

融合语料库的论文作者姓名中英自动翻译研究

于夏薇¹ 袁军鹏²

1. 中国科学技术信息研究所科学计量与评价研究中心 北京 100038;

2. 中国科学院文献情报中心 北京 100190

摘要 中文姓名英译现状非常混乱,在科研评价中根据作者英文名检索时有很大的问题。论文作者姓名中英自动翻译研究旨在为科研机构评价人才提供准确的参考,使其从繁琐的匹配论文工作中解脱出来。针对论文作者这一固定、较为封闭的群体,本文研究中、英文的姓名翻译对照规则,设计中英文姓名对照辞典;调研了中文姓名英译存在的问题,归纳总结出了适合长远发展以及推广的中文姓名英译的规则:《中国人姓名汉语拼音字母拼写法》;根据论文编号、第一作者、英文名写法筛选形成中英文姓名对照字典建设语料库,在此基础上设计与实现论文作者姓名中英自动翻译模型,规避了复姓、多音字、生僻字等一些复杂情况。该模型的设计与开发针对论文作者语料库,根据相似性算法优化匹配符合作者需求的规范化英文名,较之一般翻译软件具有较高的准确率。

关键词: 语料库; 姓名翻译; 相似度; 汉语拼音方案

中图分类号: F204

开放科学(资源服务)标识码(OSID)



Research on Chinese and English Automatic Translation Combined Authorship Corpus

YU Xiawei¹ YUAN Junpeng²

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;

2. National Science Library, Chinese Academy of Sciences, Beijing 100038, China

Abstract The situation of Chinese name translation is confusing in the evaluation of scientific research in terms of the author's English name. The author of the paper is an accurate reference for the evaluation of talents by scientific research institutions, so that it can be freed from the work of the cumbersome

基金项目: 国家自然科学基金项目: 中英文论文中的中国作者姓名消歧研究(71473236)。

作者简介: 于夏薇(1993-), 硕士研究生, 研究方向: 科学计量学、数据挖掘; 袁军鹏(1973-), 研究员, 博士, 研究方向: 科学计量学、数据挖掘、科技政策、技术创新管理, Email: yuanjp@mail.las.ac.cn。

matching papers. This paper studied the Chinese-English names in the context of the translation between Chinese and English. The research reviewed and summarized the existing problems in the translation of Chinese names. In addition, the paper summarized the rules which are suitable for long-development of the Chinese name translation: "Chinese Name Chinese Phonetic Alphabet Spelling Method". This study constructed the Chinese and English name corpus based on the ID of papers, the Chinese name and English name of the first author. On the basis of this, the design and implementation of the English automatic translation model for the author's name to avoid the compound surname polyphonic words, uncommon characters and other complex situations. The design and development of the model are aimed at the author corpus, and the similarity algorithm is used to optimize the matching English name to meet the needs of the author, which has higher practical value compared with the general translation software.

Keywords: Corpus; name translation; similarity; Chinese Pinyin program

1 引言

中文姓名的英译主要有两种表达方式：直接以汉语拼音表达，如“王晓梅”为“Wang Xiaomei”。姓名顺序保留汉语习惯，即姓在前，名在后。姓和名作为两个组合分开书写，名的部分不分开；或者起一个英文名，再与汉语拼音的姓组合，如张朝阳：Charles Zhang。随着中国科研人员在国际学术期刊上发表的文章越来越多，对应翻译的英文名混乱问题也越来越严重。以 Web of Science 为例：论文作者署名是时一般为姓全拼空格后+名首字母，根据该缩写规则查找该作者的论文会出现检索结果的混乱。在科学计量学的微观评价中，如何区别中英文人名，检索出一个作者的所有英文论文会有很大的困难。由此可见，论文姓名中英自动翻译非常重要。

本研究结合我国中文论文作者姓名自动翻译的相关标准及特点，对现有通用的姓名自动翻译原型进行归纳总结；搜集专有的中文科技

论文作者中英文名构建语料库；结合相关算法，实现高准确率的集成语料库自动翻译原型，对原有中文作者姓名翻译体系的完善提供参考。同时，在基于语料库的基础上进行论文作者的姓名自动翻译研究，旨在为科研机构评价人才提供准确的参考，使其从繁琐的匹配论文工作中解脱出来，有效地开展科研评价工作。

