广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(4):439~442
Vol. 26, No. 4 November 2010

基于概率模型的汉语和越南语的人名音译方法*

The Approach of Chinese-Vietnamese Name Transliteration Based on Probabilistic Model

申文明[1],刘连芳[1,2],黄家裕[2],温家凯[2]
SHEN Wen-ming[1], LIU Lian-fang[1,2], Huang Jia-yu[2], WEN Jia-kai[2]

(1.广西大学计算机与电子信息学院,广西南宁 530004;2.南宁平方软件新技术有限公司,广西南宁 530007)
(1. College of Computer and Electronic Information, Guangxi University, Nanning, Guangxi, 530004, China; 2. Pingsoft New Technology Co. Ltd. of Nanning, Nanning, Guangxi, 530007, China)

摘要:利用概率模型训练、学习得到基于字形的汉越音译知识,实现汉语和越南语的人名音译。音译方法简单有效,在汉译越上效果尤为显著,准确率达到97.41%。

关键词:汉越人名翻译 音译 概率模型 音译知识
中图法分类号:TP391.2 **文献标识码**:A **文章编号**:1002-7378(2010)04-0439-04

Abstract: A probabilistic model is used to get the transliteration information of Chinese-Vietnamese based on grapheme and the translation between Chinese and Vietnamese name is based on this information. The approach is simple and effective, especially on Chinese-to-Vietnamese transliteration, and the precision rate reaches 97.41%.

Key words: Chinese-Vietnamese name translation, transliteration, probabilistic model, transliteration information

随着世界经济文化的融合,大量的外来词纷纷涌现到本土文化中,这些外来词多为命名实体:人名、地名和机构名等。外来词只有部分可以直接使用,大部分都需要进行翻译。翻译外来词的方法主要有两种:意译和音译。意译是根据词语的意思把词语从一种语言翻译到另一种语言,而音译是根据其发音来翻译的。例如"Penicillin",翻译为"青霉素"是意译方法,翻译为"盘尼西林"是音译方法。由于命名实体是一个开放集,很难在词典中来获得相应翻译,因此主要使用音译方法来翻译。命名实体的音译在机器翻译、跨语言信息检索、多语言命名实体对抽取、平行语料库对齐等自然语言处理领域占据重要地位。

根据音译方法的不同,音译被分为基于语音(phoneme-based)的音译和基于字形(grapheme-based)的音译[1]。基于语音的音译需要3个转换步骤:首先将源语言词语转换成其发音符号(音素),然后再转换为目标语言的近似音素,最后再转换成目标词语。基于语音的音译需要通过机器学习来获取转换规则,转换过程中信息容易丢失,并且多个步骤错误的累加会影响整个音译的准确率。基于字形的音译是将源语言词语按照字形单元的匹配直接转换为目标语言词语,例如从人名"smith"和"史密斯"训练得到"s-mi-th"直接对应着"史-密-斯"。基于字形的音译方法将基于语音的3个步骤简化成1个步骤,减少了中间过程对整体音译性能的影响,从而提高了音译的准确率[1]。基于字形的音译方法是近些年的研究热点,主要是在联合信源通道、噪声通道等原有音译模型的基础上引入新的特征来改进音译的性能,比如音译单元相对距离、关键字信息、上下文

收稿日期:2010-08-16
修回日期:2010-10-13
作者简介:申文明(1984-),男,硕士研究生,主要从事信息检索和自然语言处理研究。
*科技部2010年度科技型中小企业创新基金(10C2614502818)资助。

