文章编号：1003-0077(2011)02-0049-06

题录信息的机器翻译方法

李贤华, 于 淼, 苏劲松, 吕雅娟

(中国科学院 计算技术研究所 智能信息处理重点实验室, 北京 100190)

摘 要：该文针对题录信息中的人名、地址、机构名和公司名的不同特征，分别设计了不同的翻译方法，并依靠词典和翻译规则，实现了大部分内容的翻译。对于人名翻译，该文设计了拼音转换、假名转换和同音转换的翻译方法；对于地址、机构名和公司名的翻译，该文提出了先切分、再翻译、最后调序的翻译流程。实验表明，利用该文的方法翻译人名、地址、机构名及公司名，能够取得不错的翻译效果。

关键词：题录信息；机器翻译；人名翻译；地址翻译；机构名翻译

中图分类号：TP391　　文献标识码：A

Approaches to Bibliographic Information Translation

LI Xianhua, YU Miao, SU Jinsong, LV Yajuan

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper proposes different machine translation approaches to translating bibliographic information, such as person names, addresses, organization names and company names according to their different features. With dictionary and translation rules, most of them can be translated properly. For name translation, we design Pinyin conversion and Kana conversion methods. For address translation, organization name translation and company name translation, we propose a procedure which includes splitting, translating and reordering. Experiments show that these approaches achieve good results.

Key words: bibliographic information; machine translation; person name translation; address translation; organization name translation

1 引言

机器翻译是使用计算机进行翻译工作的技术。从 1949 年 Weaver 提出机器翻译的概念至今，短短半个世纪中，机器翻译技术在各个领域发挥着越来越大的作用，人们在机器翻译领域取得了很多阶段性的成果[1-3]。题录信息的翻译，是机器翻译的任务之一。随着信息社会的不断发展，题录信息的翻译会有广阔的应用前景，比如名片翻译、专利人信息翻译、刊物作者信息翻译、英文信函地址翻译等。

题录信息是书籍、科技文献、专利文献等的一个重要组成部分，它通常包含标题、人名、地址、组织机构名、公司名以及邮箱地址等。本文主要处理题录信息中人名、地址、组织机构名和公司名的翻译。由于这些信息具有上下文无关性、类型容易确定等特点，其翻译相对于其他内容的翻译来说目标更明确、精度更高。在现代化的信息社会中，题录信息的翻译在政治、经济、外交、贸易、旅游、新闻出版、文化交流以及日常生活中都有着重要的应用。当前研究人名翻译的工作很多，文献[4]等提出了人名翻译的许多注意事项，但提出人名翻译的通用方法的文献较少；研究地名翻译的工作也层出不穷[5-6]，但是针对地址翻译的工作较少；还有一些工作致力于从双语

收稿日期：2010-08-01　定稿日期：2010-11-23

基金资助：国家自然科学基金资助项目(60873167, 60736014)

作者简介：李贤华(1985—)，女，硕士生，主要研究方向为机器翻译，自然语言处理；于淼(1986—)，女，硕士生，主要研究方向为机器翻译，自然语言处理；苏劲松(1982—)，男，博士生，主要研究方向为机器翻译，自然语言处理。

语料库中获得翻译[7],但这些方法受到语料规模和时期的限制。目前大量题录信息的翻译工作是人工完成的。人工翻译虽然有着较高的翻译质量,但是其耗时长,占用资源多,不适合大规模的翻译。这些都是本文将解决的问题。

本文主要设计了题录信息中人名、地址、机构名和公司名的机器翻译方法。采用的方法主要是词典查找和规则翻译等。对于中国人名,本文使用拼音转换的方法进行翻译,即通过查看汉字拼音转换表对汉字进行翻译;对于日本人名,本文设计了假名转换的方法,即首先将中文的日本人名转换为假名,再将假名转换为相应的罗马字母的方法;对于欧美国家人名,本文设计了同音转换的方法,即读音相同的欧美国家人名,其对应的译文也相同;对于地址、机构名和公司名,本文提出了先切分、再翻译、最后调序的翻译流程。用汉语拼音拼写中国人名和地址,更加有利于不同国家的人们了解中国文化,也更加方便外界与国人的沟通交流,是中国和全世界的标准。

