i(CItem_1, CItem_2, \cdots, CItem_9)$ 是 AddrX 的中文表述.

2　系统实现与实验结果

在 Window 2000 的平台上，采用 C++ 编写算法，实现了本文的地址翻译方法. 根据多次实验结果，取 $\lambda = 0.85$，$\mu = 0.6$，$\nu = 0.01$.

2.1　信函收信人地址的翻译结果

对图 2 的信函地址进行地址识别后得到：AddrX(“200120”，“SHANGHAI”，“PUDONG”，“EAST LUJIAZU1 LU”，“”，“CHINA MERCHANT TOWER”，“161/28F”，“”，“MISS CHEN FENG YU”). 与其相似度最高的地址记录为：DB_i(“200120”，“SHANGHAI”，“PUDONG”，“EAST LUJIAZUI LU”，“”，“CHINA MERCHANT TOWER”，“161/28F”，“”，“MISS CHEN FENG YU”，“200120”，“上海”，“浦东”，“陆家嘴东路”，“”，“中国招商局大厦”，“161/28F”，“”，“陈凤玉女士”). 表 2 列出两个地址对应信息项的相似度.

表 2　AddrX 和 DB_i 对应信息项的相似度

Tab. 2　The similarity between AddrX's address items and DB_i's address items

地址信息项	AddrX	DB_i	相似度
邮编	200120	200120	1.000 0
市	SHANGHAI	SHANGHAI	1.000 0
区	PUDONG	PUDONG	1.000 0
路名	EAST LUJIAZU1 LU	EAST LUJIAZUI LU	0.875 0
大楼	CHINA MERCHANT TOWER	CHINA MERCHANT TOWER	1.000 0
号码	161/28F	161/28F	1.000 0
收信人	MISS CHEN FENG YU	MISS CHEN FENG YU	1.000 0

按照平均值法，可得 $\varphi_i = 0.982\ 1 \geqslant \lambda$，则认为 AddrX 与 DB_i 相匹配，该地址的中文表述为“200120，上海，浦东，陆家嘴东路，中国招商局大厦，161/28F，陈女士”. 按照改进的平均值法，可得 $\phi_i = 6.875 \geqslant \lambda \times 7 = 5.95$ 成立，其结果与平均值法相同.

图 4 给出了从另一封真实信函图像中分割和定位出的收信人地址区域.

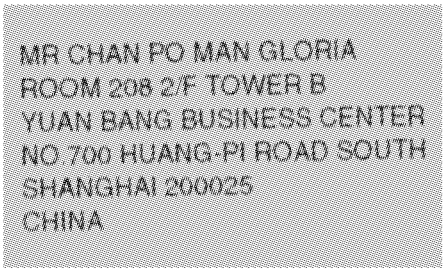

图 4　定位得到的收信人地址区域

Fig. 4　The mail address area from another letter

对图 4 进行字符识别，结果如下：

MRCHANPOMANGLORIA
ROOM2082/FTOWERB
YUANBANGBUS 1 NESSCENTER

NOJOOHUANG-P1ROADSOUTH
SHANGHAI200025
CHINA

显然,该识别结果存在多处错误,“. 700”错识成“JOO”,“PI”错识成“P”“1”,“BUSINESS”错识成“BUS”“l”“NESS”.

对该识别结果进行地址识别后得到:AddrX(“200025”,“SHANGHAI”,“”,“NOJOO HUANG P1 ROAD SOUTH”,“”,“NESS CENTER”,“B/2F/208”,“”,“MR CHAN PO MAN GLORIA”);与其相似度最高的地址记录为:DB_i(“200025”,“SHANGHAI”,“”,“HUANG PI ROAD SOUTH”,“”,“YUAN BANG BUSINESS CENTER”,“700/B/2F/208”,“”,“MR CHAN PO MAN GLORIA”,“200025”,“上海”,“”,“黄陂南路”,“远邦商务中心”,“700/B/2F/208”,“”,“陈先生”). 表 3 列出两个地址的对应信息项的相似度.

表 3 AddrX 和 DB_j 对应信息项的相似度

Tab. 3 The similarity between AddrX's address items and DB_j's address items

地址信息项	AddrX	DB_j	相似度
邮编	200025	200025	1.000 0
市	SHANGHAI	SHANGHAI	1.000 0
路名	NOJOO HUANG P 1 ROAD SOUTH	HUANG PI ROAD SOUTH	0.868 4
大楼	NESS CENTER	YUAN BANG BUSINESS CENTER	0.440 0
号码	B/2F/208	700/B/2F/208	0.666 7
收信人	MR CHAN PO MAN GLORIA	MR CHAN PO MAN GLORIA	1.000 0

按照平均值法,可得 $\varphi_i = 0.829\ 2 \leqslant \lambda$,则认为 AddrX 与 DB_i 不匹配.