本研究形成的中文作者姓名英译语料库可以规范中文科技论文的作者姓名翻译问题，使科学评价时论文作者的学术成果客观性有了增强，提高了广大科研工作者学习检索文献的效率，有助于作者中英文姓名消歧研究的进一步实施。

2 相关研究综述

2.1 中文姓名英译混乱的原因

中文姓名英译混乱的原因主要有：注音法的演变、少数民族姓名与一般汉语人名不同以及国内姓名翻译适用的不同标准的干扰等。

2.1.1 注音法的演变

威妥玛式拼音法^[1]是1867年由威妥玛等人合编的注音规则,在《汉语拼音方案》制定前被广泛用于人名、地名注音。该注音法将汉语声母z, c分成ts', ts或tz, tz'两种,把s分成s, ss和sz三种,排版印刷时容易产生混乱,如“张”、“常”不分等。历史上影响较深的注音法还有:利玛窦的拉丁字母式拼音、《国语罗马字拼音法式》、《中国拉丁化新文字的原则和规则》以及肯尼迪设计的耶鲁拼音。

2.1.2 少数民族姓名

少数民族姓名字符串较长,有中间连接符,一般来说,按少数民族语读音拼写。如阿不都热苏力 Abdurusul 和司马义·萨依木 Simayi Sayimu。少数民族人名汉字英译按照一定的规则,准确地转写,或是对别的民族的人名用本民族的语言文字准确地表达,是民族进步、社会发展的需要,也是当代语言文字发展的需要^[2]。

2.1.3 姓名英译规则

(1) 国家标准:1978年颁发的《国务院批转关于改用汉语拼音方案作为我国人名地名罗马字母拼写法的统一规范的报告》^[3]是中国人名英译的理论依据。该标准规定汉语姓名英译时,姓在前,名在后,中间为空格,复姓连写,双名连写,无连接符号“-”,姓与名的第一个字母均大写。国家规定改善了早期姓名翻译存在的问题,然而还存在复姓、双名连写问题:如无法判断“Sun Yue”是“孙玉娥”还是“孙悦”^[4]。

国家标准在国际上不断被推行,作为中文论文作者姓名署名时的参照依据:1979年,联合国秘书处采用汉语拼音作为转写中国人名、地名的标准,要求所有成员国必须执行。1982年,

《汉语拼音方案》正式成为汉语罗马字母拼写的国际标准,同时通过采用汉语拼音作为情报网络、图书馆和文献管理中拼写汉语的国际标准,确立了汉语拼音为汉语国际罗马化的唯一标准。2015年12月,国际标准化组织正式出版了《ISO 7098 信息与文献中文罗马字母拼写法》2015年修订本^[5],增加了汉语拼音分词连写的规则,增强了《汉语拼音方案》在国际语言生活中的可操作性。

(2) 中国学术期刊标准:1999年《中国学术期刊(光盘版)检索与评价数据规范》在全国近3500种入编期刊中试行^[6,7]。该标准规定姓的全部字母大写,名字的首字母大写,双名中间加连字符,名字不缩写。该方案避免了姓和名的混淆和缩写问题,但因与国家标准相冲突,使作者在署名时有很大的困扰,且存在一定的缺陷,连字符较难被识别,作者姓名检索具有较大困难。目前不同学术期刊对于作者署名的格式规范并不相同,较难统一实施。

本节总结归纳了中文姓名英译混乱的原因,由此得出亟待解决的是确立一个国家标准,来规范化不同场景下已经存在的混乱问题。汉语拼音方案已经在国际上推广且被承认,具有很好的示范价值,可以作为未来姓名规范化的标准参考;对于少数民族的人名写法,其现有的英译姓名最好保留而不加以改变;中文姓名回译困难,可将人名全称附在出版物中译名之后。