本文的组织如下:第 2 节详细介绍了人名翻译的主要方法和策略,针对中国人名、日本人名和欧美人名的特点,分别设计了相应的翻译方法;第 3 节介绍了地址翻译的方法,将地址翻译的过程分为地址切分、局部翻译、译文调序三大部分,并给出了每一步的具体过程;第 4 节主要介绍了如何翻译机构名和公司名,其翻译方法与地址翻译的方法类似。在第 5 节中,介绍了实验情况,经过人工随机抽样测试,本文设计的翻译方法能够很好的翻译人名、地址、机构名和公司名。最后一节,我们对本文的工作进行了总结,并指出未来研究工作的方向。

2 人名翻译

人名是意义相对较少的专有名词,是所指称对象的一个对应符号。一般地,人名的翻译方法主要有书写形态借用、语音借用、语义翻译三种。当两种语言处于相同或者相似的文字系统中时,一般采用书写形态借用的翻译方法;当两种语言处于不同的文字系统中时,语音借用起了很大的作用[8];当人名有着特殊的意义时,一般采用语义翻译的方法。由于汉语和英语处于不同的文字系统,本文主要采用语音借用的翻译方法。

本文主要处理三类人名:中国人名、日本人名以及欧美国家人名。人名首先经过词典进行切分查找翻译;不能通过词典得到翻译的人名,将首先通过人名分类器得到其对应的类别,然后根据类别使用不同的翻译方法进行翻译。

2.1 词典的使用

词典是在进行题录信息翻译时的辅助资源。由于题录信息的翻译相对于长句的翻译来说,内容简短、存储空间小、查询效率高,因此,题录信息的机器翻译借助于词典,显然是简单可行的方法。同时,词典提供给用户灵活添加词典词条的接口,从而极大的提高翻译质量。另外,对于一些有歧义的翻译项,将其添加进词典后,由于词典的优先级较高,译文优先选择词典内的翻译项,可以尽量避免歧义造成的干扰。

本文针对人名翻译、地址翻译、机构名和公司名翻译,分别开发了三本词典:人名词典、地名词典、机构公司词典,以此来翻译不同的内容。三部词典均存储在数据库中,其中每个词条包含如下特征:序号、中文端、英文端、所在词典、用户 ID、添加时间、是否使用、是否审批等。

除了用户词典,本文还用到了 LDC 命名实体词典①。LDC 在语料资源的开发加工方面做了大量工作,是国际上自然语言处理方向最大的资源共享发布平台。本文使用 LDC 开发的命名实体词典,来辅助题录信息的翻译。

在进入题录信息翻译模块时,首先查找词典,如果词典中已包含需要翻译的词条,那么,直接将其对应的翻译取出,作为翻译结果;否则,进入规则翻译流程,用规则方法实现词条的翻译。

使用拼音转换等方法,已经可以翻译题录信息的大部分内容,但仍有少数的翻译结果差强人意。本系统提供给用户自行添加词典词条的接口,用户可以动态地加入自定义的词典词条,从而明显提高了翻译质量。

由于在人名翻译、地址翻译、机构名和公司名翻译的模块中,对词典的使用与维护类似,因此这里一并作出论述,下面不再累述。

2.2 人名判断器

人名判断器的主要作用是判断人名所属的类别,其主要利用人名的姓氏特征、字符特征和长度特征进行判断。中国人名、日本人名和欧美国家人名

① http://projects.ldc.upenn.edu/Chinese/.

的姓氏有显著的不同, 按照姓氏特征可以基本区分这三种人名。本文收集了中国姓氏 494 个, 日本姓氏 9 973 个(其中有对应翻译的姓氏为 3 617 个), 以此识别绝大部分的中国人名和日本人名。字符特征主要用来识别欧美国家人名。欧美国家的正式人名, 姓氏与名字之间多用“ · ”间隔, 大多数名字带有字母, 这是中国人名和日本人名不具备的特征。通过符号特征可以将欧美国家人名识别出来。长度特征主要用来判断通过姓氏特征和符号特征无法识别的人名。

