基于转换规则的汉文—维文专有名词自动翻译研究

塞麦提·麦麦提敏[1]　亚森·伊明[2]

[1]中国传媒大学　北京　100024　新疆大学人文学院　乌鲁木齐　830046
[2]新疆维吾尔自治区民语委　乌鲁木齐　830000

摘　要：本文针对真实文本中出现最为频繁的人名、地名、机构名等三种专有名词，提出了一种基于转换规则的专有名词自动翻译方法。该方法根据汉语和维吾尔语的特点，将翻译过程分为三个阶段，从而实现汉语专有名词向维吾尔语的自动翻译。其不同于传统的机器翻译方法，不需要建立丰富、完整的双语词库。实验结果表明：该方法的准确率达到 90.5%，从而证明了基于转换规则的专有名词自动翻译方法的有效性，而且基于该方法的专有名词自动翻译子系统可以运用到跨语言信息检索(CLIR)、机器翻译(MT)和问答系统等多语言信息处理应用领域之中。

关键词：汉文；维吾尔文；专有名词；转换规则；机器翻译

Rule-based Approach on Chinese-Uyghur NE Machine Translation

Saimaiti Maimaitimin[1]　Yasin Imin[2]

[1] Applied Linguistics Department, Communication University of China, Beijing, 100024
[2] The Ethnic Languages and Scripts Committee of XUAR, Urumqi, 830000

Abstract: In this paper, a rule-based approach on machine translation of most frequent proper names such as person name (PER), location name (LOC) and organization name (ORG) is introduced. The approach translates NE from Chinese to Uyghur by following three-step transformation, and it is dictionary-independent and highly efficient. The experiment shows very encouraging result, we achieved a precision rate of 90.5% for Chinese-Uyghur NE translation. Machine translation systems can use such a rule-base NE translation system as a component to handle phrase translation in order to improve overall translation quality. Cross-Language Information Retrieval (CLIR) systems could identify relevant documents based on translations of NE provided by such a sub-system.

Keywords: Chinese; Uyghur; NE; Transformation Rule; Machine Translation

1　引言

在跨语言信息检索（CLIR）、机器翻译（MT）和问答系统等多语言信息处理应用领域，专有名词的自动翻译是系统的重要组成部分。这些系统必须解决专有名词的识别和翻译。国外专有名词的自动翻译研究已取得了可喜的成果。Jarmo Toivonen[2]等人提出了基于两步骤的专业术语和专有名词的翻译方法，Paola Virga, Sanjeev Khudanpur[4]等人提出了在跨语言信息处理中运用的基于统计的专有名词翻译方法，Chen, Hsin-His[1]等人又提出了基于语音匹配模型的英汉专有名词自动获取的方法。他们的研究结果表明，西班牙语、英语等同类语言的专有名词翻译已不成问题，但是使用不同语音和文字系统的语言之间的专用名词的翻译还需要进

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

一步探索。

在国内，中文专有名词的识别研究有了一定的进展，根据汉英或英汉机器翻译以及其他多语言信息处理的需要，专有名词的机器翻译研究已开展起来。关于少数民族语言的人名和地名的汉字转写，有关机构制定了规范并在此基础上开发专门软件来实现了自动转写和翻译，如：新疆民语委已开发成功维一汉，哈一汉人名的单向翻译软件并投入使用。但是汉语和维吾尔语等少数民族语言之间专有名词的双向自动翻译问题还没有有效的解决。

某些类别的专有名词初看似乎可以用专名词典，如收地名词典的方法进行翻译。但实际上，专有名词是一个开放、不稳定的类，随着时间的推移，旧的专有名词不断消失，新的不断出现，其数量十分庞大，采用列举的方法往往难以穷尽。专有名词的机器翻译有两大主流方向，即基于语料库（词典）与基于规则的方法。前者只能对已出现过的专名进行翻译，而对未登陆的专有名词的翻译效果多数情况下并不理想；后者不仅能对已出现过的专名进行很好的翻译，而且对新出现专有名词的翻译效果也令人满意。

本文针对真实文本中出现最为频繁的人名、地名、机构名等三种专有名词，提出了一种基于转换规则的专有名词自动翻译方法，该方法根据汉语和维吾尔语的特点，将翻译过程分为三个阶段，从而实现汉语专有名词向维吾尔语的自动翻译。其不同于传统的机器翻译方法，不需要建立丰富、完整的双语词库。我们从《人民日报》语料库随机选取 3 万字的语料进行测试，其试验结果表明：采用该方法的准确率达到 90.5%，从而证明了基于转换规则的专有名词自动翻译方法的有效性。初步试验结果表明该方法是可行的，而且基于该方法的专有名词自动翻译子系统可以应用到跨语言信息处理的各个领域之中。