2.2 中文姓名中英自动翻译综述

如何对庞大的中文人名群体进行统一化的

规范化音译是目前语言学, 机器翻译, 统计学尚待解决的问题之一。作者姓名中英自动翻译是命名实体翻译问题的一个特例, 是一个非常具有挑战性的问题。

2.2.1 基于算法的姓名自动翻译研究

新加坡资讯通信研究院李海洲^[8]等所著《TRANSLITERATION》中通过探索各种音译和验证技术, 探讨专有名词, 特别是个人名字的英汉对照音译范式。在自动语音识别(ASR), 文本到语音(TTS), 跨语言信息检索(CLIR)和机器翻译(MT)等语言处理中, 越来越需要翻译从一种语言到另一种语言的词汇单词。在实践中语义翻译不可用则通过音译进行翻译, 而音译是指通过保留原语言中的单词声音将一种语言翻译成另一种语言的方法。

国防科技大学李亚男等^[9]提出了一种语言映射方法, 可以将给定社交网络中的用户名的不同类型的汉字翻译成相应的拼音单词; 鉴于用户身份数量可能非常大, 使用Hadoop和Spark框架来克服效率问题。新疆大学信息科学与工程学院周颀^[10]利用VB中的字符串处理函数及SQL语句数据查询技术等, 有效处理了哈文词中每个字母因位置不同写法不同而引起的计算机编码不同的问题, 从而实现了哈萨克族人名自动翻译功能。

针对已经翻译成英文的中文人名回译, 也有相关的研究。清华大学刘颖教授^[11]使用组合方法: 熵对齐模型和网络挖掘将英文人名翻译成中文人名。熵对齐模型利用人名和姓氏字典、双向条件概率和音译相似度; 网络挖掘利用规则、线索词、音译相似度和条件概率。实验结果表明, 组合方法可以提高大规模并行语料库

的精确度和召回率。

2.2.2 基于语料库的自动翻译研究

中国台湾中央研究院李汉铭等^[12]提出了一种从网络语料库挖掘中文名称翻译的方法。该方法提供了中文名称的分类, 基于语音相似性, 最小距离和附近出现的次数三种特征, 通过使用查询扩展技术和支持向量机(SVM)来提取名称翻译候选者, 实验结果显示可以正确地翻译大多数中文姓名。北京大学常宝宝^[13]提出了基于词语关联度进行多词组合单位的识别方法, 利用汉英双语语料进行汉英翻译等价单位的抽取, 并提出利用范畴假设来改进抽取算法的效率。

新疆师范大学阿力木·木拉提^[14]通过维吾尔音节切分与维吾尔人名特征相结合方法, 基于字形的DOM音译框架进行实验, 构建维吾尔人名语料库, 提高维吾尔人名汉语自动音译系统正确率30%。借鉴少数民族语言英译软件的开发与实现, 在语料库的扩大中寻找提高准确率的实现算法, 是研究自动翻译软件的一个方向^[14]。

对于简繁体汉字自动翻译方法, 上海交通大学庞祯军^[15]教授提出一种基于对照表以及语义相关性的自动翻译方法。使用繁体字语料库构建语言模型以及收集到的维基百科词组对照表, 通过计算语言模型的分数来达到提升简体字转换成繁体字的准确率, 同时考虑一个句子中单词组合的几率, 几率越高代表越可能符合正确的文法。

对姓名的中英自动翻译相关研究较少, 其中多采用机器翻译方法, 中英对照词典映射等方法。国内学者对中文姓名的英译研究大部分

集中在翻译现状和对策,其中有一些针对少数民族语言的拼音规范化进行自动翻译软件的设计和实现,对于姓和名的区分度并不高,多音字的情况下也难以处理。

语料库是翻译研究的基础,在句子和文本的翻译中,国内公开使用的语料库特别是双语平行语料库屈指可数,自建语料库虽多,但由于人力和资源所限,在规模和加工深度上并不理想。针对句子和文本的翻译使用语料库的研究对姓名中英自动翻译提供了一个很好的参考,姓名中英语料库本身并不复杂,只需要构建一个中英文姓名对照字典,可以有效规避语料库数据庞大难以处理和运算的问题,并能充分利用语料库的特性,实现论文中文作者的姓名自动翻译。目前还没有根据语料库进行论文中文姓名进行中英自动翻译的针对性的研究。