2.3　人名翻译流程及方法

针对上述三类人名, 本文分别使用三种不同的方法进行翻译, 其主要流程如图 1 所示。

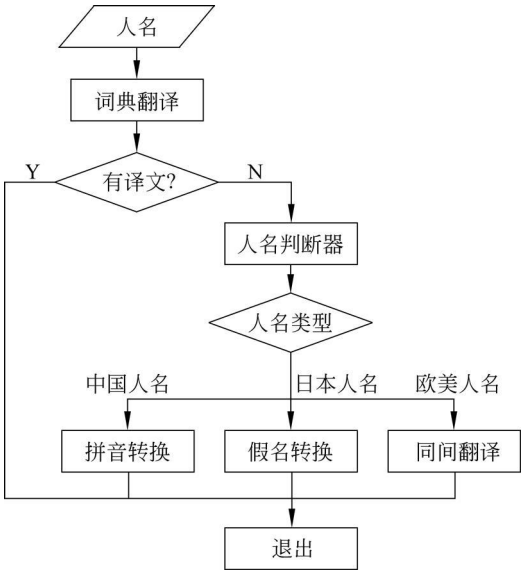

图 1　人名翻译主要流程

对于中国人名, 主要采用拼音借用的方法。汉语的文字系统和英语的文字系统虽然不相容, 但是罗马化的汉语拼音和英语形成了一种特殊的书同文关系, 中国人名通过汉语拼音直接借用为英语人名, 在理论和实际操作中都是可行的。比如中国人名“李志强”, 直接用拼音“Zhiqiang Li”作为其翻译即可。对于日本人名, 主要采用假名转换的方法。假名转换, 指的是通过假名做中间桥梁, 首先将汉字的日本人名转换为对应的假名形式, 再根据假名的发音, 将其转换为对应的罗马字母。比如日本人名“藤田良雄”, 首先将其用假名“ふじた よしお”表示, 再根据假名的读音, 转换为“Fujita Yoshio”即可。而对于欧美国家人名, 如“R. A. 戴维森”, 则根据字符“·”进行切分后, 再逐个进行翻译。对于切分后的每个单元, 将其转换为拼音, 如果与词典中词条的拼音相同, 则将其对应翻译选为译文, 这样, 可以增加词典中词条的匹配率。

3　地址翻译

地址的翻译, 指的是将中文的地址翻译为英文的地址。本文主要设计了中国地址、日本地址和欧美国家地址的翻译方法。本文收集了中国省市区县的名字共 2 381 个, 全世界国家名 220 个, 作为地址切分和翻译的基础。另外, 中文地址的书写特点是先大后小, 即将区域范围广的内容放在前面, 区域范围窄的内容放在后面; 而英文地址的书写恰好相反。一般地, 地址翻译的主要原则是先小后大, 本文采用译文调序的方法, 实现译文的先小后大排列。

地址翻译的主要流程是: 先切分, 再翻译, 最后调序。

3.1　地址切分

地址切分, 指的是将长串的地址切分为有独立意义的较小单元, 方便下一步的翻译[9]。地址切分的主要依据是词典和切分关键词。地址切分使用“前向最大匹配法”, 并优先考虑词典中的词条。由于中国地址和日本地址、欧美地址的切分关键词不尽相同, 本文设计了两组切分关键词表, 一组用来切分中国地址, 一组用来切分外国地址。部分切分关键词见表 1。

表 1　部分切分关键词

中国地址切分关键词: 中国, 省, 市, 区, 县, 镇, 市镇, 村, 院, 乡, 沟, 队, 局, 街, 路, 大道, 庄, 屯, 组, 医院, 大学, 弄, 巷, 楼, 单元, 号……
外国地址切分关键词: 国, 州, 县, 区, 市, 城, 府, 目, 番, 番地, 号房, 巷, 町, 街, 路, 楼, 号, 公园, 工业园, 工业区, 公寓, 单元, 信箱, ……

比如地址: 安徽省芜湖市新芜区莲塘村芙蓉园 6 幢 4 单元 102 室。根据关键词表, 可以切分为: 安徽省　芜湖市　新芜区　莲塘村　芙蓉园　6　幢　4　单元　102　室。