2 专有名词及其翻译方法

从实用的角度考虑，名词通常被分为专有名词和普通名词。所谓专有名词，是指人名、地名、机构团体名和其他具有特殊含义的名词或名词词组。本文将重点放在最为常见的人名(PER)、地名(LOC)、机构名(ORG)等三种专有名词上。此外，专有名词有一条重要的“名从主人”原则，即所译人名、地名，若原先不是汉语的，应按照其原来的读音处理，而非汉语中的发音。由于对这些外来专有名词的处理方法不同于其他专有名词，在识别过程中就要加以区分，根据它们的特点采取适当的翻译策略。因此，本文提出的翻译算法的对象不包括外来专有名词。下文就结合汉语专有名词的特点，介绍汉文专有名词向维吾尔文的翻译规则和翻译方法。

2.1 专有名词的特点

一个完整的汉语人名是由姓和名两部分组成的，其中大部分姓是一个字，少部分是两个字；名两个字的多，一个字的少。相对来说，人名的自动翻译比其他专有名词容易，通常是音译。中国地名的内模式结构要比中国人名复杂，大部分地名由专名和通名组成的。如：“上海市、张北县、安徽省”中的“市、县、省”是通名，前部分是专名。地名作为一种专有名词，也主要以音译为主，当然也存在一些特殊情况。当地名中的专名是单音节时，应当将通名视为专名的组成部分，先音译并与专名的音译连写，后重复意译，分写。

组织机构名是指机关、团体或其他企事业单位的名词，包括：学校、公司、厂矿、银行、医院、研究所以及政府部门机关的厅、局、部、委办等等。组织机构名的翻译与人名、地名相比难度最大，其构成也最不稳定，随着社会的发展，新的组织机构不断涌现。但机构组织名称也具有许多特性，其末尾的特征词一般都会出现，其构成有时也有一定规律可循。比如：“北京华建集团”是由一个地名，一个专名和一个通名组成的机构名。从此可见，地名和机构名由专名和通名两部分组成。音译通常是针对专名的，而对表示特征、性质意义的通名要进行意译。因此，专有名词翻译应包括专名的音译和通名的意译。通名的意译可用查词典的方法，但是由于专名通常是未登陆词，查词典的方法往往不够有效。所以，我们在本研究中采用了基于转换规则的专有名词音译法，本文也重点讨论专有名词的音译问题。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2 音译转换规则

音译是把一种语言中的词、字母的读音译成另一种语言的读音，或者用目标语的文字符号来表现原语的发音，主要用于专有名词的翻译上。在汉维专有名词翻译系统中，汉语是原语，维吾尔语是目标语。该系统对专有名词进行翻译时，模仿人的翻译，通过一系列转换过程完成。翻译人将一个汉语专有名词翻译成维吾尔语的时候，首先发出该词在汉语普通话中的标准音，其后为使该发音适应维吾尔语的语音习惯，而进行一些必要的转换。在此过程中，一方面尽力保留原语的发音特点，另一方面将该发音适应目标语的语音系统。汉语和维吾尔语的语音不是一一对应的。比如：维吾尔语没有汉语语音/zh/、/sh/，汉语词语中的这些语音，在维吾尔语通常用/j/、/ 来表达，汉语的/a/在维吾尔语对应/a/、/æ/等两个音。维吾尔语的/Ø/、/q/在汉语没有对应的语音。因此，音译时需要对上述现象进行必要的转换处理，而且这些转换有一定的规律。

汉维专有名词翻译系统基于转换规则进行专有名词的翻译。所谓转换规则，是指两种语言字符串之间的对应关系，一般用对应表来描述。汉语和维吾尔语语音（或音节）之间的转换有规律，而且维吾尔语和汉语的语音，在大多数情况下是比较对应的，转换规则的获取也相当容易和简单，通过观察分析可以归纳出一系列对应关系。我们在对汉维语专有名词进行语音学比较的基础上制定了以汉语音节声母和韵母为单位的转换规则。具体如表 1 和表 2。

表 1　汉语声母和维吾尔语语音的对应表

汉语声母	b	p	m	n	f	d	t	n	l	w	g	k	h	j zh	q ch	x sh	r	z	s c	y
维语语音	*b*	*p*	*m*	*n*	*f*	*d*	*t*	*n*	*l*	*w*	*g*	*k*	*x*	*j*	*ch*	*sh*	*r*	*z*	*s*	*y*

