英文维文人名机器翻译算法的研究与实现*

艾山·吾买尔,吐尔根·伊布拉音
(新疆大学 信息科学与工程学院计算机系,新疆 乌鲁木齐 830046)

摘 要: 本文提出了一种基于规则库的多层过滤进行机器翻译的算法.该算法实现了英文人名向维吾尔文人名的自动翻译.该算法不同于传统的英维人名翻译系统所采取的方法,不需要建立丰富、全面的两种语言人名对齐词库.本算法在对已翻译的大量实例进行统计并分析的基础上,找出了英文与维文人名翻译的三层规则并设计出了本算法.

关键词: 英文人名; 维吾尔人名; 规则库; 机器翻译
中图分类号: TP311.53 文献标识码: A 文章编号: 1000-2839(2007)01-0097-05

Researching and Implementation of the English to Uyghur Personal Name Machine Translation Algorithm

Hasan Umar, Turgun Ibrayim
(*College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang* 830046, *China*)

Abstract This paper presents a new algorithm called Multi-Layer Filtering Rule Based Machine Translation. This algorithm mainly solves the English personal name auto translation to Uyghur personal name. What makes the difference from the traditional translation method is that the new one is just based on rules which obtained by a large scale of statistical work, while the traditional one relies on the dictionary.

Key words English personal name; Uyghur personal name; Rule; Machine Translation; Multi-Layer Filtering.

0 引 言

近年来机器翻译技术取得了很大的进步,但是仍然很不成熟.众所周知,机器翻译有两大主流方向,即基于语料库的方法和基于规则的方法.从算法的实现角度来讲,基于语料的方法比较容易实现,但是需要建立比较完整的语料库,基于规则的方法其难点在于规则不好发现,算法也比较难实现.基于规则的方法与语料的方法,除了实现方法之外,还有一些差别.一般基于语料的机器翻译只能对已出现过的人名进行翻译,而对新出现人名的翻译结果多数情况下比较差,但是基于规则的机器翻译方法不仅把已出现过的人名翻译的很好,而且对新出现人名的翻译结果与基于语料的机器翻译相比准确度高得多.

现有的英维机器翻译方法都是基于语料的,翻译的质量跟语料库的质量及量有直接的关系.这种翻译方法的核心问题是收集及建立比较全面的对齐语料库,实现两种语言的排序和快速查询等.本文中所提出基于规则的多层次过滤算法只需要建立准确的规则库,实现算法的难度在于如何研究英文人名所对应维文人名之间的各种对应关系,从而建立比较准确的规则库,并且根据三个规则的复杂度进行分类及确定优先级.本文根据规则复杂度,把所有的规则分成了三个子库: 单字符对应关系库、双字符对应库和特殊字符对应关系库.

1 统计已翻译的人名与分析翻译规则

英语是由 26个字母来表达所有词汇的语言,维吾尔语也是由 32个字母来表达所有词汇的语言.往往

* 收稿日期: 2006-09-05
基金项目: 国家自然科学基金项目[60263004]资助
作者简介: 艾山·吾买尔(1982-),男,硕士研究生,主要从事信息处理研究.

在很多语言中,人名、地名等名称是根据读音进行翻译的.本算法的出发点也是根据英文人名在维吾尔文中读音找出字符各种组合的对应关系.统计过程如下:从语料库中分离出维文翻译完全相似的英文人名,然后选出所有的只有一个维文字母不同的所有的英文人名,以此类推把语料库根据不同字符的数量分成几个子库,然后从不同字符数量的子库开始分析所不同的字符特征.通过这种方法获取最简单的规则库进行实验性翻译并对翻译结果进行比较与分析.通过对551830个已翻译的人名进行统计与分析,实验最终确定了比较完整的对应关系规则库.这些规则库根据复杂度分成如下三个子规则库:

单字符对应关系库

序号	英文	维文	序号	英文	维文	序号	英文	维文
1	a	ا	10	z	ز	19	n	ن
2	b	ب	11	s	س	20	o	و
3	p	پ	12	c	س	21	u	ۇ
4	t	ت	13	x	س	22	v	ۋ
5	j	ي	14	f	ف	23	w	ۋ
6	q	ق	15	k	ك	24	i	ي
7	h	ھ	16	g	گ	25	y	ي
8	d	د	17	l	ل	26	e	ې
9	r	ر	18	m	م			