3.2　局部翻译

局部翻译, 指的是将切分得到的各个单元分别翻译为英文。翻译的主要方法是查词典和拼音转换: 对于切分后得到的每一个单元, 首先通过查词典

得到翻译;对于查字典后没有翻译的单元,将符合条件的单元进行拼音转换,以得到其对应翻译。此处的符合条件,指的是该单元的最后部分在拼音转换关键词表中。部分拼音转换关键词见表2。

上述切分后的地址:安徽省 芜湖市 新芜区 莲塘村 芙蓉园 6幢 4单元 102室,经过地址翻译后的结果为:Anhui Province, Wuhu City, Xinwu District, Liantang Residential, Furongyuan, Building 6, Unit 4, Room 102.

切分关键词表以及拼音转换关键字表可以随时进行修改和维护,以提高地址翻译的准确率。

表2 部分拼音转换关键词

省	Province	楼/栋/幢/座	Buiding
市	City	巷/弄	Lane
区	District	信箱	Mailbox
县	County	院	Yard
镇	Town	单元	Unit
村	Residential	层	Floor
路	Road	室	Room
街	Street	号	No.

3.3 译文调序

译文调序,即将翻译后的译文进行一定的调序,使其符合英文的书写习惯。译文的调序需要满足一定的规律,上述翻译后的地址:Anhui Province, Wuhu City, Xinwu District, Liantang Residential, Furongyuan, Building 6, Unit 4, Room 102,进行倒置后得到:Room 102, Unit 4, Building 6, Furongyuan, Liantang Residential, Xinwu District, Wuhu City, Anhui Province。此即原地址"安徽省芜湖市新芜区莲塘村芙蓉园6幢4单元102室"的最终翻译结果。国外地址与国内地址一样,只是使用了不同的切分关键词表。

地址除了包含地址信息外,还会包含机构名、公司名等复杂地址信息。这部分内容的翻译,将在第4部分得到阐述。

4 机构名和公司名翻译

机构名和公司名的翻译过程,与地址的翻译过程类似,即遵循"先切分,再翻译,最后调序"的翻译流程。相对地址翻译而言,机构名和公司名的切分比较简单,倒置规则也较简单。由于机构名和公司名中有些部分不能直接用拼音转换的方法进行翻译,其对词典的依赖程度相对较高。

本文收集并整理了常见的机构和公司后缀名327个,并设计了他们的对应翻译。常见的机构和公司的后缀名及其翻译如表3所示。

表3 常见的机构和公司的后缀及其翻译

重点实验室	Key Lab	基金会	Foundation
大学	University	联合会	Federation
铁道部	Ministry of Railways	办公室	Office
劳动部	Ministry of Labor	株式会社	Co., Ltd.
化学公司	Chemical Company	有限公司	Co., Ltd.
研究所	Institute	研发中心	R&D Center

对于机构名和公司名,首先识别其所属地信息并进行切分,再结合用户词典、LDC词典以及后缀表进行切分,接着对各个部分进行翻译,最后进行一定的调序。

例如机构名"深圳华为通信技术有限公司",经过切分后,得到"深圳 华为 通信技术有限公司";经过翻译后,得到"Shenzhen Huawei Communication and Technology Co. Ltd";经过调序,得到最终翻译"Huawei Communication and Technology Co. Ltd, Shenzhen"。

5 数据与实验

本文对于人名翻译、地址翻译、机构名和公司名翻译这三个模块分别进行了测试。本文所用的样例来自于13 477篇专利文件的题录信息。经过去重处理,最终得到总数据量为22 705个人名、6 431个地址以及7 709个机构名和公司名。测试样例从数据中随机抽样产生,分别抽取了中国人名、日本人名、欧美国家人名、地址、机构名和公司名各200条。本文所进行的测试,均是在没有使用词典的基础上进行的。如果添加词典,翻译效果将会得到极大的提升。

5.1 人名翻译模块的测试

对于人名翻译模块,本文主要测试了人名判断器的正确率以及人名的翻译率。人名判断器的正确率对于人名翻译有着重要的意义,因此必须保证人名判断器有较高的准确率。人名翻译率主要是测试在没