信息相似度计算、前/后向音节映射模型等[2]。

命名实体中的人名在语料中所占的比例很高，庞薇等通过对NIST测试语料中随机挑选3940句的统计发现人名共计2312个，占总词数的2.7%[3]。人名的音译是命名实体翻译中需要解决的重要问题。当前人名音译的研究主要集中在英语和汉语之间，邹波等[4]把英汉人名音译问题转化为序列标注问题，使用最大熵模型和条件随机场模型训练得到从英文字母直接转换成汉字的音译知识；庞薇等[3]建立以基于字符和发音的转换模型为核心，通过加权有限状态转换器将多模型进行融合的人名翻译框架，实现英汉人名音译；周美玲[2]用统计机器翻译方法实现一个英汉人名音译系统；艾山·吾买尔等提出了一种基于规则库的多层过滤的人名翻译算法，通过建立3个规则库实现英语到维吾尔语的人名音译[5]。

相对英语和汉语之间的人名音译取得众多成果的现状，目前汉语和越南语之间的人名音译研究还是个空白。和汉英人名音译相比，汉越人名音译在语言、人名文化、音译知识等方面存在着较大的差别：(1)英语和汉语属于不同的语系，英语的一个单词(word)可以对应着一个或多个连续的汉字；汉语和越南语的词素分别是汉字和音节(syllable)，一个汉字对应一个音节。(2)英语和汉语之间的相互影响相对较小，人名文化上的共同点很少；越南和中国的人名在构成以及起名的禁忌、习惯和规则上都极具相似[6]。(3)英语单词和汉字在读音上不存在固定的对应关系，需要用音译模型对汉英人名进行训练来获取汉英音译知识；而越语音节和汉字之间存在着固定的对应关系，这些对应关系可以作为汉越人名音译知识。

汉越人名音译与汉英人名音译的本质差别使得汉越人名音译必须根据汉语、越南语的特点和相互之间的关联来获取汉越音译知识，从而实现汉越人名音译。本文利用概率模型对现有的汉越音译表进行训练、学习得到越南语音节直接转换为汉字的汉越音译知识，实现汉语和越南语的人名互译。

1 获取汉越音译知识

越南语是越南的国语，也是中国京族使用的语言，而且在老挝、缅甸、柬埔寨等东盟国家和地区有一定影响力。越南处于汉文化圈中，越南语本质上主要受汉语的影响。经过上千年的文化交流和融合使得很多越南语的音节和汉字之间产生了一定的对应关系，例如越南语的“h ạo”对应着汉字“浩、昊、皓、灏、颢”，这些对应关系将是汉越音译知识的主要来源。

基于概率模型的音译知识训练就是通过对训练语料进行统计得到一个越南语音节到任意汉字转换的概率，一个越南语音节 a 翻译转换为任意汉字 b_i 的概率为：$P(a \mid b_i)=P(a,b)/P(b_i)$。在汉越人名语料中计算所有越南音节到每个汉字的翻译概率，并根据概率大小对汉字进行排序，最终得到一个越南语音节和按转换概率排序的汉字集合的对应关系，即汉越人名音译知识。

使用来源于《现代越汉词典》[7]的“汉越字表”和《Chữ’ Hán và Tiếng Hán-Việt》[8](汉字与汉越音)的汉越音译表，把《实用汉越分类词典》[9]的“越南历史名人翻译对应表”作为训练语料，最后共得到1768个越南语音节和6250个汉字的对应关系。以后通过对双语词典或高质量平行语料库中提取的汉越人名对进行学习，得到越南语音节和汉字新的对应关系，把语言专家审核认可的加入到音译知识中。

2 汉越人名音译原理

在汉越人名音译系统的框架(见图1)中，翻译的核心模块包含2个主要步骤：(1)姓的翻译。利用常用姓氏库来确定姓名的姓氏部分并翻译，如果不在常用姓氏库中则直接跳向第2个步骤。(2)名的翻译。姓名里确定姓氏后余下的就是名字，利用汉越音译知识对名字的每个词素进行翻译。例如，在汉译越中翻译“司马相如”，首先通过查询中越常用姓氏库来确定姓氏是复姓“司马”并翻译为“Tu’ Mã”，然后再利用汉越音译知识把名字部分的“相如”翻译得到“Tu’o’ng Nhu’”，合并姓和名后得到翻译结果为“Tu’MãTu’o’ng Nu’”。