3 基于语料库的作者姓名自动翻译研究

论文作者是一个比较固定且封闭的群体,将论文作者姓名翻译成英文具有一般姓名翻译的特征。不同之处是论文作者自己提供的对应英文名具有较高的准确性,规避了一些复姓、多音字、生僻字的情况。如果加以优化和格式规范,即可得到规范的英文名。所以可以参照现有的中文论文数据库中的作者自己提供的中英文姓名来构造语料库^[16]。

本文通过构建论文作者中英文姓名语料库,缩小姓名中英自动翻译适用范围,更加专业化和针对性,结合相关相似度算法,匹配修正英译名,实现能提高检效率与准确率的集成语料

库自动翻译原型,对原有中文作者姓名翻译体系的完善提供参考。

3.1 语料库建设

语料库^[17]是用科学的方法系统化收集起来的语言文集,存放的是在语言的实际使用中真实出现过的语言材料;真实语料需要经过加工才能成为有用的资源^[18]。语料库中包含了大量的相关信息,可以研究各篇章的语言特征和组合情况,从而分析构建篇章的语篇或语体特征^[19]。

由于本文研究对象是论文作者这个特殊群体,构建的语料库充分体现了作者姓名的特点,分析论文作者所附带的各种属性:如比较稳定封闭,一般不会改变,只需做定期添加更新等来规范与缩小语料库的范围。

3.1.1 数据来源

数据来源为截止2015年的中国科学技术信息研究所CSTPCD数据库,筛选所有中文科技论文的中文姓名和对应的英文姓名,统计中英文姓名相同的记录去重,设计中英文姓名对照字典。

3.1.2 语料库规模

将所有论文记录的唯一ID、第一作者名、英文名写法筛选形成中英文姓名对照字典,即语料库^[20],去重后共得到4085160条记录。如图1所示:

由于中文姓氏存在复姓,如“司马”、“诸葛”等,为了简化流程,增加姓氏判断的准确性,构建复姓字典,在模型设计时直接调用即可。共有82个复姓姓氏,前20个常用复姓如表1所示:

	作者	ZZe	出现频次
19	李明	LI Ming	7284
20	张静	ZHANG Jing	7241
21	王涛	WANG Tao	7168
22	张军	ZHANG Jun	6795
23	李莉	LI Li	6793
24	李敏	LI Min	6681
25	李刚	LI Gang	6590
26	张鹏	ZHANG Peng	6570
27	张杰	ZHANG Jie	6527
28	王敏	WANG Min	6479
29	刘勇	LIU Yong	6454
30	王健	Wang Jian	6447
31	刘军	LIU Jun	6365
32	王鹏	WANG Peng	6182
33	张健	ZHANG Jian	6180
34	张华	ZHANG Hua	6167
35	王辉	Wang Hui	6165
36	李斌	LI Bin	6131
37	李娜	LI Na	6071
38	李清	LI Tao	6052
39	张强	ZHANG Qiang	5980
40	李鹏	LI Peng	5966

图1 中英文姓名对照字典

表1 复姓对照表

复姓姓氏			
欧阳	东方	夏侯	澹台
太史	独孤	诸葛	皇甫
端木	南宫	尉迟	宗政
上官	万俟	公羊	濮阳
司马	闻人	赫连	公冶
.....			

3.1.3 语料库类型

语料库主要有以下四类：

- (1) 单语语料库 (Monolingual Corpus), 仅收集一种语言的语料库。
- (2) 双语 / 多语语料库 (Bilingual / Multilingual Corpora), 指两种或两种以上的语言文本构成的语料库, 有平行语料库和类比语料库两种形式。
- (3) 平行语料库 (Parallel Corpora), 由原文文本及其平行对应的译语文本构成的双语语料库。
- (4) 类比语料库 (Comparable Corpora), 由不同语言的文本或同一种语言不同变体的文

本所构成的两个或两个以上的语料库。

用于翻译研究的语料库, 一般是一个双语库或者多语库 (Bilingual / Multilingual Corpora)。本文构建的中英文姓名对照表属于双语语料库和平行语料库, 即只有两种语言, 由原文及其平行译文组成。复姓对照表属于单语语料库。

3.2 姓名中英自动翻译模型设计

3.3.1 操作环境

该模型以 Windows 系统为操作平台, 使用关系数据库 SQL SERVER 2008 进行 SQL 语句查询处理等数据库操作, 以 python 程序语言为开发工具, 使用数据库连接技术、字符串处理函数和相似度算法, 从而基本实现了中文作者姓名的中英自动翻译功能。