有外加词典的情况下，人名得到正确翻译的情况。

本文随即抽取了中国名字、日本名字、欧美国家名字各 200 个。通过人名判断器后，统计得到人名判断器的分类正确率，再经过人名翻译模块，最终翻译结果的翻译率进行统计（见表 4）。

表 4　人名翻译模块测试结果

人名类别 / 测试类型	中国人名	日本人名	欧美国家人名
判断器正确率	100%	100%	100%
判断器召回率	100%	100%	100%
翻译率	100%	46.5%	

实验结果表明，人名翻译模块在没有添加词典的情况下，可以很好的完成中国人名的翻译；日本人名的姓氏基本可以得到翻译，名字则需要借助词典；欧美国家的人名则主要依赖词典中的词条。由于文中实验均在不加词典的前提下进行，因此欧美国家人名的翻译率并没有进行测试。

5.2　地址翻译模块的测试

对于地址翻译模块，本文主要测试了地址的切分正确率以及翻译正确率。地址的切分正确率指的是在地址切分过程中的正确率。如果将人工切分得到的地址块数量记为 N，机器切分的地址块中，与人工切分相同的地址块数量记为 n，则地址切分正确率为：$n/N\times 100\%$。地址的翻译正确率率指的是正确切分并正确翻译的地址块部分，占人工切分的地址块的比例。

本文随机抽取了 200 条地址作为测试语料进行翻译。通过人工分析，最终测得该模块的切分正确率为 92.2%，翻译正确率率为 84.8%。切分错误大多数是由地址信息本身较为复杂引起的，比如地址“广东省深圳市福田区福华三路与民田路交界处星河国际花园 A2 座 11”，此处的“福华三路与民田路交界处”并不符合一般的地址写法，因此发生切分错误。但是，如果把“福华三路与民田路交界处”作为词条加入词典，则根据词典的优先权，此地址可以得到正确的切分。

实验表明，中国的绝大部分地址，都可以通过拼音转换的方法得到对应的翻译。外国地址的翻译，主要依靠词典以及规则翻译。

5.3　机构名和公司名翻译模块的测试

机构名和公司名翻译模块的测试方法与地址翻译模块的测试方法类似。本文同样测试了该模块的切分正确率和翻译正确率。

本文随机抽取了 200 条机构名及公司名作为测试语料，经过切分、翻译以及人工分析，得到该模块的切分正确率为 99.2%，翻译正确率为 63%。实验表明，在机构名和公司名中，所属地信息基本都可以得到翻译，后缀的翻译效果也较好，翻译正确率率偏低主要是受机构名和公司名中间的名称部分的影响。比如公司名“吴江福华织造有限公司”，切分后得到“吴江 福华织造 有限公司”；此时“吴江”和“有限公司”可以得到很好的翻译，“福华织造”的翻译则需要依靠词典。

实验表明，机构名和公司名的大部分可以通过拼音转换的方法得到对应翻译，如果加入词典，则翻译效果可以得到极大的提升。

6　总结及将来的工作

本文主要针对题录信息的不同特征，使用词典以及规则，设计了不同的方法对人名、地址、机构名和公司名进行翻译。对于人名翻译，本文提出了词典查找、拼音转换、假名转换和同音转换的方法；对于地址、机构名和公司名，本文采用了先切分、再翻译、最后调序的流程。实验结果表明了上述方法的可行性和有效性。

为了进一步提高题录信息翻译的质量，还需要收集和整理更多有关题录信息的资料，采用更细致的方法针对性地翻译题录信息，进一步提高系统性能。在本文中，为了体现系统的稳健性，并没有充分利用词典资源，在实际系统中，可靠的词典资源将对系统性能产生重要的影响。近年来，随着网络信息资源的爆炸式增长，研究人员开始在实际系统中引入各种有用的网络资源，下一步的工作可以考虑将可靠的网络资源引入系统中，更好的提升系统的翻译质量。

参考文献

[1]　冯志伟. 机器翻译研究[M]. 北京：中国对外翻译出版社，2004.

[2]　刘群，张华平，骆卫华，孙健等译，刘群审校. 自然语言理解[M]. 北京：电子工业出版社，2005.

[3]　宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008.