双字符对应关系库

序号	英文	维文	序号	英文	维文	序号	英文	维文
1	aa	ا	9	ee	ې	19	oe	ۆ
2	ai	اي	10	eh	ئې	20	oo	و
3	ao	ئاۋ	11	ei	ې	21	ou	ۇ
4	au	و	12	eo	يۇ	22	pp	ف
5	ch	چ	13	eu	ئېي	23	rh	ربخ
6	cs	س	14	gh	غ	24	sh	ش
7	ea	ى	15	ia	يا	25	ue	ۈ
8	ee	ي	16	ih	يخ	26	zh	ج
9	ee	ې	17	je	ژ			
10	eh	ئې	18	ng	ڭ			

特殊字符对应关系库

序号	前后元音	维文
1	S	ز
2	C	ك

在这三个子规则库中,特殊字符对应关系库比较复杂,当在英文人名中出现 s和 c字母时,需要判断它前面的字符和紧跟它后面的字符是不是英语元音字母,如果是元音字母,那么可以直接根据规则库中所对应的关系进行翻译.跟据这三个子规则库的复杂度及优先使用级别,给它们分配了如表中所示的优先级编号.

表 1 三个子规则库优先级表

规则库优先级别	规则库名称
0	特殊字符规则库
1	双字符规则库
2	单字符规则库

2 机器翻译算法的设计

算法的主要目的是把英文人名语料翻译成维吾尔文,其实现步骤如下: 首先根据特殊对应关系库进行过滤并把符合规则的英文字母或字母组合与相应的维吾尔文字母或字母组合进行替换;第二,根据双字符规则库进行过滤并把符合规则的英文字母或字母组合与相应的维吾尔文字母或字母组合进行替换;第三,根据单字符规则库进行过滤并把符合规则的英文字母或字母组合与相应的维吾尔文字母或字母组合进行替换. 一个英文人名通过这三层过滤后的结果就是它的维吾尔文翻译.

以下是本算法的实现:

```
string EUPN_MT(string strEnPersonalName) //英文维文人名机器翻译算法
{
string strTmp= ConvertSuzuk(strEnPersonalName); //根据规则库三进行翻译,在翻译过程只对规则库中有字符进行分析与替换
strTmp= ConvertTwoChar(strTmp); //根据规则库二对没被上一个翻译过程翻译的双字符进行翻译
strTmp= ConvertSingleChar(strTmp); //对所有仍没被替换的字符进行分析与替换
return strtmp;
}
String ConvertSuzuk(string strEnPersonalName); //根据库对前后有元音字符进行替换
String ConvertTwoChar (string strEnPersonalName); //根据库对双字符进行替换
String ConvertSingleChar (string strEnPersonalName); //根据库对单字符进行替换
```

以上三个函数都实现了一层的过滤,而且每个函数都返回通过过滤的字符串. ConvertSingleChar返回的是最终结果.

为了对此规则进行测试,开发出了一个简单的英文维文人名机器翻译系统.以下是翻译系统的流程图:

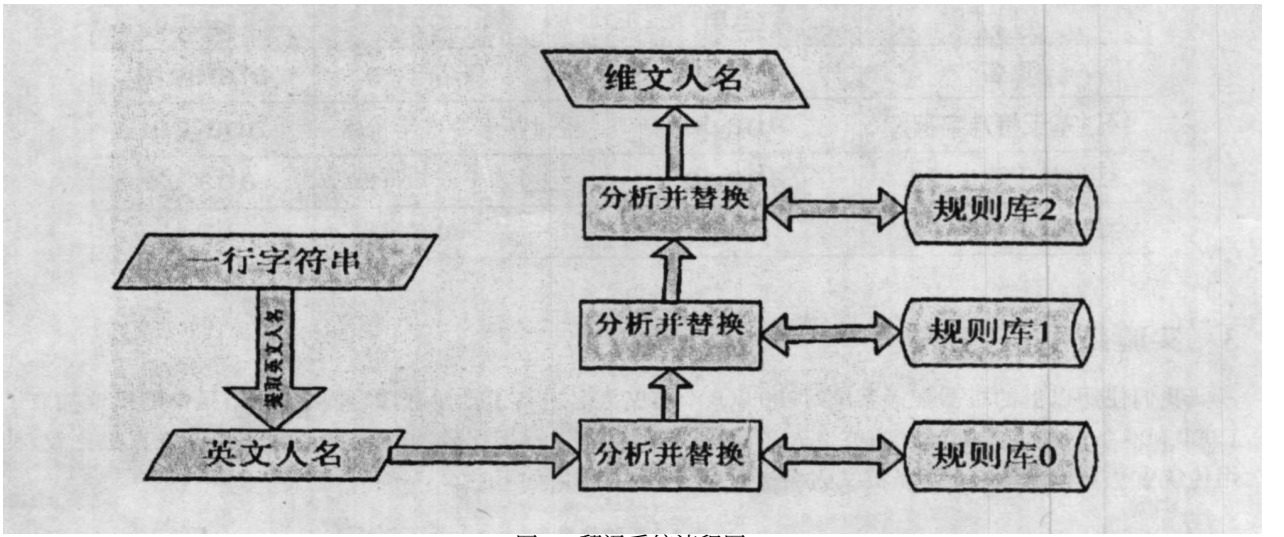
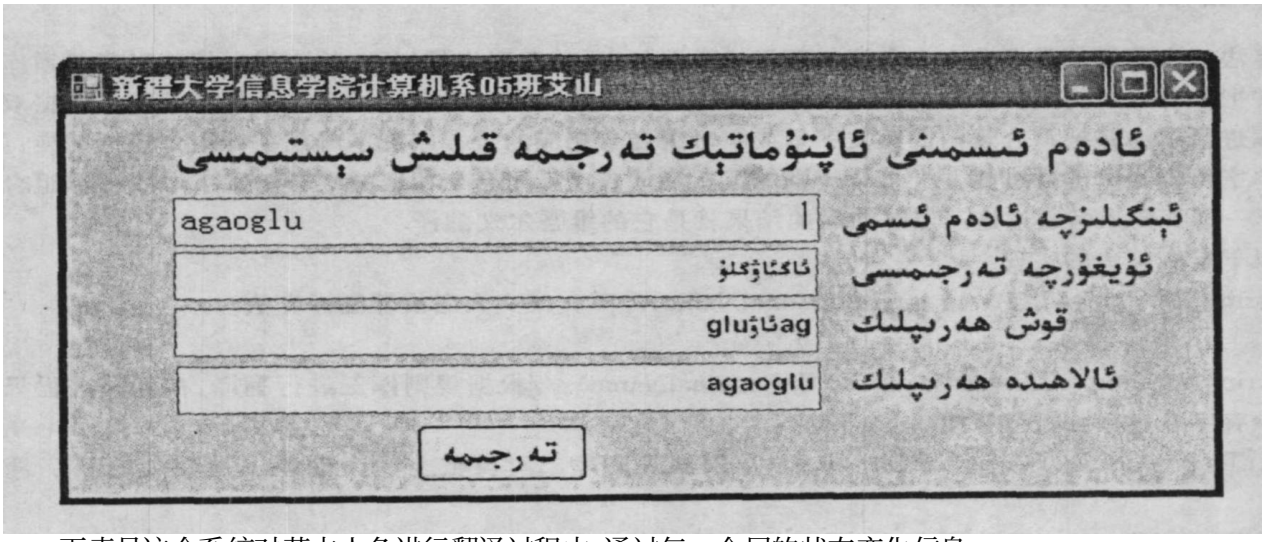


图 1　翻译系统流程图

基于规则库的多层过滤英维人名机器翻译系统界面如下:



下表是这个系统对英文人名进行翻译过程中,通过每一个层的状态变化信息.