3.3.2 中文姓名切词

中文姓名的分词是姓名自动翻译的关键步骤。SQL 语言是一个通用的、功能极强的关系数据库语言^[21], 集数据查询、数据操纵、数据定义和数据控制功能于一体。本文中使用 SQL 语句对数据库中的中文姓名进行切词处理, 同时将每个文字转变为对应的拼音。

从给定的拼音词中根据“转拼音规则”采用正向最大匹配法切分出正确的拼音字, 分为姓和名两部分。按照字符转拼音法则将每个字符转化为拼音, 作为待处理翻译结果。具体语句以首字母大写转化为例如下:

```
begin
set @crs=substring(@str,@str_len,1)
select @str_len=@str_len-1,@result=
case
```

```

when @crs>='发' then 'F'
when @crs>='婀' then 'E'
when @crs>='叮' then 'D'
when @crs>='噤' then 'C'
when @crs>='八' then 'B'
when @crs>='吖' then 'A' .....
else @crs end+@result

```

end

3.3.3 相似度算法

该模型中的相似度算法采用 Jaro Winkler 算法, 该算法主要用于记录链接领域^[22]。设两字符串的 Jaro Distance 是 d_j , 两字符串拥有的共同前缀的长度为 L 。Jaro-Winkler 算法认为, 起始部分相似的字符串, 其相似的可能性更高, 给予 L 一个权重, 即前缀范围因子 p 。对于要比较的两个字符串, Jaro-Winkler Distance 为:

$$d_w = d_j + (lp(1 - d_j))$$

其中: 是最后得分; L 是前缀部分匹配的长度, 最大是 4 个字符; p 是一个范围因子常量, 用来调整前缀匹配的权值不能超过 0.25, 一般 p 赋值为 0.1。越大, 说明两字符串相似度越大。特点: 速度快, 两个被比较的字符串越是靠前, 则对该两字符串打分越低。该算法主要适用于短字符串的比较, 如人名等。

3.3.4 功能流程设计

该模型使用 Python 语言能够快速搭建窗体与运行界面, 结合了用户操作简单、响应速度快捷, 界面设计要用户友好、简洁明了等特点。根据应用需求分析, 翻译界面包括查询浏览与自动翻译, 后台数据处理主要包括数据库添加修改等信息传送的服务功能。翻译界面功能模块如图 2 所示

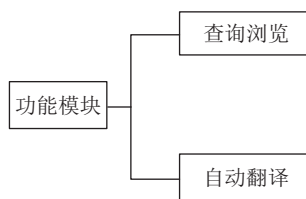


图2 翻译界面功能模块图

设计论文作者姓名中英自动翻译流程图如

下图 3 所示:

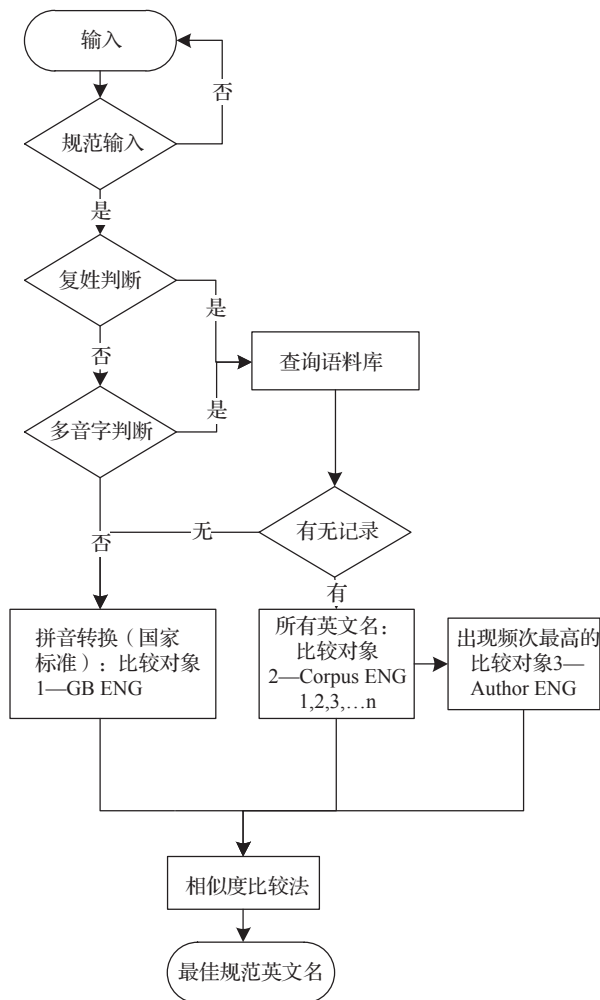


图3 论文作者姓名中英自动翻译流程图

具体步骤如下:

- (1) 输入中文姓名, 判断是否规范输入, 是则下一步, 否则重新输入;
- (2) 系统根据复姓表判断姓氏是否为复

姓，是则转（3），否则转（7）

（3）查询语料库中该中文名的所有英文名以及出现频次，所有英文名为作为比较对象2：

Corpus ENG 1, 2, 3……, n;

（4）若有记录，选取频次最大的英文名作为比较对象3：Author ENG；无记录则转（5）

（5）系统将输入的中文名按照拼音转换翻译为国家标准下的英文姓名作为比较对象1：

GB ENG；

（6）相似度比较法：

将比较对象1：GB ENG 和比较对象3：Author ENG 进行相似度判断，设定相似度阈值设定为0.9，大于阈值则认为两个对象一致则输出规范英文名 GB ENG；

小于阈值为不一致，将比较对象2：

Corpus ENG 1, 2, 3……, n 和比较对象1：

GB ENG 进行中文切词，形成姓与名两部分。

对 Corpus ENG 中所有的姓和 GB ENG 的姓进行相似度计算，再分别对名进行相似度计算，阈值设定与之前相同。

Corpus ENG 中的每个英文名与 GB ENG 的相似度计算公式为：姓的相似度加名的相似度，选取相似度最大的作为最佳规范英文名；

（1）系统判断姓氏是否为多音字，是则转（3），否则转（5）

（2）若无比较对象则按照国家标准生成的 GB ENG 输出；

3.3.5 模型功能展示

自动翻译功能如下图4所示：

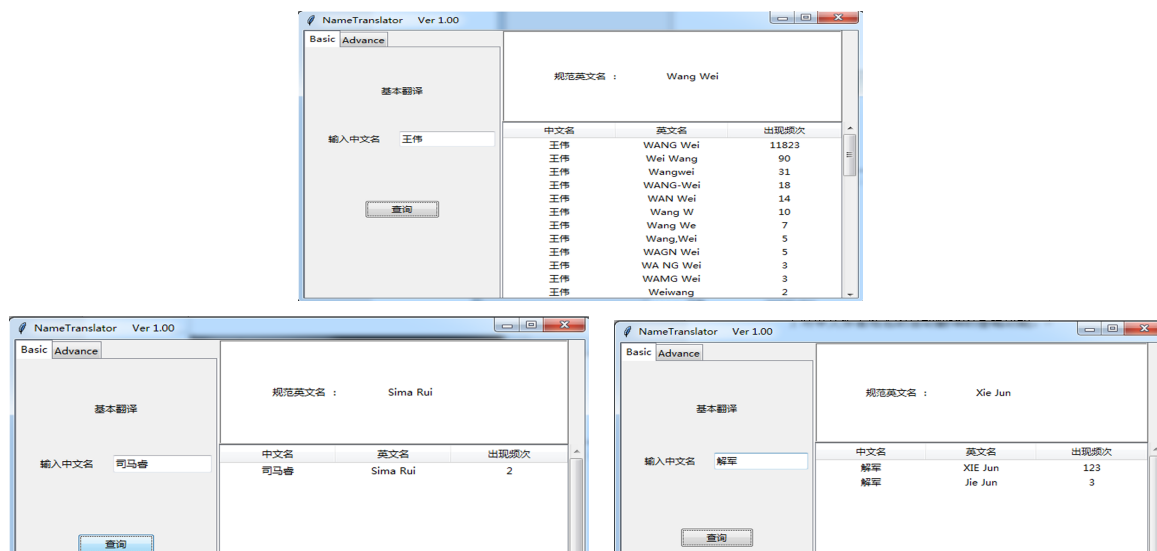


图4 自动翻译界面展示

3.3 评价测试

人名的专有性限制了翻译的规则和方法。在线翻译软件采取规则较为单一，对于一些复姓和多音字等具有变化的姓名难以做到准确的翻译^[19]，较难满足论文作者姓名自动翻译的需