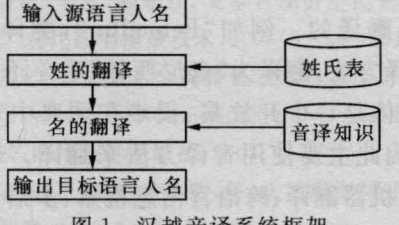

图1 汉越音译系统框架

由于汉字对应唯一的越南语音节，在汉译越时，如果音译知识中有该汉字则可以直接翻译为对应的越南语。反之，每个越南音节对应一个或多个汉字。因此，在越译汉时译者得到的是音译知识中越南音节对应的汉字集合。如果集合中汉字多于一个，译者需要从多个候选汉字中选择，这就需要译者熟悉

中国人的命名习惯。很明显在翻译"La Li ệt"的时候"罗烈"比"罗列"要更符合中国人的习惯。

由于当前的汉越音译知识的不足,出现很多汉字没有对应的越南音节和越南语音节没有对应的汉字的情况,例如以"im"为韵母的越南语音节几乎都不存在。因此,当翻译的汉字或越南语音节不存在的时候需要用相似替代法进行翻译,相似替代的策略主要有语音相似和语义相似[6]。对当前的汉越音译知识进行统计发现一个越南音节所对应的汉字几乎都是同音字,因此用语音相似的替代策略,在汉译越的时候选择同音而且同声调的汉字来替代,例如:用"荣"来代替"镕",在越译汉的时候选择读音相似的音节来替代,例如用"dăn"替代"dân"。

3 实验与分析

实验中使用的汉越姓氏表主要来自于互联网和文献6附录提供的"京族人姓氏与汉语对照表",共有汉越对应的姓氏551个。使用的人名来自于维基百科(Wikipedia),维基百科作为互联网上最大和最广泛使用的多语种百科全书,很多条目都存在着对应的多种语言版本。维基百科中的越南语和汉语对应的条目既可以作为我们实验数据的来源,也可以作为评价音译效果优劣的参考标准。从维基百科离线数据(http://dumps.wikimedia.org/backup-index.html)中得到2010年6月12日的汉语数据库(zhwiki)和越南语数据库(viwiki),把抽取到的1004对人名分为越南语人名集VData和汉语人名集CData。

音译分为正向翻译和反向翻译,正向翻译是指将源语言词音译为目标语言词的过程,反向翻译是指将目标语言的音译词翻译为目标语言词的过程[3]。以汉译越为例:把"邓小平"翻译为"D ặng Ti ểu Bình"是正向音译,而把来自于越南的"阮晋勇"翻译为"Nguy ễn T ân Dũng"是反向音译。为了实验正向翻译和反向翻译的效果,把越南语人名集VData按照来源分为来自越南的VData1和来自中国的VData2,分别有236个和768个,同理也把汉语人名集CData分为来自于中国的CData1和来自越南的CData2,分别有236个和768个。

实验分为越译汉和汉译越,分别翻译越南语人名和汉语人名。当前英汉人名音译的评价存在双语词典和评测集,例如《外国人名的汉语音译》、ACE 2007的ET测试集等。但是汉越人名翻译还没有比较权威的评测集,我们的实验采用维基百科提供的翻译作为音译结果评价标准,并用准确率作为衡量音译的效果的指标:准确率=翻译正确的个数÷总个数×100%。

汉译越:对CData中的汉语人名分别进行正向和反向音译,汉译越的整体翻译效果非常好,正向音译的首项准确率达到97.39%,前五项准确率达到98.30%,反向音译的首项准确率达到95.57%,前五项准确率达到96.46%,正向音译比反向音译的准确率略高。总体上汉译越的首项准确率达到97.01%,前五项准确率为97.41%。分析发现汉译越的错误都是在相似替代翻译时出错,例如对"荀彧"中的"彧(Ng ọc)",用"豫(Vũ)"替代翻译。因此在相似替代翻译上需要寻找更好的相似替代策略。