按照改进的平均值法,可得 $\phi_i = 4.975\ 1 \leqslant \lambda \times 6 = 5.10$. 其中 $\mathrm{Sim}(\mathrm{Sec}_6, \mathrm{Item}_6)$为非零最小值,即 $\mathrm{minSim}_6 = \mathrm{Sim}(\mathrm{Sec}_6, \mathrm{Item}_6) = 0.440\ 0 \leqslant \mu$,因此重新设置 ϕ_i,Δ,m 和 $\mathrm{Sim}(\mathrm{Sec}_6, \mathrm{Item}_6)$的值,使 $\phi_i = \phi_i - \mathrm{minSim}_6 = 4.537\ 1$,$\Delta = \Delta + \nu = 0.01$,$m = m - 1 = 5$,$\mathrm{Sim}(\mathrm{Sec}_6, \mathrm{Item}_6) = 0$. 由于不等式 $\phi_i = 4.537\ 1 \geqslant m \times (\lambda + \Delta) = 4.3$成立,因此认为 AddrX 与 DB_i 相匹配,AddrX 的中文表述是“200025,上海,黄陂南路,远邦商务中心 700/B/2F/208,陈先生”.

2.2 地址翻译性能比较

采集 500 封真实信函的图像,对分割定位出的收信人地址区域进行 OCR 识别,由地址识别模块将非格式化的文本地址转换为若干地址信息项,采用平均值法,λ 值分别取1.00,0.95,0.90,0.85,0.80,进行 5 次地址翻译,结果见图 5.

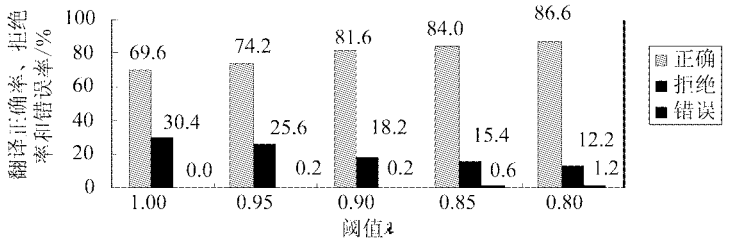

图 5 λ 不同取值的翻译结果的比较

Fig. 5 The comparison of translation performances with different λs

由图 5 可知,λ 越小,系统正确率越高,拒绝率越低,但同时错误率会有所增加. $\lambda = 0.80$ 时的正确率高于 $\lambda = 0.85$时的正确率,但前者的错误率比后者的错误率高出 6 个百分点,因

此λ取0.85较为合理.

对同样的500封图像，分别采用平均值法和改进后的方法进行翻译（λ = 0.85），结果见图6. 由此可见改进后的方法在保证错误率不变的情况下提高了翻译的正确率.

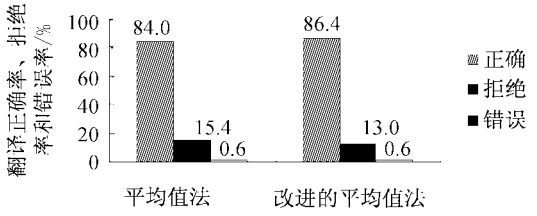

图6　两种地址相似性度量方法的比较

Fig. 6　The comparison of translation performances between two methods

3　小　　结

针对OCR识别结果的地址翻译是图像识别技术与机器翻译技术的有机结合，拓宽了机器翻译技术的应用范围. 本文针对邮政信函的地址自动翻译这一实际应用问题，定义和归纳了地址不同于自然语言的用语规则，并提出了一种利用非精确字符串匹配技术，由关键词引导的地址翻译方法. 利用非精确字符串匹配技术可以有效地减少识别错误带来的影响，在保证错误率较低的情况下正确率达到85%. 实验结果证明了该方法的有效性和可行性. 本文的工作还是初步的，今后的研究方向可以考虑应用基于统计的机器翻译技术，完善地址识别功能，进一步提高系统性能.

[参　考　文　献]

[1]　冯志伟. 机器翻译研究[M]. 北京：中国对外翻译出版社，2004.

[2]　赵铁军. 机器翻译原理[M]. 哈尔滨：哈尔滨工业大学出版社，2000.

[3]　SELLERS P. The theory and computation of evolutionary distance [J]. Pattern Recognition, Journal of algorithms, 1980 (1): 359-373.

[4]　PETER F B, STEPHEN A, DELLA P, et al. The mathematics of statistical machine translation: parameter estimation [J]. Computational Linguistics, 1993, 19(2): 263-311.

[5]　FRANZ J O, HERMANN N. Discriminative training and maximum entropy models for statistical machine translation[C]// Proc of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002). Philadelphia: ACL, 2002: 295-302.

[6]　NAGAO M. A framework of a mechanical translation between Japanese and English by analogy principle[C]// Artificial and Human Intelligence. North-Holland, Amsterdam: NATO Publications, 1984: 173-180.

[7]　杜祝平，吴保民，张连海，等. 英汉机器翻译系统中基于规则的词法分析[J]. 信息工程大学学报，2003，4(3)：89-92.