求。本文采用基于语料库的姓名中英自动翻译方法，研究开发论文中文作者姓名自动翻译模型，使作者姓名在一个规范的准则里得到一个尽可能贴近作者本身的准确结果，使之与作者本身提供的英文名相符，能够被广大科研工作者使用和采

纳,具有较高的可操作性和应用价值。

以太阳风名址翻译机为例,进行比较分析。在输入例如“张华、赵丽”等常见人名时,该翻译网站得出的译文结果显示为机器翻译,根据该网站的翻译规则进行论文作者的逐一翻译评测,与本文设计的基于语料库的姓名自动翻译模型相对比,由于所涉姓名较多,无法将特殊情况一一列举,只选取一些常见情况部分展示如表2。

表2 不同姓名翻译软件以及与Author ENG的对比

输入	太阳风名址翻译机	基于语料库的最佳规范英文名	Author ENG
王敏	Wang Min	Wang Min	WANG Min
刘勇	Liu Yong	Liu Yong	LIU Yong
解军	Xie Jun	Xie Jun	XIE Jun
司马睿	Sima Rui	Sima Rui	Sima Rui
阿布拉提	Abra Ti	Ablatti	Ablatti
李雨中生	Liyu Zhongsheng	Li Yuzhongsheng	/
朝乐门	Chaolemen	Chao Lemen	Chao Lemen
王伟	Wang Wei	Wang Wei	WANG Wei
王辉	Wang Hui	Wang Hui	Wang Hui
李斌	Li Bin	LI Bin	LI Bin
王斌	Wang Bin	Wang Bin	Wang Bin
李宁	Li Ning	Li Ning	LI Ning
张辉	Zhang Hui	Zhang Hui	Zhang Hui
杨静	Yang Jing	Yang Jing	Yang Jing
李磊	Li Lei	Li Lei	LI Lei
刘艳	Liu Yan	Liu Yan	LIU Yan
.....			

两种翻译软件下的不同翻译结果

(1) 一眼看去不像名字的姓名:如“朝乐门”,翻译结果为:“Chaolemen”,明显不符合人名拼音规范;

(2) 四个字的单姓:如“李雨中生”,翻译结果为“Liyu Zhongsheng”,误以复姓规则进

行英译。

(3) 知名人物的论文姓名:不符合特殊的论文姓名拼写规范。

而在基于语料库的论文作者姓名中英自动翻译模型中,这些问题都得到了很好的解决。通过采用复姓表,预先处理姓名的复姓情况;同时通过《中国人名汉语拼音字母拼写法》规范中文姓名翻译成英文的格式,并且通过语料库的高度可参照性能够规避一些多音字、方言、中西结合的情况,提高翻译的准确率。相似性的比较能在作者提供的 Corpus ENG 中选择一个最匹配的英文名来进行之后的规范化流程。阈值设定在合理范围内,使结果既具有较高的准确率,也避免了作者自己提供的英文名出现的误差和错误情况。

在模型输出最终规范英文名后与对该英文名和统计频次最高的英译结果 Author ENG 进行大规模比对,发现除去格式外,两个比较对象的写法总是一致的,添加更新一些中英文姓名对照记录,再次比对,结果也不会改变。由此看出,基于语料库的设计能够保证得到的规范英文名拥有论文作者自己提供的准确写法,一般认为作者自己提供的是真实可用的写法;在国家标准的格式规范下,重组 Author ENG,即能够得到最佳规范英文名。双重设计对于翻译模型的准确率有了很大的提高,优化了自动翻译原型的结果。

