英文-维吾尔文人名机器翻译算法的研究与实现[1]

艾山·吾买尔，吐尔根·依布拉音

新疆大学信息科学与工程学院　乌鲁木齐，830046

Email: Hasan1479@163.com　Turgun@xju.edu.cn

摘要： 本文提出了一种基于规则的多层过滤进行机器翻译的方法。该方法实现了英语人名向维吾尔语的自动翻译。该方法不同于传统的机器翻译方法，它不需要建立丰富，完整的词库。本机器翻译系统对现有的已翻译的英文人名进行大量的统计并分析的基础上，找出英语人名跟翻译成维文的名称之间的关系。根据这些关系的复杂程度，将把规则分了三个层次，从而实现了基于规则的多层次过滤机器翻译系统。初步的试验结果表明该系统性能较好，而且使用此方法实现的英维人名机器翻译系统，跟基于词库实现的机器翻译系统做对比，结果表明基于规则的多层次过滤机器翻译系统明显提高了翻译水平。

关键词： 英文人名，维吾尔人名，规则库，机器翻译，多层过滤。

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

Hasan UMAR Turgun IBRAYIM

Abstract: In this paper we propose a new algorithm called Multi-layer filtering Based Rule manchine translation.This algorithm mainly solves the English personal name auto translation to Uighur language.What makes the difference from the traditional machine translation algorithm is that the new one just based on rules which obtained by a large number of statistical work,while the traditional one relies on the name dictionary.There is lots of rules between the English name and Uighur translation,and we divided them into three groups according to their complexity.The preliminary experimental results express that our algorithm achieves a good performance in doing the translation.Moreover,the translations generated by this algorithm are much better than the results generated by the old one.

Key Words: English personal name, Uighur personal name, rule database, machine translation,Multi-layer filtering.

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与基于规则的方法。从算法的实现角度来讲，基于语料的

[1]作者简介：艾山·吾买尔，男，1981 年生，新疆大学信息科学与工程学院计算机系硕士研究生，研究方向为信息处理。吐尔根·伊布拉音，男，1958 年生，新疆大学信息科学与工程学院副院长，教授，博士生导师。研究方向为计算机应用，自然语言处理，软件工程。

方法是比较容易实现，但是需要建立比较完整的语料库，基于规则的方法的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的过的人名进行翻译，而对新出现人名的翻译结果多数情况下并不太好，但是基于规则机器翻译方法不仅能对已出现过的人名进行很好的翻译，而且对新出现的人名翻译也比基于预料的方法更准确。

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有直接的关系。这种翻译算法的核心问题是实现两种语言的排序和快速查询。本文中所提出的基于规则的多层次过滤算法只需很少的规则库，实现算法的难点在于如何研究英文人名跟维文翻译之间的各种对应关系，从而确定规则，并根据规则的复杂度，优先级等进行分类。根据规则的复杂度优先级，我们把规则库分成三个子库，单字符对应关系库，双字符对应关系库，特殊对应关系库。

本文第一部分给出了统计已有语料并分析规则；第二部分给出了机器翻译算法的设计；第三部分给出了初步的实验结果，最后给出了相关结论与展望。

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字母来表达所有词汇的语言。往往在很多语言中，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人名中出现的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小库，然后从不同字符数量小的子库开始分析所不同字符的特征。通过这种方法获取的最简单的规则库进行实验性翻译并对结果进行比较与分析。通过对551830个已经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：

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	ۈ
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			

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

双字符对应关系库

序号	前后元音	维文
1	S	ز
2	C	ك

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	但字符规则库

特殊对应关系库

在这三个库中第三个库比较特殊，当在英文人名中出现 s 和 c 字符时我们要判断它前面的字符与紧跟它后面的字符是不是英语元音字母，如果是那么我们可以根据表中所显示的对应关系进行翻译。因为这三个规则库的优先考虑的级别不同，我给这三个库分配了从零开始的优先级编号

2. 机器翻译算法的设计

算法的主要目的是把提供的英文人名语料翻译成维吾尔文。算法的实现步骤如下：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。一个英文人名通过了这个三层的过滤后的结果就是它的维吾尔文翻译。

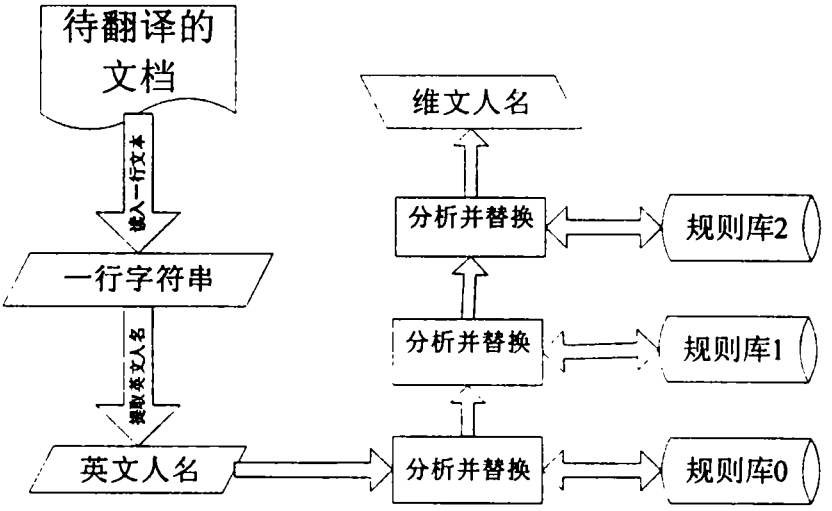
```
string EUPN_MT(string strEnPersonalName)//English Uighur personal Name
                                          Machine //Translatio
 {
string strTmp=ConvertSuzuk(strEnPersonalName);//根据规则库三进行翻译，在翻译
                                                过程
                                              //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。

为了对我们所确定的规则进行测试，简单使用 C#的实现了一个批量英文人名翻译系统，这个系统从指定的文本文件以行为单位进行翻译。

以下是翻译系统的的流程图：



以下表图所示的是这个系统对某个英文人名进行翻译过程中通过每个过滤层的状态变化信息。

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	agۇئاglu
经过基于单字符库	ئابباگنارا	经过基于单字符库	ئاگئاۇگلۇ
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	abaجeal
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	ئاباژل

3. 实验结果

我们使用以上所说的人名机器翻译系统对 700000 个英文人名进行了翻译，并从翻译结果中随机的抽取了 100000 条人名的翻译，进行了人工测试。为了跟传统的基于语料库进行翻译的方法做对比，我们还使用传统的方法进行了翻译以人工测试。以下是实验结果准确度表。

基于传统方法准确度	基于规则的方法准确度
56.1%	86.6

为什么会有这么大的区别，因为目前我们有的英维文语料库是 90 年代初被人文科学界进行翻译的，内容比较老，在我们测试时用的文档内容是包含很多新的人名的。可是，我们从实验结果发现基于规则人名翻译系统还有一个比较明显的缺陷。

4. 结论

人物名称翻译是在各种翻译过程常见的而且对准确率要求比较高的一项工作。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，单是这种产品无法满足现代社会上的情报信息翻译，期刊翻译，图书翻译等行业的需求。利用本算法实现的应用程序不仅可以很准确的翻译已有的人名，也能比较准取得翻译未来有可能的出现大多数英文人名。因此能给翻译行业带来很大的便利。

参考文献

[1] 新疆大学学报 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 2005 年 8 月

[2] 计算机语言学研讨会论文集 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 2004 年 8 月

[3] 计算机语言学研讨会论文集严蔚敏 吴伟民 清华大学出版社 1992 年 6 月

[4] 计算机语言学研讨会论文集 周亚 宗成庆 徐波 基于多层过滤得统计机器翻译 2004 年 8 月

[5] 段慧明，俞士汶，机器翻译评测报告，《计算机世界》报 1996 年 3 月 25 日，第 183 页

[6] Martin Kay, Unification in Grammar, In V.Dahi & P.Saint-Dizier(Ed.) Natural Language Understanding and Logic Programming, North Holland, Amsterdam, The Netherlands.

[7] 俞士汶，关于计算语言学的若干研究，语言与文字应用，1993 年第 3 期

[8] 计算机语言学研讨会论文集 颜伟 荀恩东 基于 WordNet 的英语词语相似度计算 2004 年 8 月