[4]　吕永进，郑承萍. 外国人名、地名翻译中的汉字应用问

题[C]//第三届全国语言文字应用学术研讨会, 2003: 372-383.
[5] 孙乐乐. 中文地名翻译浅谈[J]. 科技经济市场, 2006: 358-359.
[6] 彭月华, 张向京. 地名短语文化内涵与翻译[J]. 中国科技翻译, 2008: 54-57.
[7] Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora[C]//Proc. of the 31st Annual Meeting of the ACL. 1993: 17-22.
[8] 蒋龙, 周明, 简立峰. 利用音译和网络挖掘翻译命名实体[J]. 中文信息学报, 2007, 21(1): 23-29.
[9] 苗文彦, 赵铁军, 周向根, 等. 面向机器翻译的中文机构地址切分方法[C]//2009 国际信息技与应用论坛, 2009.

《中文信息学报》征稿简则

一、《中文信息学报》主要刊登中文信息的基础理论、应用技术、中文信息处理系统及设备、中文信息的自动输入和人工编码输入、汉字字形信息、自然语言处理、计算语言学及民族语言文字信息处理及网上信息处理等方面的研究论文、技术报告、综述、通讯、简报、国内外学术活动等。

二、来稿要求和注意事项

1. 来稿内容力求正确, 论点明确, 文字简练, 数据可靠, 图表清晰, 字数不超过 8 000 字。

2. 投稿要一式三份, 计算机打印, 亦接收电子投稿。文章题目不超过 20 个字, 须有 200 字中文摘要和英文摘要。英文文摘应符合英文语法, 概括论文内容, 包括研究目的、方法、结果和结论。中英文摘要均应包括题目、作者姓名、单位名称、城市名、邮编、摘要、关键词。写明中图分类号。

有基金项目支持的写明基金名称、编号。

给出作者信息, 包括姓名, 出生年, 性别, (学位), 职称, 主要研究方向。

3. 文中图、表放在文稿中相应位置, 并注明图号、图注。图中文字用六号宋体。

4. 文中外文字母、符号要分清大小写、正斜体; 上下角标的位置高低应区别明显; 容易混淆的字母数字, 用铅笔批清(如 O[英大] 0[数字]); 文中需用黑体字之处, 在字下加波纹线。全文计量单位要一致, 或中文, 或符号。

5. 参考文献只列最主要的, 必须是已公开发行的书刊才能列入, 最少不得少于 5 条。文献按文中出现先后次序编排, 书写格式为:

专著:[序号] 作者. 题名[M]. 出版地: 出版者, 出版年: 起止页码.

期刊:[序号] 作者(多作者用逗号分开, 超过 3 个者用“, 等”代替). 文章题目[J]. 刊物名称, 年代, 卷数(期数): 起止页码.

论文集:[序号] 作者. 题名[C]//编者. 论文集名. 出版地: 出版者, 出版年: 起止页码.

学位论文:[序号] 作者. 题名[D]. 保存地点: 保存单位, 年份.

报告:[序号] 作者. 题名[R]. 保存地点: 保存单位, 年份.

报纸文章:[序号] 作者. 题名[N]. 报纸名, 出版日期(版次).

标准:[序号] 制定单位. 标准编号, 标准名称[S]. 出版地: 出版者, 出版年.

专利:[序号] 专利所有者. 专利题名: 专利国别, 专利号[P], 公开日期.

电子文献: 主要责任者. 电子文献题名[电子文献标识/载体类型]. [发表或更新日期]. 电子文献的出处或可获得地址.

电子文献标识:[DB] —数据库 [CP] —计算机程序 [EB] —电子公告

电子文献载体类型:[OL] —联机网络 [MT] —磁带 [DK] —磁盘 [CD] —光盘

6. 来稿请勿一稿二投, 文责自负。不录用稿件概不退还, 请自留底稿。来稿一经发表, 按规定付给稿酬, 并赠送单行本 2 册。

来稿请寄: 北京 8718 信箱《中文信息学报》编辑部收, 邮政编码 100190, 电话: 010-62562916。本刊也接收电子投稿, 请以附件方式, 将 WORD 文档发至: cips@iscas.ac.cn。请写明作者工作单位、通信地址(邮政编码)、电话(手机)、E-mail。