跨语言命名实体翻译对抽取的研究综述*

王志娟[1,2]，李福现

1. 中央民族大学信息工程学院，北京 100081；2. 国家语言资源监测与研究中心少数民族语言分中心，北京 100081

摘要：跨语言命名实体对于机器翻译、跨语言信息抽取都具有重要意义，本文从命名实体的音译、基于平行/可比语料库的跨语言命名实体对齐、基于网络挖掘的跨语言命名实体对翻译抽取三个方面对跨语言命名实体翻译对抽取的研究现状进行了总结。音译是跨语言命名实体翻译对抽取的重点内容之一，基于深度学习的音译模型将是今后的研究重点；目前跨语言平行/可比语料库的获取和标注直接影响基于语料库的跨语言命名实体对齐的深入研究；基于信息检索和维基百科的跨语言命名实体翻译对抽取研究将是跨语言命名实体翻译对抽取研究的趋势。

关键词：命名实体翻译对；音译；命名实体对齐；网络挖掘

中图分类号：TP391　　**文献标识码：**A

Survey on cross-language named entity translation pairs extraction

Zhijuan Wang[1,2], Fuxian Li

(1.The College of Information Engineering, Minzu University of China, Beijing, 100081, China;
2. Minority Languages Branch, National Language Resource Monitoring & Research Center, Beijing, 100081, China)

Abstract: Cross-language named entity translation pairs are very important for machine translation, cross-language information retrial and so on. We made a survey of cross-language named entity translation pair extraction on three aspects: named entity transliteration, named entity alignment based on parallel/comparable corpus, and cross-language named entity extraction based on web mining. Transliteration is vital for cross-language named entity translation pair extraction. Transliteration model based on deep learning will be the key in future studies. The constructing and annotation of cross-language corpus are bottle necks for research on named entity alignment based on parallel/comparable corpus. And cross-language named entity extraction based on information retail and Wikipedia will be the trade in the future.

Key words: named entity translation pairs; transliteration; named entity alignment; web mining

1 引言

命名实体是承载文本信息的重要语言单位，主要包括人名、地名和组织机构名三种类型。命名实体识别（named entity recognition）研究开始于1991年，Lisa F. Rau 在第七届 IEEE Conference on Artificial Intelligence Application 上介绍了从文本中抽取公司名的方法[1]（一种组织机构名），1996 年的 MUC（Message Understanding Conference）正式提出了命名实体（Named Entity）的定义。

随着时间的推移，有关命名实体识别的研究从英语扩展到汉语、朝鲜语、阿拉伯语、日语等，近年来，一些资源较少语

* **收稿日期：** 2016.6.5　　**定稿日期：2016.8.15**

基金项目：国家自然基金重点项目（61331013）；国家语委科研项目（WT125-46）

作者简介：王志娟（1977——），女，博士/副教授，研究方向为信息抽取、机器翻译；李福现（1973——），男，硕士/高级工程师，研究方向为数据挖掘、机器学习。

言（Low Resource Languages）的命名实体识别引起了一些研究者的关注，如豪萨语、希腊语、越南语、保加利亚语等命名实体识别，我国一些学者也开展了面向我国少数民族语言如藏语[2-3]，维吾尔语[4-5]、蒙古语[6]的命名实体抽取研究。

随着各语种命名实体抽取研究的深入，人们渴望消除语言障碍、实现跨语言的命名实体抽取的愿望越来越迫切。最早的跨语言命名实体抽取研究开始于 2002 年的英语和阿拉伯语之间，Al-Onaizan 等[7]使用音译模型以及词典实现了英语和阿拉伯语之间的命名实体对翻译，之后，一些学者对英语-日语、英语-朝鲜语、英语-汉语的跨语言命名实体翻译对抽取进行了深入研究。