4 结论与展望

本文调研并完善了中文姓名英译存在的问题,归纳总结出了适合长远发展以及推广的中

文姓名英译的规则:《中国人姓名汉语拼音字母拼写法》。并在基于语料库的基础上设计与实现论文作者姓名中英翻译模型,较之一般在线翻译软件准确率有了很大的提高,规避了复姓、多音字、生僻字等一些复杂的翻译情况。该模型的设计与开发极具针对性地通过相关算法得出符合作者自身需求的规范化英文名,较之一般翻译软件具有较高的准确率。

论文作者姓名中英翻译模型建设具有积极意义,对于作者的检索时间也将大大减少。今后可以考虑将不同规则下的标准纳入考虑范围,扩大语料库,使用户可以自由选择不同规则下的规范英文名,从而具有更高的应用价值,减少作者署名时发生的错误。同时可以扩大语料库的容量,及时更新,在获取真实的中英文姓名的语料基础上,使该模型发挥更大的作用。

参考文献

- [1] 孙伟杰. “威妥玛式”拼音研究[D]. 长春: 吉林大学, 2009.
- [2] 陈毓贵. 规范少数民族人名汉字音译转写促进语言文字规范化建设[J]. 语言与翻译, 2003(2):3-6.
- [3] 原新梅. 学术期刊作者姓名的拼写规范问题[J]. 语言文字应用, 2002(1):66-70.
- [4] 黄一心. 国内期刊英译中国人姓名拼写格式应该规范[J]. 中华流行病学杂志, 2003(7):100.
- [5] 李宇明. 中华文化迈向国际新步伐——写在中文罗马字母拼写法国际标准(ISO 7098:2015) 修订出版之时[N]. 光明日报, 2016-05-01(7).
- [6] 辛献云. 谈汉语人名翻译的规范问题[J]. 解放军外国语学院学报, 2005(1):65-69.
- [7] 缪宏建, 徐鸿飞, 沈志宏. 关于中国学术期刊(光盘版)检索与评价数据规范的几点思考[J]. 中国科技期刊研究, 2000,11(6):376-377.
- [8] Li H, Bai S, Kuo J S. TRANSLITERATION[M]. 2006.
- [9] Li Y, Zhu J, Zhou Z, et al. Connecting Chinese Users Across Social Media Sites[C]. International Conference on Material, Mechanical and Manufacturing Engineering. 2015:1273-1279.
- [10] 周颢. 哈萨克族人名规范化译写软件设计[J]. 计算机与现代化, 2004(3):34-36.
- [11] Liu Y, Xiao T. Translation of English-Chinese Person Name Based on Dictionary, Bilingual Corpus and Web Mining[C]. International Conference on Natural Computation, 2014.
- [12] Yang K H, Chen W D, Lee H M, et al. Mining Translations Of Chinese Names From Web Corpora Using A Query Expansion Technique And Support Vector Machine[C]. Web Intelligence & Intelligent Agent Technology Workshops IEEE/WIC/ACM International Confere, 2007:530-533.
- [13] 常宝宝. 基于汉英双语语料库的翻译等价单位自动获取研究[J]. 术语标准化与信息技术, 2002(2):24-29.
- [14] 阿力木·木拉提, 艾孜尔古丽, 玉素甫·艾白都拉. 现代维吾尔语人名汉字音译转写关键技术研究[J]. 计算机工程与应用, 2014(9):209-213.
- [15] 庞祯军, 姚天昉. 基于对照表以及语义相关性之简繁汉字转换[J]. 计算机工程与应用, 2015(4): 115-119.
- [16] 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(1):97-101.
- [17] 闻永毅, 樊新荣. 浅谈建设中医英语语料库的意义[J]. 陕西中医学院学报, 2003(5):76-77.
- [18] 李文良, 朱燕. 英美文学语料库构建及其应用价值探析[J]. 唐山师范学院学报, 2009(4):121-123.
- [19] 何安平. 语料库研究的层面和方法述评[J]. 外国语, 1999(3):10-17.
- [20] 庞祯军, 姚天昉. 基于对照表以及语义相关性之简繁汉字转换[J]. 计算机工程与应用, 2015(4): 115-119.
- [21] 萨师煊, 王珊. 数据库系统概述(第3版)[M]. 北京: 高等教育出版社, 2000.
- [22] Winkler W E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage[EB/OL]. [2015-09-09]. http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf.