汉文专有名词藏文音译的研究与实现

梁会方,黄鹤鸣,杨 峰
(青海师范大学 计算机学院,青海 西宁 810000)

摘 要:汉藏语言的文化交流,少不了语言之间的互译。为了汉藏音译规范化,提出了一种基于规则的汉藏音译方法,根据目前汉藏的音译情况以及汉藏拼音相似性制定了汉藏音译的规则集—汉文对应的拼音和拼音相应的藏文对照表。对于一个汉文存在多个拼音的情况,则要采用统计的方法,依赖上下文相关的词组等选取合适的拼音,然后再根据规则集翻译出所对应的藏文。在音译算法上,文中在汉藏音译的规则制定的基础上,对于存在的约定俗成译法词组优先处理,以及汉文的多音字结合了统计的多音字语料词组,提高音译系统的性能以及其音译的准确性。该算法实现简单,准确率高。

关键词:音译;汉藏音译;规则
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2014)12-0192-04
doi:10.3969/j.issn.1673-629X.2014.12.045

Research and Implementation of Chinese-Tibetan Transliteration about Proper Noun

LIANG Hui-fang, HUANG He-ming, YANG Feng
(College of Computer, Qinghai Normal University, Xining 810000, China)

Abstract: The cultural exchange of Chinese and Tibetan language, translation between languages is essential. In order to be standardization of transliteration, propose a Tibetan transliteration method based on rules, according to the current Chinese and Tibetan transliteration as well as the similarity of Chinese and Tibetan, constitute the rule set about Chinese-Tibetan Pinyin transliteration, that is Chinese Pinyin and corresponding Tibetan table. For situations where there are multiple phonetic Chinese are using statistical methods, relying on context sensitive phrases to choose the appropriate Pinyin, then according to the rule set to translate into Tibetan. On the transliteration algorithm, on the basis of rule-making in Chinese and Tibetan transliteration, for priority conventional translation of phrases, as well as an understanding of how polyphone combines statistical polyphone corpus phrases, improve the accuracy and the performance for its phonetic transliteration system. The algorithm is simple and accurate.

Key words: translation; Chinese-Tibetan transliteration; rules

0 引 言

藏文信息处理一直是少数民族语言研究的重要部分,随着信息技术的飞速发展,在长期的汉藏文化交流中,藏民族曾从汉语中引进了许多词汇,随着科学技术的进一步发展,新事物不断的涌现,那么从汉语中借用的新词将不断增加,这些新词大多直接音译使用,但对同一新词,不同的人可能有不同的音译结果。虽然我国汉藏翻译工作很早就开始了,但是由于藏族居住分布较为广泛,导致藏语在语言上具有地域差异性,主要是藏语的方言差异。藏族主要有三大方言区,它们是康巴、安多以及前后藏三大方言区,而各大方言区又有很多小的方言区[1],且经过多年的变迁,发音也发生了一定的变化,这就导致了音译同一汉字时出现不同的藏文,而且汉藏译史已经有千年历史,很多名词已经存在约定俗成的译法,如果对此不了解,可能导致一词多种翻译的问题。并且汉藏音译中不能形成用词统一的标准,故存在音译上的混乱[1-2]。因此制定统一的音译规范对于汉藏文化的信息交流具有重要意义。在信息处理自动化的时代,则要快速准确地实现汉藏的音译,提高工作效率,针对此研究并实现汉文专有名词基于规则的汉藏音译。

目前有关汉英的音译研究,以及汉文对其他语言的音译尤其是人名、地名的音译[3-4]问题的研究比较

收稿日期:2014-01-10 **修回日期**:2014-04-15 **网络出版时间**:2014-10-23
基金项目:国家自然科学基金资助项目(60963016);青海省普通高等学校研究生创新研究项目
作者简介:梁会方(1988-),女,湖北天门人,硕士研究生,CCF 会员,研究方向为藏文信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1047.003.html

深入,其他语言的专名识别,例如泰语的专名识别研究是基于特征,侧重于专名的外部信息,为了从语料中提取所需的特征,就要利用机器学习算法来学习和训练。利用特征的专名识别方法可以识别已知词或者未知串构成的专有名词[5-6]。这些方法总结如下:

(1)基于统计的方法。建立语料库,用统计模型对语料库进行训练,获取其上下文关系,然后设定其阈值判定。该方法在中英文人名识别等方面得到了较好的使用[7-10]。但是由于该方法依赖语料库的完整性,因此,需要不断完善语料库。

(2)基于规则的方法。该方法要根据音译的特征建立规则集,该方法准确率极高,使用范围广[11],但是规则集的建立需要大量人工对这些规则进行归纳总结,而且规则难于发现,算法实现也困难。

(3)规则和统计相结合的方法。一般运用统计的方法建立语言模型,降低了手工建立规则的复杂性,结合规则的方法可以减少统计的方法对语料库的依赖性,统计与规则的方法互补在藏文识别和分词中能得到较好的效果[12]。

根据汉藏音译的特点,总结出汉藏音译采用基于规则的方法更合适,研究汉文专有名词和藏文音译之间的对应关系,并确定规则,建立汉藏音译的规则库,根据此规则库集来实现算法。实验结果表明,算法实现简洁轻便,对系统资源要求低,且准确度高。

1　汉藏音译规则

汉藏音译规则的建立依赖大量的藏文专家以及学者,翻译工作者的考察、整理,汉藏音译转写规则主要总结为以下几点:

(1)在使用藏文字上,尽量选用目前还没有实际意义的藏文字符,尽量避免使用已经具备了一定汉语意义的藏文字符,避免歧义性。

(2)汉文字符应该输入为简体汉字,即汉文字符编码的常用汉字,避免输入不在字符编码范围的字符。

(3)目前汉藏音译混乱,其中之一就是发音不准,而音译是根据拼音进行翻译的,这就导致很多同一名字多种译法。因此汉文字符发音应以普通话为准则,即要同音同字,相同发音的汉文名字采用相同的藏文字符对应音译,而且不考虑汉文拼音中的音调。

(4)对于已经存在约定俗成译法的词语尽量保留原译法,即很多汉文对应的藏文翻译已经沿用很久,普遍被大众接受的,虽然在译法上不符合音译规则库,但是仍然保持该译法。

上述规则制定了汉藏音译时可能出现的情况,有很好的指导意义,使音译规范化、标准化。即后续的音译算法实现过程以此音译规则为准则。汉藏音译对应关系即汉藏音译规则库的建立依照上述规则。

汉藏音译规则库,汉文拼音对应的藏文,汉藏音译拼音对应关系其中一部分如图 1 所示。

	b	p	m	f	d	t	n	l	g	k
a	བ	པ	མ	ཕ	ད	ཏ	ན	ལ	ག	ཁ
o	བོ	པོ	མོ	ཕོ						
e			ཨེ		ད	ཏ	ན	ལ	ག	ཁ
-i										
ai	བའི	པའི	མའི		དའི	ཏའི	ནའི	ལའི	གའི	ཁའི
ei	བེ	པ	མ	ཕ	དེ		ནེ	ལ	གེ	ཁ
ao	བའོ	པའོ	མའོ		དའོ	ཏའོ	ནའོ	ལའོ	གའོ	ཁའོ

图 1　汉藏音译对照表(部分)

该表是在遵循以上规则的情况下建立的,表中根据汉语拼音的声母和韵母划分音节,该表中包含了汉文中常出现的 400 多个音节,这些拼音包括了现在常用汉字的所有拼音,可以实现目前常用汉语拼音的藏文音译,即可以实现现有的和未来将出现的所有汉文人名、地名的准确藏文音译。假如使用语料库的方法,则需要建立上万条对应的语料词条,而且还需要不断更新完善,增加新出现的人名、地名等专有名词新词条。而对于过分依赖上下文的方法,先采用上下文相关标示出了可能的专名,然后再根据更加宽松的上下文规则判定出一部分,最后根据规则库识别出最后的一部分,侧重外部信息,依赖性强,而过于依赖外部信息容易导致系统不稳定。

目前汉藏存在的约定俗成译法的词语,不能单个字处理,需要根据上下文,判断该词组是否已经存在译法,给出其相应的音译。例如“青海省”,其中“青海”,人们已经有了其对应藏文“ མཚོ་སྔོན་ ”,如果根据上面的规则,则是“ ཆིང་ཧའེ་ ”;为了方便人们的使用,对于这些在藏区已经存在俗成译法的词语,要保留原译法。对其进行统计,对这些词语制定了特殊对应关系库,其中部分统计的结果如表 1 所示。

表 1　约定俗成的汉藏音译举例(部分)

汉文词	音译结果
青海	མཚོ་སྔོན་ཞིང་།
西藏	བོད་རང་སྐྱོང་ལྗོངས
四川	སི་ཁྲོན་ཞིང་།
云南	ཡུན་ནན་ཞིང་

对于约定俗成译法的词语,这里主要统计了地名,大部分是使用藏语的地区,保留原译法,在后面算法上

既要保留这些词语的译法,又与音译系统不矛盾。

由于汉文存在多音字,并且遵循的音译规则主要是根据汉文的拼音对应进行翻译的,所以汉字输入如果存在多个音节就要进行音节选取了。首先选取的时候根据大众习惯,例如“强”,这个字在名字中出现时,普遍读作“qiang”而不应按其另一音节“jiang”来进行音译。再次根据上下文字符,选取当时合适的音节,例如“柏乡县”,在这个地名里,“柏”应该读作“bo”,而“柏油马路”里“柏”则对应拼音“bai”。最后确定音节后都按照音译规则输出藏文字符。

2 算法实现

首先在藏文信息处理实现要以 Unicode 编码标准为基础,使藏文字符集可以正常显示。汉藏音译原理主要是汉文字符转换为其对应的汉语拼音,然后由拼音翻译其对应的藏文字符,其中准确的拼音是关键,汉文字符对应的拼音,系统采用的汉语拼音词典包括了现在常用汉字六千多个,包括了现在常用人名、地名的用字。能够正确给出系统输入的对应汉文的汉语拼音,系统能根据该词典很好地获取汉文字符的拼音字符,保证后续音译藏文的准确性。

首先考虑到很多已经存在约定俗成译法的词语,对这些词进行排除处理,根据上面设定的汉藏音译对照关系进行相关汉文字符的音译。汉藏音译技术原理:获取汉文字符音节,对于多个音节选取合适音节,按照汉藏音译对应规则表给出汉文音译的藏文。算法描述如下:

Step1:获取输入的汉文字符,判断输入的词语是否在统计的已经存在译法的特殊词语表中,如果是,查询特殊藏文词语库转到 Step4;

Step2:对输入汉文字符进行单个字符处理,并且从第一个字符开始进行,判断当前字符是否是最后一个字符,如果是,获取拼音及其对应的藏文字符,跳转到 Step4;

Step3:根据汉语拼音词典获得汉文字符拼音,如果是多音字,选取其合适拼音,然后再根据其拼音,按照图 1 中所示的拼音到藏语的对应关系,得到相应的藏文字符,添加藏文音节符,重复此过程直到最后一个字符;

Step4:输出获取的藏文字符串和结束符。

算法流程图如图 2 所示。

该算法非常有效地处理了汉文音译为藏文的过程,算法不依赖预先建立的语料库,对已有的和将来可能出现的专有名词的翻译结果都将是准确的。在此算法中,特殊的词语处理方法—即约定俗成音译的词语表,能很好地避免音译的多异性。如:“四川”,俗成的译法是“ཞི་ཁྲོན་”;“青海”—“མཚོ་སྔོན་”;如果遵循规则的音译过程,如:“湖北”,分解该词为“湖”和“北”,其对应一个个字符处理,“湖”—“hu”—“ཧུ་”、“北”—“北”—“བེ་”,每个输入词组音译结束最后输出结束符“|”。