跨语言命名实体翻译对抽取对于很多跨语言自然处理研究如跨语言信息检索、机器翻译等都具有重要意义，文献[8]指出：跨语言命名实体翻译对抽取方法主要有命名实体的直接翻译、基于跨语言语料库的命名实体对齐以及基于网络挖掘的跨语言命名实体对抽取三种，本文将从这三个方面系统、深入介绍跨语言命名实体翻译对抽取相关的国内外研究。

2 基于翻译的跨语言命名实体翻译对抽取

命名实体直接翻译是早期获取跨语言命名实体翻译对的主要方法，由于跨语言命名实体对的翻译通常基于音译（人名、简单地名）和音译/意译组合（组织机构名和复杂地名）。因此本文重点介绍基于音译的命名实体翻译。

2.1 目前音译研究涉及的语言

根据文献[9-10]，目前音译研究涉及的语言有 24 种、33 种语言对（language pair），表 1 所示为音译研究涉及的语言的基本情况。由表可见：

(1)目前音译研究主要围绕英语展开，一共有 14 种语言、22 个语言对，其中英语到其他语言的双向音译研究有8种语言，分别是：汉语、日语、朝鲜语、泰语、西班牙语、阿拉伯语、印地语、旁遮普语，另外还有6种语言以英语作为音译源语言或者目标语言进行音译研究。

表 1 音译研究涉及的语言情况

序号	语言对	说明
1	英语-汉语	
2	汉语-英语	
3	英语-朝鲜语	
4	朝鲜语-英语	
5	英语-泰语	
6	泰语-英语	
7	英语-日语	
8	日语-英语	
9	英语-西班牙语	
10	西班牙语-英语	
11	英语-俄语	
12	英语-希伯来语	希伯来语为以色列官方语言
13	英语-拼音	
14	乌尔都语-英语	乌尔都语为巴基斯坦官方语言
15	波斯语-英语	波斯语为伊朗、阿富汗、塔吉克官方语言
16	孟加拉语-英语	孟加拉语为孟加拉国官方语言、印度方言
17	芬兰语-瑞典语	
18	西班牙语-汉语	
19	维吾尔语-汉语	维吾尔语为中国少数民族语言
20	拼音-汉语	
21	阿拉伯语-英语	阿拉伯语是埃及、沙特阿拉伯、摩洛哥、伊拉克等 20 余个国家的官方语言
22	英语-阿拉伯语	
23	阿拉伯语-法语	
24	英语-印地语	印地语和英语是印度的官方语言，其余为印度方言。
25	印地语-英语	
26	英语-旁遮普语	
27	旁遮普语-英语	
28	英语-奥里亚语	
29	英语-泰卢固语	
30	英语-泰米尔语	
31	英语-卡纳达语	
32	马拉雅拉姆语-英语	
33	旁遮普语-印地语	

(2)印度是一个语言资源丰富的国家，除了印地语和英语两种官方语言外，还有 20 余种方言被广泛使用，目前和印度语言相关的音译一共涉及 8 种语言、有 11 种语言对。

(3)和汉语相关的音译涉及英语、西班牙语、维吾尔语、拼音 4 种语言，5 个语言对。

2.2 音素 vs 形素

在音译的过程中，音译可以分为基于音素的音译（源语言→源语言发音→目标语言发音→目标语言）、基于形素的音译（源语言→目标语言）以及结合音素和形素的混合音译。

分析文献[11]提到基于不同音译模型、57 个语言对的音译发现：约 30%的语言对采用基于音素的音译，约 63%的语言对采用基于形素的音译，还有 7%的语言对采用基于音素和形素的混合音译。由于基于音素的音译比基于形素的音译会带来更多的翻译误差，因此基于形素的音译是目前音译研究主要采用的方法。

2.3 音译模型[9]