越译汉:对VData中的越南语人名分别进行正向和反向音译,越译汉的效果良好,正向音译的首项准确率达到85.84%,前五项准确率达到91.155,反向音译的首项准确率达到83.08%,前五项准确率达到88.28%,正向音译比反向音译的准确率略高。整体上越译汉的准确率达到83.83%;前五项的准确率达到89.83%。分析发现越译汉出现的错误往往是翻译时的汉字选择不恰当,例如把"Tr ịnh Doanh(郑楹)"翻译为"郑莹",而"Doanh"对应的汉字有"莹、营、赢、瀛、盈、楹、茔";另外由于无法判断人名的性别,也加大了选字的难度。因此在下一步工作中,一方面要设法扩大汉越音译知识,另一方面需要在更大规模的人名语料库上进行训练,以挑选到更好的候选字。

汉译越准确率比越译汉的准确率高出近10个百分点,主要原因是:(1)一个汉字对应唯一的越南语音节,而一个越南语音节却对应多个汉字,从多个候选汉字中选择一个就增加了翻译的错误率。(2)中国人名文化博大精深,在越译汉中,选择汉字的时候仅考虑了翻译的概率,没有考虑到中国人起名的禁忌和风俗习惯。

4 结束语

汉越人名的翻译是汉语和越南语互译中一个亟待解决的重要问题,汉越人名音译的研究不仅有利于推动越汉互译的进步发展,也有利于扩大中越两国经济、文化的进一步交流。

利用概率模型对现有的汉越音译表进行训练得到基于字形的汉越音译知识,实现汉越和越汉人名音译。音译方法简单有效,在汉译越上效果尤为显

著，但是越译汉的效果还需要进一步提高。下一步我们将通过对高质量的平行语料库中抽取人名对训练、学习来扩大汉越音译知识，从而提高音译的整体效果。另外，本文的音译方法不仅可以用于汉越人名的翻译，还可以应用于汉越机器翻译和汉越双语命名实体对的抽取。

参考文献：

[1] 王蕾. 基于字形的英汉机器音译改进研究[D]. 哈尔滨工业大学，2007.

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

[3] 庞薇，徐波. 基于多模式融合的人名翻译系统[J]. 中文信息学报，2009，23(1)：44-49.

[4] 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，2008：232-238.

[5] 艾山·吾买尔，吐尔根·伊布拉音. 英文维文人名机器翻译算法的研究与实现[J]. 新疆大学学报：自然科学版，2007，24(1)：97-101.

[6] 吴妙玲. 越语人名与汉语人名的对比与翻译问题[D]. 桂林：广西师范大学，2008.

[7] 雷航. 现代越汉词典[M]. 北京：外语教学与研究出版社，1998.

[8] Phạm Văn Hải，Lê Văn Dặng. Chữ Hán và Tiếng Hán-Việt[EB/OL]. http://www.viethoc.org/eholdings/PhamVanHai/ChuHanvaTiengHanViet.pdf. 2005.

[9] 梁远. 实用越汉分类词典[M]. 北京：民族出版社，2007.

（责任编辑：邓大玉）

（上接第435页）

参考文献：

[1] Ricardo Baeza-Yates，Berthier Ribeiro-Neto，王知津，等. 现代信息检索[M]. 北京：机械工业出版社，2005.

[2] 何伟，薛素静，孔梦荣，等. 基于Lucene的全文搜索引擎的设计与实现[J]. 情报杂志，2006(9)：88-89.

[3] 管建和，甘剑峰. 基于I.，ueene全文检索引擎的应用研究与实现[J]. 计算机工程与设计，2007(2)：490-491.

[4] 赵汀，孟祥武. 基于Lucene API的中文全文数据库的设计与实现[J]. 计算机工程与应用，2003，20：179-183.

[5] 葛帅. 开方源代码的全文检索引擎[EB/OL]. http://www.lucene.com.cn/about.htm 2006.09.

（责任编辑：邓大玉）