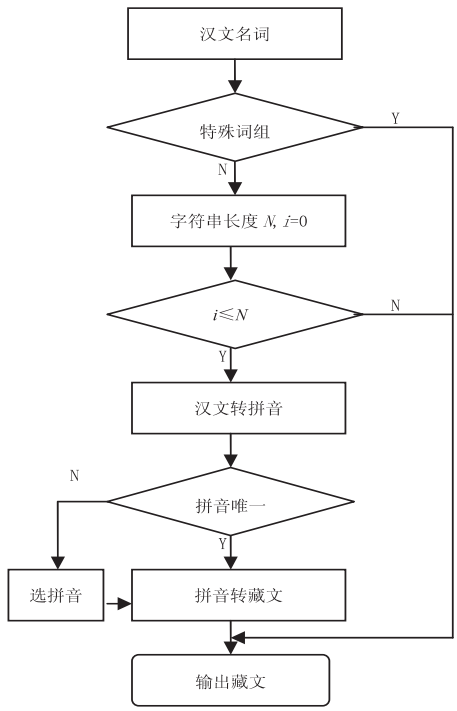

图 2 汉文专有名词藏文音译流程图

3 音译实现和结果分析

运用了基于规则的方法,并结合上述算法,采用 Visual Studio 2010 开发了汉藏音译系统,该系统采用 C++编写,结合后台数据库 Access 2003,运行的系统平台为 Windows 7。对现在存在的一万多个人名进行了批量测试,翻译结果符合上述规则库。对于特殊的词组、字符,有代表性的汉文字符,也进行了测试,符合特殊词语库的规则。如果将特殊词语库和汉文的多音字的规则进行完善还可以提高准确率。除了专有名词的特殊词语库,还需要统计其他类型的词组。

该系统选用字符集符合藏文字符国际标准,藏文字符显示选择微软目前的喜马拉雅字体。汉藏音译系统部分运行结果举例如表 2 所示。下面实例主要包括约定俗成特例、多音字和常规翻译词。

4 结束语

在汉藏翻译中,人名、地名等专有名词的翻译很常见,而且对翻译的要求也相当高。目前,很多音译的实现是基于预先建立的语料库实现人名、地名的翻译,此方法针对已经存在的名称进行翻译,而对新出现的名称需要进行语料库的完善[13]。文中深入研究了汉藏

表 2　汉藏音译结果部分实例

汉文词	转换对象	音译结果
青海省	青海'sheng	མཚོ་སྔོན་ཞེང་ཅིང།
福建省	fu'jian'sheng	སྤུ་ཇན་ཅན་ཅིང་
东莞	dong'guan	ཏུང་ཀོན།
柏林	bo'lin	སྤོ་ལིན།
柏树	bai'shu	པའེ་ཤུའུ།
李克强	li'ke'qiang	ལི་ཁེ་ཆང་
梁会方	liang'hui'fang	ལྷང་ཧུའེ་ཐྭང

音译的规律,提出了基于规则的音译算法。该规则包括通俗译法的词语以及常用汉文字符拼音对应的藏文字符。实验结果表明,该算法不需要依赖预先建立的语料库,对于目前已经有的和将来出现的人名、地名的翻译结果都会是准确的,而且具有普遍性。该规则不仅仅适用于人名、地名的音译,对非专用汉文字符对藏文的音译,该规则都适用,只是对于其他的专用词语需要完善前面约定俗成的词语库规则,并在长句中,存在俗成译法词语组合的拆分。该音译系统的实现给现在信息翻译带来了很大的便利,满足很多期刊、媒体等汉文的音译的需求。并且都遵循上述音译对照表以及规则,能让大众音译规范化。

参考文献:

[1] 巴松拉姆. 关于汉藏翻译中音译规范化问题研究[J]. 剑南文学:下半月,2011(1):93-93.

[2] 多杰太. 关于汉藏翻译中音译规范化问题[J]. 青海民族学院学报(社会科学版),2007,33(1):151-153.

[3] Kuo J S,Li Haizhou,Yang Y K. A phonetic similarity model for automatic extraction of transliteration pairs[J]. ACM Transactions on Asian Language Information Processing,2007,6(2):1-24.