音译模型是抽取音译知识、执行音译过程的核心，音译模型主要有基于规则、统计、机器学习、深度学习四大类。

（1）基于规则

基于规则的音译模型采用人工建立源语言与目标语言之间的音译规则。如 Davis 等[12]提出了面向对象的以规则为基础的文字输入系统的音译模型、Bhalla 等[13]基于规则实现了英语到旁遮普语的音译、Wan 等[14]人工编制的映射了英文音节到汉语拼音的转换规则，实现了英语到汉语的音译。基于规则的音译方法符合音译特点，但是人工建立的规则复杂，不能覆盖所有情况；建立的规则只能适用于某两种语言，对于其它语言对需要重新建立音译规则。

（2）基于统计模型[11]

基于统计的音译模型主要有噪声信道模型和信源信道模型两大类。

噪声信道模型（Noisy channel model）假设一段目标语言文本 T，经过某一噪声信道后变成源语言 S，也就是说，假设源语言文本 S 是由一段目标语言文本 T 经过某种奇怪的编码得到的，那么翻译的目标就是要将 S 还原成 T，这也就是就是一个解码的过程。该模型于 1993 年由 Peter E Brown 等人[14]用于法语到英语的音译，近年来被广泛应用于英语-汉语、英语-日语、英语-海地语以及西班牙语-英语、旁遮普语-英语、旁遮普语-印地语等的音译研究中。

信源信道模型（Source Channel Model，SCM）模型把音译过程看成是一个信息传输的过程，假设源语言文本 S 是由一段目标语言文本 T 经过某种编码得到的，那么音译的目标就是要将 S 还原成 T，这也就是一个解码过程。信源信道模型最早由 Kevin Knight 等人[15]于 1998 年用于英语到日语的音译研究中，Gao[16]等人采用该模型研究英语到汉语的音译。

Li 等人[17]于 2004 年在信源信道模型的基础上提出了联合信源信道模型（Joint Source Channel model），目前该模型被用于是英语-汉语的双向音译以及日语-日本汉字、英语-印地语等的音译，是目前常用的基于统计的音译模型。

（3）基于机器学习

基于机器学习的音译模型较多，常用的主要有：1）隐马尔科夫模型，该模型于 1999 年由 Jeong 等人应用于朝鲜语-英语的音译，该模型还被用于英语-朝鲜语/海地语的音译研究中；2）条件随机场模型，该模型于 2008 年由 Ganesh 等人用于英语-海地语的音译，Reddy 等人[18]和 Ying 等人[19]分别基于条件随机场研究了英语-海地语/泰米尔语、英语到汉语的音译；3）最大熵模型，Kato N 等人[20]应用该模型研究了英语-日语的音译模型，Oh 等人基于 MEM 研究了英语-朝鲜语/日语的音译；4）支持向量机，Rathod 等人[21]基于该方法研究了海地语/马拉地语-英语的音译；5）决策树，Kang 等人[22]基于该方法研究了英语-朝鲜语的的自动音译。

（4）深度学习

Kim [23]于 1999 年基于深度置信网络（Deep Belief Networks，简称 DBNs）研究阿拉伯文-英文的音译。Finch A 等人[24]将递归神经网络（Recurrent Neural Network，简称 RNN）应用于汉语-英语、英语-阿拉伯语等 14 种语言对的音译中，Lemao Liu 等[25] 提出 Additive Neural Networks 研究汉语-英语、日语-英语的音译。日本情报通信研究机构在 2015 年的 ACL 机器音译评测中采用神经网络对之前的音译系统进行了改进，在 14 个评测任务中的 11 个任务取得了第一的成绩[26]。

2.4 音译评测[27]

2009 年以来，ACL 会议共组织了 5 次机器音译（Machine transliteration）的评测（2009、2010、2011、2012、2015）。2015 年共有 6 个团队（瑞典乌普萨拉大学 UPPS、印度理工学院孟买 IITB、北京交通大学 BJTU、日本情报通信研究机构 NICT、国立台湾大学 NTU、加拿大阿尔伯塔大学 UALB 对 12 种语言、14 个语言对进行了音译评测，表 2 为 2015 年机器音译评测的基本情况，可见：