表 2　汉语韵母和维吾尔语语音的对应表

汉语韵母	a	o	e	i	u	ü	er	ai	ao	ou	an	en	ang
维语语音	*a*	*o*	*é*	*i*	*u*	*ü*	*ér*	*ey*	*aw*	*u*	*en*	*én*	*ang*
汉语韵母	ian	in	iang	ing	iong	ua	uai	uan (üan)	un (ün)	uang	ui	ue (üe)	ei
维语语音	*yen*	*in*	*yang*	*ing*	*yung*	*ua*	*uey*	*üen*	*ün*	*uang*	*üy*	*ö*	*éy*
汉语韵母	ong	ia	iao	iu	uo	un (ün)	uan (üan)	eng	ie				
维语语音	*ung*	*ya*	*yaw*	*yu*	*o*	*ün*	*üen*	*éng*	*yé*				

例如：(1) *au → **aw* (2) *ian → **yen* (3)*ue→ * *ö*

代码形式为：If yunmustr= au　then yunmustr= aw　....

翻译过程

专有名词翻译系统输入的是已识别、标注好的中文专有名词，输出的是翻译成维文的专有名词。翻译过程如图 1 所示。

如流程图所示，人名翻译是音译，地名和机构名的翻译不是简单的音译问题。地名和机构名进入分析器，通过分析、匹配对其通名（特征词）部分以查词典的方式进行意译，剩下的部分转到音译系统，音译过程完成了以后，将音译结果与意译部分合并形成一个完整的专有名词。很多汉语分词标注系统识别专有名词时就切开其专名和通名部分，以便翻译系统能够辨别专名和通名部分。通名或特征词的翻译是任何翻译系统都应具备的一项功能，其翻译方法也简单。因此，在本文中只讨论专名部分的音译问题。

音译系统将翻译过程分为三个步骤进行。

步骤 1：汉字转换为拼音

汉文不是拼音文字，文字系统所使用的汉字不是发音。所以，汉语专名的音译需要用拼音来转写汉字。在第一步，音译系统的拼音转写模块对要音译的汉字进行转写（不转写声调）。例如：周恩来 → /zhou en lai/

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

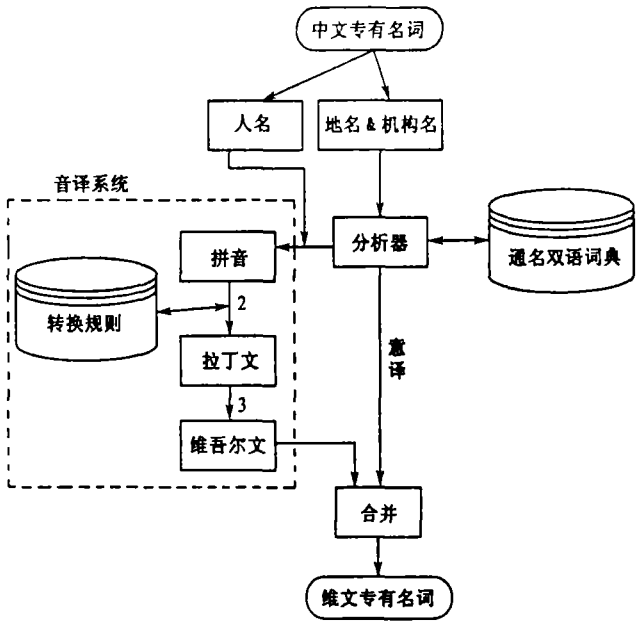

图 1　汉维专有名词翻译系统流程图的流程图

步骤 2：拼音（PINYIN）转换为维语拉丁文（LSU）

这是音译过程的关键一步，在此步骤将拼音转换为代表维吾尔语发音的中介文字—维语拉丁文（Latin Script of Uyghur，简称 LSU）。首先代表一个汉字的拼音（一个音节）被切分为声母和韵母，然后根据表 1,2 所示的转换规则进行转换。

例如：/zhou en lai/ →*ju énlei*. 转换过程: /zh/→*j*, /ou/→*u*, /en/→*én*, /l/→*l*, /ai/→*lei*.

步骤 3：LSU 转换为维吾尔文（ASU）

现在维吾尔语使用文字由阿拉伯文转变而来，即所谓的 ASU (Arabic Script of Uyghur)，输出结果应该为用 ASU 拼写的专有名词。因此，在最后一步骤，音译系统将 LSU 转换为 ASU。LSU 与 ASU 之间的对应关系如表 3 所示。转换过程：*ju énlei*→。

表 3　拉丁文和维吾尔文字母的对应表

ASU																
LSU	a	e	b	p	T	j	ch	x	d	r	z	zh	s	sh	f	q
ASU																
LSU	k	g	ng	l	M	n	h	gh	o	u	ö	ü	w	é	i	y