[4] Karimi S,Scholer F,Turpin A. Collapsed consonant and vowel models:new approaches for English-Persian transliteration and back-transliteration[C]//Proceedings of the 45th annual meeting of the association of computational linguistics. [s. l.]:[s. n.],2007:648-655.

[5] Charoenpornsawat P,Kijsirikul B,Meknavin S. Feature-based proper name identification in Thai[C]//Proceedings of national computer science and engineering conference'1998. Thailand:[s. n.],1998.

[6] Charoenpomsawat P,Kijsirikul B. Feature-based Thai unknown word boundary identification using winnow[C]//Proceedings of the 1998 IEEE Asia-Pacific conference on circuits and systems. Thailand:IEEE,1998.

[7] 邹　波,赵　军. 英汉人名音译方法研究[C]//第四届全国学生计算语言学研讨会会议论文集. 出版地不详:出版者不详,2008.

[8] 赵明明. 英汉命名实体翻译方法研究[D]. 苏州:苏州大学,2011.

[9] 周美玲. 英汉人名音译方法的研究与实现[D]. 苏州:苏州大学,2009.

[10] 钱　晶. 汉语专名识别与音译方法研究[D]. 上海:复旦大学,2006.

[11] 牛小莉,谢新卫. 谈《维吾尔人名汉字音译转写规则》的重要意义[J]. 语言与翻译,2003(2):7-10.

[12] 窦　嵘,加羊吉,黄　伟. 统计与规则相结合的藏文人名自动识别研究[J]. 长春工程学院学报(自然科学版),2010,11(2):113-115.

[13] 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,2010,27(8):86-87.

(上接第 191 页)

[3] Zheng Junjie. An effective submarine detection technology:underwater sensor networks[C]//Proc of 2011 IEEE international conference on information theory and information security. Hangzhou:IEEE,2011:417-421.

[4] 郑君杰,李延斌,尹　路,等. 水下传感器网络系统架构与体系结构研究[J]. 计算机科学,2013,40(06A):251-254.

[5] 郑君杰,刘志华,刘　凤,等. 基于水下三维传感器网络的海洋环境立体监测系统关键技术研究[J]. 海洋技术,2012,31(4):1-4.

[6] Gao T, Greenspan D, Welsh M,et al. Vital signs monitoring and patient tracking over a wireless network[C]//Proceedings of the 27th IEEE annual international conference on EMBS. [s. l.]:IEEE,2005:102-105.

[7] 郑德忠,韩昭明,王　聪,等. 基于无线传感器网络的 CO 监测系统设计[J]. 传感技术学报,2007,20(4):925-928.

[8] 江　杰,张云飞. 基于无线传感器网络的水文监测系统[J]. 工业控制计算机,2011,24(7):68-70.

[9] 赵　明,徐科军,倪　伟,等. 一种无线传感器网络节点设计和通信协议研究[J]. 仪器仪表学报,2005,26(z2):630-632.

[10] 卢　崇,马建仓,王吉富. 基于 ATmega128L 与 CC2420 的无线传感器网络节点的研究与实现[J]. 电子技术应用,2006,32(12):130-133.

[11] 雷昌有,蒋　英,史东华. 北斗卫星通信在水情自动测报系统中的研究与应用[J]. 水利水电快报,2005,26(21):26-28.

[12] 黄小波. 基于 GPRS 的无线传感器网络网关的设计与实现[J]. 自动化应用,2010(7):57-59.

[13] 周延年,叶　松,郑君杰,等. 利用流星余迹通信系统传输海洋数据设计[J]. 仪器仪表学报,2008,29(8):486-489.

[14] 卡　勒. 无线传感器网络协议与体系结构[M]. 邱天爽,译. 北京:电子工业出版社,2007.

汉文专有名词藏文音译的研究与实现

作者： 梁会方， 黄鹤鸣， 杨峰， LIANG Hui-fang， HUANG He-ming， YANG Feng
作者单位： 青海师范大学 计算机学院,青海 西宁,810000
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(12)