(1)英语-汉语之间的音译是该评测的热点，有 6 个团队参加了英语-汉语的音译评测、4 个团队参加了汉语-英语的音译评测；

(2)日本的情报通信研究机构参加了所有 14 个音译语言对的评测，并在 11 个语言对的音译评测中取得了第一的成绩成绩；

(3)音译 F 值超过 90%的有 5 个语言对，分别是：英语-波斯语（95%）、阿拉伯语-英语（94%）、英语-印地语（93%）、英语-泰米尔语（92%）、英语-孟加拉语/卡纳达语（90%）；音译 F 值在 80%～90%的有 2 个语言对，分别是：英语-日语（0.82）、英语-希伯来语（0.81）；音译 F 值在 70%～80%的有 5 个语言对，分别是：泰语-英语（0.78）、英语-泰语（0.76）、汉语-英语（0.73）、英语-朝鲜语（0.71）、日语到日语汉字（0.70）；英语-汉语的音译挑战较大，参加评测的 6 支团队均参加了英语-汉语的音译评测，最高 F 值为 67%。

表 2 2015 年 ACL 音译评测语言对及参评单位情况

语言对	参加测评的单位					
	UPPS	IITB	BJTU	NICT	NTU	UALB
英语-汉语	√	√	√	√*/0.67	√	√
汉语-英语	√	√	√*/0.73	√		
英语-朝鲜语				√*/0.71	√	
英语-日语				√*/0.82		√
日语-日语汉字				√*/0.70		
阿拉伯语-英语				√		√*/0.94
英语-印地语		√		√*/0.98		√
英语-泰米尔语		√		√*/0.92		√
英语-卡纳达语		√		√*/0.90		√
英语-孟加拉		√		√		√*/0.90
英语-希伯来语		√		√*/0.81		√
英语-泰语		√		√*/0.76		√
泰语-英语		√		√*/0.78		√
英语-波斯语		√		√*/0.95		√

3. 基于跨语言语料库的命名实体对齐

除了命名实体的直接翻译，基于跨语言语料库的命名实体对齐也是抽取跨语言命名实体翻译对的重要方法。根据是否对所有语言的语料进行命名实体识别，基于跨语言语料库的命名实体对齐有两种方法：基于对称策略的命名实体对齐和基于非对称策略的命名实体对齐。

3.1 基于对称策略的命名实体对齐

基于对称策略的命名实体对齐首先在各个语言上进行命名实体识别，然后根据跨语言命名实体的特征计算它们之间的相似度，进而实现跨语言命名实体的对齐。

基于对称策略的命名实体对齐的公式如下：

$$Score(NE_1, NE_2) = \sum_{i=1}^{n} \mu_i Feature_i\ (NE_1, NE_2)$$

其中，$Score(NE_1, NE_2)$是待对齐命名实体NE_1, NE_2的相似度分值；$Feature_i\ (NE_1, NE_2)$是待对齐命名实体NE_1, NE_2第i个特征的特征值，μ_i是第i个特征的权重。

目前常用于跨语言命名实体对齐的特征主要有三类：实体自身特征（包括音译特征、意译特征、实体类型特征、词长特征等）、实体上下文特征、实体位置特征。

由于平行语料的命名实体出现的位置具有很强的一致性，Kumano[28]基于平行语料、根据命名实体翻译对在文档中出现顺序的相似性实现了从文档对齐的双语平行语料库中提取英语-日语命名实体翻译对。Huang[29-30]使用音译特征、实体标注特征和意译特征、共现频率特征在平行语料上抽取汉语-英语实体的翻译等价对。Lu M 等人[31]等利用意译特征、翻译特征、词长特征、上下文特征计算各语言命名实体之间的相似度，进而在可比语料上实现双语命名实体的对齐。Fung[32]提出了词的分布假设，通过上下文的相似度来判断两个词语的翻译关系，进而基于上下文特征从可比语料中抽取词的双语翻译对。

3.2 基于非对称策略命名实体对齐

基于非对称策略的命名实体对齐首先在一种语言的语料上进行命名实体识别，然后根据多种策略在其他语言的语料库中获取命名实体。

Feng[33]首先自动识别出了英语命名实体，然后通过一些规则在汉语文本上建立汉语实体的候选项（没有对汉语进行词的切分和实体标注），最后利用最大熵模型进行特征融合计算，进而实现了英语-汉语命名实体的对齐。Moore[34]标注了英语命名实体，然后使用了一系列逐步递进的代价模型在法语中寻找对应的命名实体，这种方法在很大程度上依赖于语言信息，比如两种语言中都出现的字母序列或大写字母标志，对于不属于同一语系的两种语言是不适用的。杨萍等[35]首先对汉语端进行命名实体标注，然后基于双语 HMM 词对齐结果利用滑动窗口的方法抽取所有候选命名实体翻译对，最后基于融合 5 种特征的最大熵模型对所有候选翻译单位进行过滤，选取与汉语端命名实体相对应的置信度最高的新蒙古语命名实体翻译单位。

4 基于网络挖掘的跨语言命名实体对抽取

最早基于网络挖掘的双语命名实体对抽取开始于 2002 年 Al-Onaiza 的研究[7]。图 1 是通过意译、音译模型和互联网抽取阿拉伯文-英文命名实体翻译对的流程图，该方法首先构建阿拉伯语-英语可比语料库；并抽取阿拉伯语的命名实体；然后对于给定的阿拉伯语命名实体，应用音译模型、双语字典和英文新闻语料库，产生对应的英文翻译的排序列表，这里对人名使用了音译模型，对地名和机构名采用了音译和意译的组合模型；最后通过网络资源上的单语信息对翻译列表进行重新排列后，得到了比较好的翻译结果。

目前，基于网络挖掘的跨语言命名实体翻译对抽取方法主要有两类方法：基于网络搜索引擎、基于维基百科。

4.1 基于网络搜索引擎

Al-Onaizan Y 和 Knight 的英语-阿拉伯语命名实体抽取是最早的基于网络搜索引擎的跨语言命名实体抽取应用。Lam[36]将音译和意译融合在一个统一的框架下来抽取命名实体对，首先从指定的新闻网站上抓取新闻网页，然后将这些新闻网页基于