过程	结果	过程	结果
原名	abbagnara	原名	agaoglu
经过基于特殊字符库	abbagnara	经过基于特殊字符库	agaoglu
经过基于双字符库	abbagnara	经过基于双字符库	agۆئاglu
经过基于单字符库	ئابباگنارا	经过基于单字符库	ئاگائاۆگلو
过程	结果	过程	结果
原名	abaco	原名	abaseal
经过基于特殊字符库	abaەo	经过基于特殊字符库	abaseal
经过基于双字符库	abaەo	经过基于双字符库	abaازی
经过基于单字符库	ئاباكو	经过基于单字符库	ئاباسىل

3 实验结果

我们使用以上的机器翻译系统对700 000个英文人名进行了批量翻译,并从翻译结果中随机地抽取了1 000 000条已被人工翻译过的英文人名翻译,与语料库进行了比较.除此之外,为了跟传统方法比较,使用传统方法对以上700 000条英文人名进行翻译.实验结果如下:

表 2 实验结果表

基于语料库的传统方法	基于规则库的多层过滤方法
56.10%	86.60%

目前语料库是基于上世纪90年代人文科学界人士手工进行翻译的资料为基础而建立的.可是,在我们进行测试时使用文档是包含了新的人名,这就是导致这个差别的原因所在.从实验结果可得,基于语料的翻译系统往往落后于现在,它需要不断地去补充内容,基于规则的翻译系统虽然实现起来难度高,可是效率和生命周期远远高于基于语料库的翻译系统.

4 结　论

人物名称翻译是翻译工作过程中常见的、而且对准确率要求较高的一项工作.目前,已开发的各种词典都是基于词库而实现翻译,所以这种产品只能对已有的人物名称进行翻译,但是这种产品无法满足现代社会的情报信息翻译、期刊翻译、图书翻译等工作的需求.利用基于规则的多层过滤机器翻译算法的系统不仅很准确的翻译已有的人名,也能较准确的翻译未来可能出现的英文人名.因此给翻译行业带来了很大的便利.

参考文献:

[1]吐尔根·伊布拉音,阿不力米提·阿布都热依木.维汉机器翻译词典的结构设计与实现[J].新疆大学学报,2005,22(3):258~262.
[2]艾尼瓦尔·麦麦提,吐尔根·伊布拉音.维吾尔文字母频率统计与应用[C].计算机语言学研讨会论文集,2004年8月.
[3]严蔚敏,吴伟民.数据结构[M].清华大学出版社,1992年6月.
[4]周亚,宗成庆,徐波.基于多层过滤的统计机器翻译[C].计算机语言学研讨会论文集,2004年8月.
[5]段慧明,俞士文.机器翻译评测报告,计算机世界报,1996,12 183.
[6] Martin Kay, Unification in Grammar, In V. Dahi & P. Saint-Dizeir(Ed.) Natural Language Understanding and Logic Programming, North Holland, Amsterdam, The Netherlands.
[7]俞士文.关于计算语言学的若干研究[J].语言与文字应用,1993,(3):55-64.
[8]颜伟,荀恩东,基于WordNet的英语词语相似度计算[C].计算机语言学研讨会论文集,2004年8月.

责任编辑: 闫新云

(上接第80页)

参考文献:

[1]刘忠渊,张富春,王芸,等.昆虫抗冻蛋白的研究[J].生物技术, 2004, 14(3): 73-75.
[2] Graether SP, Sykes BD. Cold survival in freeze-intolerant insects: The structure and function of β-helical antifreeze proteins[J]. Biochemistry, 2004, 271: 3285-3296.
[3] Zhang DQ, Liu B, Feng DR, He YM, and Wang JF. Expression, purification, and antifreeze activity of carrot antifreeze protein and its mutants[J]. Protein Expression and Purification, 2004, 35: 257-263.
[4] Chao H, Davies PL, Carpenter JF. Effects of antifreeze proteins on red blood cell survival during cryopreservation[J]. The Journal of Experimental Biology, 1996, 199: 2071-2076.
[5] Gabriel A, Liana H, Boris R et al. Subzero nonfreezing cryopreservation of rat hearts using antifreeze protein I and antifreeze protein III[J]. Cryobiology, 2004, 48: 273-282.
[6] Tao H, Jessie N, Daniel GZ. Expression of an insect (Dendroides canadensis) antifreeze protein in Arabidopsis thaliana results in a decrease in plant freezing temperature[J]. Plant Molecular Biology, 2002, 50: 333-344.
[7]赵干,马纪,薛娜,等.新疆准噶尔小胸鳖甲抗冻蛋白基因的克隆和抗冻活性分析[J].昆虫学报. 2005, 48(5): 667-673.

责任编辑: 周　蓉