3　试验结果与分析

为了验证这种基于转换规则的专有名词翻译方法，我们对小规模语料进行了试验。

3.1　测试语料

本次试验用的测试语料来源于北京大学计算语言学研究所已分词和词性标注过的《人民日报》语料库（网址：http://icl.pku.edu.cn/icl_res/）。我们从语料库中随机抽选 3 万词的语料并对其中的 1765 条专有名词进行了自动翻译。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3.2 试验结果

自动翻译以后，对其结果和人工翻译进行了比较。研究结果如表 4 所示。根据试验结果，该系统的人名、地名和机构名的翻译准确率分别为 92.8%、89.7%和 84.9%，我们得到了 90%以上的准确率。

表 4 试验结果表

专有名词	PER	LOC	ORG	总数
数量	905	561	299	1765
准确数量	840	503	254	1597
准确率 (%)	92.8	89.7	84.9	90.5

3.3 错误分析

试验中出现了一些翻译错误，机构名的错误率高于人名和地名的错误率。这些翻译错误可归纳为以下几种：

第一、最常见的错误发生在对外来专有名词的翻译上。由于系统无法辨别出汉语专有名词和外来专有名词，这种错误很难避免。例如：莎士比亚(Shakespeare，拼音为 shashibiya)是外国人名，错误翻译为 shashibiya，应该是 shikispér。雅加达 (Jakarta, 拼音为 yajiada) 系统翻译为 *yajyada*，在维吾尔语里应该为 jakarta，这些错误都是由外来专有名词的发音在汉语中的不同导致的，为了解决这些问题需要采取特殊标注、还原等其他措施。

第二、翻译一些地名时又出现了错误。汉语中“河、山、村”等有些汉字可以是另一个地名的一部分，这时需要音译加意译；但是一般的处理方法却是单纯的意译。比如：庄河不是河名，而是市名，错误翻译为“*juangxé deryasi*”（*deryasi* 意思为河），“河”应该音译为“*xé*”，在“黄河”的“河”应该翻译为 *deryasi*。

第三、最少见的一个错误是不规则专有名词的翻译。例如：新疆(xinjiang)习惯上应该翻译为“shinjang”，不是“shinjyang”；香港(xianggang)应该翻译为“shanggang”，不是“shyanggang”。这种错误出现的原因在于这些地名现在通用的翻译不符合上文指出的转换规则。它们应按照约定俗称的原则，而不是转换规则来翻译。实际上，前两种错误是由专有名词的识别，以及音译、意译部分的区分造成的，只有第三种才是翻译系统出现的错误。

4 结论和未来工作

本文提出了基于转换规则的专有名词翻译方法，该方法高效而不依赖于双语词典。它作为其他机器翻译系统的辅助工具，在提高翻译质量方面将起到重要作用。跨语言信息检索系统也使用该方法能够准确地辨别出两种语言的相关信息。

本次试验的结果与专有名词翻译的最终目标相比，尚有距离，需要我们继续努力。将来我们进一步研究外来专有名词和机构名的自动翻译问题，继续进行其他补充性研究。

参考文献

[1] Chen, Hsin-His. Proper Name Translation in Cross-Language Information Retrieval. *Proceedings of 17th COLING and 36th ACL*, pp. 232～236, 1998.

[2] Jarmo Toivonen .etc: Translating cross-lingual spelling variants using transformation rules, *Information Processing and Management: an International Journal;* Vol 41; pp: 859 – 872, 2005.

[3] Knight, K. and Graehl. J：Machine Transliteration. *Computational Linguistics*: 24(4), 1998.

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[4] Paola Virga, Sanjeev Khudanpur: Transliteration of proper names in cross-lingual information retrieval; *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition* - Volume 15

[5] Yan Qu. Automatic Transliteration for Japanese-to-English Text Retrieval. *Proceedings of 21st COLING and 44th ACL*, pp: 1129～1136, 2006.

[6] 艾山·吾买尔，吐尔根·依布拉音. 英文－维吾尔文人名机器翻译算法的研究与实现. 中文信息处理学术研讨会论文集（2006），北京，2006 年 11 月

[7] 张玥杰，徐智婷，钱晶，张涛. 自然语言处理中专名识别方法的研究. 中文信息处理学术研讨会论文集（2006），北京，2006 年 11 月

[8] 方小兵. 专有名词音译探讨. 皖西学院学报，2002（2）

[9] 郭曙纶. 汉语人名标注及其方法. 零陵学院学报，2003（3）

[10] 秦贻. 专有名词的翻译原则和技巧. 湖北工学院学报，2004（6）

[11] 张国喜. 英藏命名实体在机器翻译系统的实现. 青海师范大学学报(自然科学版)，2004（3）

[12] 王兴义. 基于模式匹配的中文专有名词识别. 硕士学位论文，山西大学，2005 年

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net