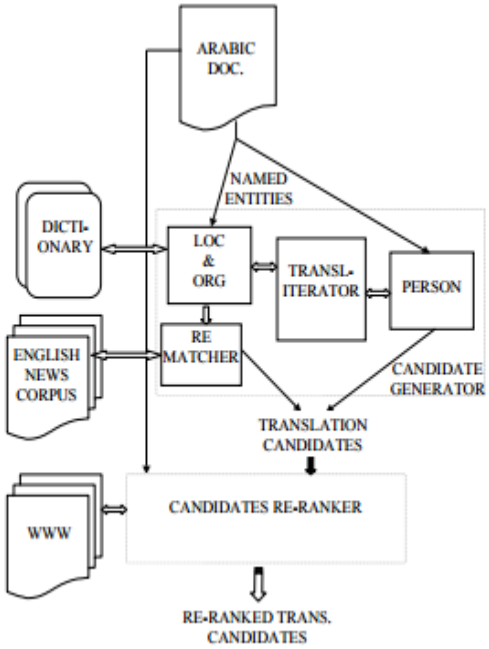

图 1 Al-Onaizan 等人的命名实体翻译系统图

主题进行聚类，在每个主题的文档中抽取出汉语-英语命名实体，最后通过混和音译匹配算法判定哪些双语命名实体对是互为翻译的。Huang[37]基于网络资源，采用语音相似性和语义相似性结合的方法翻译不常见的命名实体，给定一个源语言命名实体和它的上下文，首先根据上下文的翻译产生目标语言的查询词，在目标语言语料库中寻找相关的文档，然后将抽取到的目标语言文档中的命名实体和源语言命名实体进行语音和语义上的相似性比较，最后筛选出最佳目标语言翻译。在此基础上，Huang[38]又给出了一个从网络语料中挖掘关键词翻译的新框架。给定一个中文检索关键词，首先用搜索引擎检索出含有该关键词的中文网页片断；再从这些网页片断中提取与这个关键词主题相关的中文线索词，通过双语词典将这些中文线索词译为英文，作为扩展查询词；然后重新利用搜索引擎检索出含有这个中文关键词和英文扩展查询词的汉英混合网页；最后基于这个混合网页，根据音译，意译和频率等特征进行中文关键词的翻译。Zhang[39]通过对中英混合网页的考察发现，中文网页中如果出现英文词，往往是代表英文前面一个中文词的翻译，并常常被括号括起来。同样地，Cheng[40]发现如果中文出现在英文网页中，它的翻译一般就包含在相同的网页中。根据以上规律，Zhang[41]根据中文查询词的信息来扩展英文关键字，然后检索得到包含

这个查询词的中英混合文档（查询词的翻译往往包含在内），最后从混合文档中确定这个查询词的翻译，采用这种方式翻译OOV词是比较有效的。

4.2 基于维基百科的跨语言命名实体翻译对抽取

截止2016年5月，维基百科中有285种语言（包括10种暂停使用的）的词条，其为跨语言命名实体翻译对的抽取提供了丰富的资源。

目前基于维基百科的跨语言命名实体翻译对抽取主要涉及以下两方面：(1)基于维基百科构建跨语言语料库，基于语料库的跨语言命名实体抽取是跨语言命名实体翻译对抽取的重要方法，但是语料库的构建及标注是一项需要大量人力、财力、时间的工作，通过维基百科可以一定程度上快速构建部分标注的跨语言语料库；（2）基于维基百科的多语言链接直接抽取跨语言跨语言命名实体翻译对，维基百科的词条存在多语言链接，因此可以根据词条的多语言链接直接抽取跨语言命名实体翻译对。

Kim等人[42]提出了一种基于半监督学习和维基百科元数据识别多种语言命名实体的方法。Nothman等人[43]首先基于维基百科构建了一个大规模的已标注的语料库，然后基于命名实体类型对语料进行了分类，最后通过将目标文本的分类信息投射到锚文本上将文本的链接变换到命名实体标注，进而获取了多语言的的命名实体翻译对。

5 总结与展望

跨语言命名实体翻译对的抽取对于机器翻译、跨语言信息抽取、跨语言问答系统、跨语言信息检索等具有重要意义，本文从命名实体的音译、基于跨语言语料库的命名实体对齐、基于网络挖掘的跨语言命名实体抽取三个方面对跨语言命名实体翻译对的抽取研究进行了梳理，可以看出：

（1）基于深度学习的音译将是跨语言命名实体翻译对抽取今后的研究重点

无论是基于翻译的命名实体翻译对获取还是基于语料库的跨语言命名实体翻译对抽取，音译都是其中必不可少的一部分；联合信源信道模型以及基于CRF、HMM等监督式学习算法的音译模型是目前较为成熟的音译模型，而鉴于深度学习目前在音译领域取得的成绩，基于深度学习的音译将是今后音译的研究重点。

（2）基于语料库的跨语言命名实体翻译对抽取研究面临已标注语料库构建困难的问题

跨语言语料库的构建和标注是基于语料库的跨语言命名实体翻译对抽取的基础，除了个别资源较多的语言，对大部分语言、尤其是资源较少语言而言，获取大规模、高质量的跨语言语料库难度较大，并且标注语料的人力、经济、时间成本也一定程度上制约着基于跨语言语料库的命名实体翻译对抽取研究。

（3）基于网络挖掘的跨语言命名实体翻译对抽取将是跨语言命名实体翻译对抽取研究的趋势。

网络中丰富的资源为跨语言命名实体翻译对的抽取和消歧提供了基础，从首个跨语言命名实体翻译对抽取研究至今，基于信息检索的命名实体翻译对抽取一直是跨语言命名实体翻译对抽取研究的热点，除此以外，基于网络知识库——维基百科的跨语言命名实体翻译对抽取研究目前也成为跨语言命名实体翻译对抽取研究新的热点，因此，基于网络挖掘的跨语言命名实体翻译对抽取将是今后跨语言命名实体翻译对抽取研究的趋势。

参考文献

[1] Rau L F. Extracting company names from text[C]//Artificial Intelligence Applications, 1991. Proceedings Seventh IEEE Conference on. IEEE, 1991, 1: 29-32.

[2] 华却才让,姜文斌,赵海兴,刘群. 基于感知机模型藏文命名实体识别[J]. 计算机工程与应用, 2014,15:172-176.

[3] 加羊吉,李亚超,宗成庆,于洪志. 最大熵和条件随机场模型相融合的藏文人名识别[J]. 中文信息学报, 2014, 01:107-112.

[4] 木合塔尔·艾尔肯. 基于条件随机场的维吾尔语人名识别[D].新疆大学,2013.

[5] 米日姑·肉孜. 维吾尔文机构名识别研究[D]. 新疆大学,2013.

[6] 通拉嘎. 基于蒙古文语料库的人名自动识别[D]. 中央民族大学,2013.

[7] Al-Onaizan Y, Knight K. Translating named entities using monolingual and bilingual resources[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 400-408.

[8] 王欣欣. 统计机器翻译中命名实体处理研究[D]. 哈尔滨工业大学, 2009.

[9] Dhore M L, Dhore R M, Rathod P H. Survey on Machine Transliteration and Machine Learning Models[J]. 2015, 4(2):9-30.

[10] 阿力木·木拉提. 基于音节切分的维吾尔人名汉字音译研究与实现[D].新疆师范大学,2014.
[11] Karimi S, Scholer F, Turpin A. Machine transliteration survey [J]. ACM Computing Surveys, 2011, 43(3):194-218.
[12] Davis M E, Lin J. Object-oriented rule-based text input transliteration system: US, US 5432948 A [P]. 1995.
[13] Bhalla D, Joshi N, Mathur I. Rule Based Transliteration Scheme for English to Punjabi [J]. International Journal on Natural Language Computing, 2013, 2(2):67-73.
[14] Wan S, Verspoor C M. Automatic English-Chinese name transliteration for development of multilingual resources[C]// COLING 98 International Conference on Computational Linguistics. 2002: 1352-1356.
[15] Knight K, Graehl J. Machine transliteration[C]// Eighth Conference on European Chapter of the Association for Computational Linguistics. 1997:599-612.
[16] Gao W, Wong K F, Lam W. Improving Transliteration with Precise Alignment of Phoneme Chunks and Using Contextual Features[J]. Lecture Notes in Computer Science, 2004, 3411:106-117.
[17] Haizhou L, Min Z, Jian S. A joint source-channel model for machine transliteration[C]// Meeting on Association for Computational Linguistics. 2004:1190 - 1194.
[18] Reddy S, Waxmonsky S. Substring-based transliteration with conditional random fields[C]// Named Entities Workshop: Shared Task on Transliteration. 2009:92-95.
[19] Dhore M L, Dixit S K, Sonwalkar T D. Hindi to english machine transliteration of named entities using conditional random fields [J]. International Journal of Computer Applications, 2012, 48(23): 31-37.
[20] Goto I, Kato N, Uratani N, et al. Transliteration considering context information based on the maximum entropy method[J]. Proc of Ixth Mt, 2003:125--132.
[21] Rathod P H, Dhore M L, Dhore R M. HINDI AND MARATHI TO ENGLISH MACHINE TRANSLITERATION USING SVM [J]. Aircccse Org, 2013, 2(4):55-71.
[22] Kang B J, Choi K S. Automatic transliteration and back-transliteration by decision tree learning[C]// 2000.
[23] Kim J J, Lee J S, Choi K S. Pronunciation unit based automatic English-Korean transliteration model using neural network[C]//Proceedings of Korea Cognitive Science Association (in Korean). 1999.
[24] Finch A, Dixon P, Sumita E. Rescoring a phrase-based machine transliteration system with recurrent neural network language models[C]//Proceedings of the 4th Named Entity Workshop. Association for Computational Linguistics, 2012: 47-51.
[25] Liu, Watanabe T, Sumita E, et al. Additive Neural Networks for Statistical Machine Translation[C]// Meeting of the Association for Computational Linguistics. 2013:791-801.
[26] Finch A, Liu L, Wang X, et al. Neural Network Transduction Models in Transliteration Generation[C]// Named Entity Workshop.2015:61-66.
[27] Banchs R E, Zhang M, Duan X, et al. Report of NEWS 2015 Machine Transliteration Shared Task[C]//Proceedings of NEWS 2015 The Fifth Named Entities Workshop. 2015: 10.
[28] Kumano T, Kashioka H, Tanaka H, et al. Acquiring Bilingual Named Entity Translations from Content-Aligned Corpora[J]. Natural Language Processing – Ijcnlp, 2005, 3248:177-186.
[29] Huang F. Improved named entity translation and bilingual named entity extraction[C]// IEEE International Conference on Multimodal Interfaces, 2002. Proceedings. IEEE, 2002:253-258.
[30] Huang F, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[C]// ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition. Association for Computational Linguistics, 2003:9-16.
[31] Lu M, Zhao J. Multi-feature Based Chinese-English Named Entity Extraction from Comparable Corpora [J]. Proceeding PACLIC, 2006, 6: 134-141.
[32] Fung P, Cheung P. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and E.[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. 2004:57--63.
[33] D. H. Feng, Y. J. Lv and M. Zhou. A New Approach for English-Chinese Named Entity Alignment. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004: 372-379
[34] Moore R C. Learning Translations of Named-Entity Phrases from Parallel Corpora [J]. Proc of Eacl, 2003:259--266.
[35] 杨萍,侯宏旭,蒋玉鹏,申志鹏,杜健. 基于双语对齐的汉语－新蒙古文命名实体翻译[J]. 北京大学学报(自然科学版),2016,01:148-154.
[36] Lam W, Huang R, Cheung P S. Learning phonetic similarity for matching named entity translations and mining new translations[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004: 289-296.
[37] Huang F, Vogel S, Waibel A. Improving Named Entity Translation Combining Phonetic and Semantic Similarities[C]//HLT-NAACL. 2004: 281-288.
[38] Huang F, Zhang Y, Vogel S. Mining key phrase translations from web corpora[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 483-490.
[39] Zhang Y, Vines P. Using the web for automated translation extraction in cross-language information retrieval[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004: 162-169.

[40] Cheng P J, Teng J W, Chen R C, et al. Translating unknown queries with web corpora for cross-language information retrieval[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004: 146-153.
[41] Zhang Y, Vogel S, Waibel A. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system [J]. In Proceedings of Proceedings of Language Resources and Evaluation, 2004:2051--2054.
[42] Kim S, Toutanova K, Yu H. Multilingual named entity recognition using parallel data and metadata from Wikipedia [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 694-702.
[43] Nothman J, Ringland N, Radford W, et al. learning multilingual named entity recognition from Wikipedia [J]. Artificial Intelligence, 2013, 194: 151-175.