基于网络挖掘的英汉人名翻译[1]

刘颖，曹项
清华大学中文系，北京，100084
yingliu@tsinghua.edu.cn

摘要：本文利用搜索引擎从网络中挖掘英语人名的中文翻译。该方法综合利用翻译辅助词、规则、音译相似度和条件概率。首先，利用搜索引擎从互联网上搜索英文人名及其中文翻译候选。把汉语人名标注结果、翻译辅助词、中英文人名共现规则和发音音节长度结合起来提取翻译候选词。翻译辅助词有助于搜索与英文人名更相关的信息，中英文人名共现规则和发音音节长度进一步缩小英文人名翻译的范围，使得英文人名翻译的搜索符合人名共现规律和发音规律。然后，根据音译相似度和条件概率对候选词进行排序。人名翻译的绝大部分是根据发音翻译过来的，音译相似度就是判断两个词在发音上的相似性。条件概率从统计上判断两个词的共现信息。实验结果表明，翻译辅助词、规则、音译相似度和条件概率都有助于提高人名翻译的正确率。

关键词：人名翻译；音译相似度；规则；条件概率
中图分类号：TP391　　文献标识码：A

English-Chinese Name translation Based on web mining

Abstract: We propose a method to translate English name into Chinese name using the search engine. The method makes use of supporting word, co-occurrence rules of English and Chinese name, transliteration similarity and conditional probability. First, the translated candidates of English names are obtained by means of the search engine. We use the name tagging results, supporting words, co-occurrence rules of English-Chinese name and the length of syllable to obtain translated candidates from online corpus. Supporting words help to search more correlative names. Co-occurrence rules and the length of syllable make translations of an English name follow the regularities of co-occurrence and transliteration. Then the translated candidates are sorted according to transliteration similarity and the conditional probability. English names are almost translated according to their pronunciations and the transliteration similarity help to judge the similarity of their pronunciations. We use the conditional probability to obtain the co-occurrence information of two words statistically. The experimental results show supporting word, co-occurrence rules, transliteration similarity and conditional probability are all help to improve the precision of name translation.

Key words: name translation; transliteration similarity; rule; conditional probability

1 引言

随着互联网和搜索引擎的广泛使用，网络语料库规模日渐庞大。借助于先进的搜索引擎工具如Google、Baidu 等，我们可以更好地运用网络语料库，从中提取有价值的信息。许多学者利用网络语料库进行了命名实体翻译研究。

目前，对命名实体翻译的主流方法是统计方法，主要有基于音译的统计方法、基于双语语料库的统计方法和基于网络挖掘的方法。

利用基于音译的统计方法把一种语言的命名实体 A 翻译到另一种语言命名实体 B 的主要思路是：首先把命名实体 A 转换成 A 的发音，然后把 A 的发音转换成 B 的发音，再把 B 的发音转换成命名实体 B。代表性的工作有[1][2][3][4][5][6][7][8]。基于统计的音译方法准确率较低。

基于双语语料库的方法主要有基于平行语料库的统计方法和非平行的双语语料库方法。

基于平行语料库的统计方法主要从双语语料库对齐的命名实体中统计多个特征，然后综合利用这些特征对齐新的命名实体。一般统计的特征主要包括：音译相似度、共现频率、互信息、对齐概率和语义相似度等。基于双语语料库的统计研究工作有[9][10][11][12]。基于平行的双语语料库统计方法可获得高质量的命名实体翻译，但大规模的双语平行语料库比较缺乏。

[1]**基金项目：** 受国家自然科学基金“基于语用信息的交互行为与语言特征的建模研究”（61171114）和教育部自主科研项目“基于大规模语料库的社会语用信息网的构建”（20111081010）支持

基于可比较的双语语料库统计方法主要利用实体间、实体的上下文以及实体的关系来发现其他实体间的翻译。代表性的工作有[13][14][15][16][17]。可以容易地获取大规模的非平行双语语料库，但由于两个语料库的实体及实体关系不是严格的一对一关系，导致该种方法的实体翻译准确度不高。

基于搜索引擎翻译的基本处理步骤为：（1）输入查询词，获取含有相关内容的双语语料。（2）从提取出的双语语料中生成相应的翻译候选词。（3）排序候选词并挑选出合适的翻译结果。因此，如何能搜集到相关的双语语料和候选词、进行有效的排序并发现合适的翻译是进行基于搜索引擎的人名翻译的基础。Lu, W. H.从链接锚点文本（指链接和用来说明链接网页中包含内容的文本）中提取查询词的翻译。利用链接结构和网页空间来统计与查询词共现词的概率，以此来选取查询词的翻译词[18]。Zhang, Y.使用隐马尔科夫模型提取中文未登录词的英文翻译，通过搜索引擎检索中文未登录词，从检索到的结果中，根据长度和频度特征提取中文未登录词的英文翻译[19]。Wang J.H.针对只包含单语言文档的数字图书馆给出实现多语言搜索解决方案。对未登录词可实现跨语言检索。把对称条件概率和 n 元语法相结合来抽取翻译候选词，把瞄点文本、上下文向量和χ^2 检验进行线性结合来排序候选词[20]。Zhang Y. 结合了音译、意译和词频距离等来排序候选词。同时利用辅助搜索词来获取更相关的搜索结果[21]。蒋龙[22]根据音译特征搜索网络生成翻译候选词集，再用熵模型对其进行排序。郭稷融合了共现频率、候选翻译长度、命名实体判定、词性以及上文词等多个特征，从带有括号和英文的中文命名实体受限网页中自动抽取双语翻译对 [23]。赵明明利用 n 元模型实现的音译系统抽取命名实体单字，利用搜索引擎搜索包含命名实体单字的 N 元字串，再利用编辑距离和χ^2对候选翻译进行排序[24]。我们可以利用网络上超大规模资源的优势，翻译出词典中未收录的一些人名，并且可以发现人名的多种翻译结果。

本文实现的基于搜索引擎的英汉人名翻译方法充分结合了网络语料库、搜索引擎、翻译辅助词、中英文人名共现规则、音译、共现频率等多种知识。利用翻译辅助词使得搜索结果中包含更相关的双语语料。将中英文人名共现规则与音译翻译长度相结合来提高候选词提取的精度和效率。把基于最小编辑距离的音译和条件概率等特征相结合来对候选词进行综合排序，可把最相关的翻译结果排在前面。与他人工作相比，本文结合的知识源更多，不但利用了语言学知识（规则、人名长度、翻译辅助词和音译规律），而且利用了和人名翻译最为相关的统计知识（音译相似度和条件概率）。

2 基于网络搜索的英汉人名翻译的基准方法

基于网络的英汉人名翻译的基准方法是通过以下三个步骤来完成。

（1）获取网络语料库。向搜索引擎提交英文人名查询词，返回前 100 个搜索结果，去除文本中的 HTML 标记，只保留纯文本字符格式。若搜索结果不够 100 个，则保留所有结果。搜索引擎首先利用百度搜索引擎，若搜索结果不够 100 个，再利用 google 搜索引擎。

（2）根据人名翻译规律和翻译候选词长度生成中文翻译候选词集合。

人名翻译的基本规律主要为以下几种情况：中文人名紧邻英文人名，英文人名紧邻中文人名，中文人名与英文人名之间插入符号“（”、“-”、“/”等。我们只考虑这些情况的中英文人名互译。

利用音节方法，来预估中文候选词的最大长度和最小长度。把英文人名进行音节分解，如 Smith 史密斯”有 S，mi 和 th 三个音节，其所对应的中文名字的最大长度应为音节数目 3，最小长度应为元音的音节数目 1。

（3）排序中文翻译候选词并输出结果。利用公式（1）来对每一个候选词进行排序。

$$P(CN|EN) = f(CN \cap EN)/f(EN) \quad (1)$$

其中 CN 代表中文人名翻译候选词，EN 代表英文查询人名，$f(CN \cap EN)$表示在所有出现 EN 的结果中 CN 和 EN 共同出现的频率，$f(EN)$是 EN 出现的总次数。

3 基于网络搜索的英汉人名翻译

给定一个英文人名，下面给出如何应用搜索引擎自动生成相对应的中文名字。

3.1 基于网络搜索的人名翻译的基本过程

（1）向搜索引擎提交英文待翻译人名，收集前 100 个搜索结果，将结果去除 HTML 标记得到纯文

本，并用切分和词性标注软件 ICTCLAS[①]对其进行切分和标注。如果搜索结果少于 100 个，则保留所有结果。ICTCLAS 对于人名给出标注结果。

（2）从搜索结果中提取翻译辅助词，对辅助词进行排序并选前 3 个。

（3）提交待翻译人名和每个辅助词的组合进行网络搜索，每组搜索返回 100 个结果。若返回结果不够 100 个，则返回所有结果。预处理所有搜索结果。

（4）根据规则和预估计的翻译长度来提取中文人名翻译候选词集合。

（5）把音译相似度和条件概率相结合对翻译候选词排序。

（6）去除噪音，输出排在前面的候选词。

下面对上面的过程详细化。

3.2 获取辅助查询词

翻译辅助词就是与英文人名搜索词相关度较高并经常共现的词，在搜索过程中将英文人名和辅助词一并输入搜索引擎，返回的搜索结果将更相关，便于提取有价值的信息。如当搜索“Jennifer Lopez”的英文名字，由于其是美国歌星和影星，与其经常共现的词汇包括“明星”和“歌星”等，我们便将这些词作为辅助词，与英文名字一起输入搜索引擎。

获取辅助查询词的具体实现过程如下：

（1）首先将从搜索引擎搜索获取的前 100 个结果中，所有标记为名词且非停用词表中的词汇提取出来作为辅助词的候选词。停用词包括标点、连词、语气词、代词、副词、拟声词、时间词、地点名词等共 506 个。停用词表参照了哈尔滨工业大学信息检索研究室提供的停用词表。

（2）对于辅助候选词的排序，我们借鉴并改进了 Ricardo Baeza-Yates 提出的关联群簇方法，其用来计算辅助词与源英文人名查询词的关联度分值 $Score_{e,f}$ [25]。

$$Score_{e,s} = \frac{W_{e,s}}{W_{e,e} + W_{s,s} - W_{e,s}} \qquad (2)$$

其中，e 代表源英文人名查询词，s 代表辅助词的候选词，$W_{e,s}$计算方法如下：

$$W_{e,s} = \sum_{r_m \in R_n} f_{e,m} \times f_{s,m} \qquad (3)$$

r_m 代表第 m 个搜索结果，R_n 是所有搜索结果，$f_{e,m}$ 源英文人名查询词 e 在第 m 个结果中出现的频率，$f_{s,m}$ 是辅助词 s 在第 m 个结果中出现的频率。

（3）选取前 3 个中文词作为辅助词。

将获取的 3 个辅助词与源英文人名分别组合输入搜索引擎，每一组合提取前 100 个搜索结果，去除 HTML 标记，将其转换为纯文本。

3.3 用规则提取中文人名翻译候选词

（1）提取翻译规则

我们将从网络语料库、百科全书及线下语料库中提取的 1000 多个人名对作为提取规则的训练语料库，共提取了 120 多条规则，表 1 是出现最多的前 8 条规则，可以覆盖 90%的中英文人名共现情况，其中 CN 代表中文人名，EN 代表源英文人名查询词，x 代表一个汉字或英文单词。

表 1 主要规则形式

规则	概率	范例
CNEN	26.7%	比尔 盖茨 Bill Gates
CN(EN	23.2%	威廉 亨利 盖茨（William Henry Gates
ENCN	16.1%	Willis Carrier 威利斯 开利
EN(CN	10.4%	Bill Gates（比尔盖茨）
CN-EN	6.5%	比尔盖茨-Bill Gates
CN/EN	3.2%	比尔 盖茨/Bill Gates
CNxxEN	2.4%	史蒂夫 乔布斯传记 Steve Jobs
ENxCN	1.9%	Warren Buffett Speaks 巴菲特

（2）利用音节方法，来预估中文候选词的最大长度和最小长度。把英文人名进行音节分解，如 Jennifer 有 Je，nni 和 fer 三个音节，其所对应的中文名字的最大长度应为音节数目 3，最小长度应为元音的音节数目 3。Jennifer Lopez 的最大长度是 6，最小长度是 5。Bill Gates 的最大长度是 5，最小长度是 3。

（3）根据 ICTCLAS 切分和标注结果、预估的翻译长度和翻译规则生成中文翻译候选词集合。

如：“……做出集成电路（芯片），比尔盖茨（Bill Gates)做出视窗，……”

Bill Gates 翻译候选词的最大长度为 5，最小长度为 3。根据规则 CN（EN，可以提取出前后紧挨着 Bill Gates 的大于等于 3 小于等于 5 的汉字序列（遇到标点符号、英文数字等停止），生成候选词集合“比尔盖茨”、“做出视窗”、“比尔盖”“尔

[①] http:// ictclas.nlpir.org/

盖茨”“做出视”“出视窗”等。若所选的词串序列已被 ICTCLAS 标注为人名，则可直接选为该人名。

若候选词以总统、经理、总裁、歌星及影星等常见人名称呼开头，我们将其去除并生成新的候选词。

3.4 对中文人名翻译候选词进行排序

Fei Huang[10]、陈钰枫[12]指出人名翻译主要是音译形式。陈钰枫[12]对 LDC 机构发布的汉英双语命名实体语料库（LDC 2005T34）进行统计，发现人名翻译对音译词占 100%。所以，我们判断一个中文候选是不是给定英文人名的翻译，主要依靠两者之间的音译相似度和条件概率。

$$Score(CN,EN) = W_1 \times Score_{ED} + W_2 \times P(CN|EN) \quad (4)$$

$$W_1 + W_2 = 1$$

P(CN|EN)是在给定 EN 的情况下，检索出的页面中出现 CN 的概率，其计算见公式（1）。

$\text{Score}_{\text{ED}}(CN,EN)$是基于最小编辑距离（ED）的音译相似度[14]，见公式（5）。本文实验 W_1=0.7, w_2=0.3。

$$Score_{ED}(CN,EN) = 1 - \frac{ED(EN,PY_C)}{max(Num(EN),Num(PY_C))} \quad (5)$$

*EN*是源英文人名查询词，*CN* 代表中文人名翻译候选词，PYc是 *CN* 的拼音序列，$ED(EN,PYc)$是他们之间的最小编辑距离，即从*EN*到PYc的最小编辑操作数量，包括插入，删除及替换等。Num(x)代表中文拼音序列或英文人名 x 去除空格、点号和标点符号后字母的个数即此字符串的长度。比如中英文名字对比尔·盖茨 — Bill Gates, PY_c=（bi, er, gai, ci）和 EN=（Bill, Gates）的最小编辑距离 ED 是 5，最佳的编辑路径是“Bill”—“Bi er”,ED 为 2;“Gates”—“gai ci”ED 为 3。所以，比尔·盖茨与 Bill Gates 的音译相似度为 0.44。

3.5 去除噪音，输出翻译结果

在人名候选词生成阶段，可能会产生很多冗余信息。对于冗余信息需要进行降噪处理，如果翻译候选词 A 是翻译候选词 B 的子集，且翻译候选词 A 的排序值低于翻译候选词 B，我们便将翻译候选词 A 视为噪音并删除，如“比尔盖”是“比尔盖茨”的子集，并且“比尔盖”的排序值低于“比尔盖茨”的排序值，则将其视为噪音。

4 实验与结果分析

本文通过网络语料库、百科全书及线下语料库提取出了 1000 多个中英文人名翻译等价对作为训练语料库，从该训练语料库中提取中英文人名共现规则。对其中的 1/10 作为测试语料。

对人名翻译使用正确率来进行评价，正确率 P 是指已正确翻译的英文人名个数占翻译的所有英文人名的百分比。对于排序的中文人名翻译候选词，只要前 N 个结果中包含正确的翻译，则可算进 Top-N 结果的正确率中。 Top-N 的正确率记为 P_{Top}-N。

4.1 不同组合模块下的翻译效果评估与对比

为了对比利用辅助词、规则库、通过音译和统计特征排序以及噪音除噪的效果，我们分别与基准方法叠加组合计算出 P_{Top-N}，实验结果如表 2。基准方法是为了与添加辅助词、规则和统计排序等进行比较而进行的基本实验，过程如第 2 部分。

表 2　不同模块组合下的 Top-N 翻译正确率

不同方法	P_{Top-1}	P_{Top-3}	P_{Top-6}
基准方法	65.5%	69.7%	74.4%
基准方法+辅助词	68.6%	74.3%	78.7%
基准方法+辅助词+规则	74.8%	78.9%	82.1%
基于网络搜索的人名翻译	81.3%	83.5%	88.4%

从表 2 可以看出，对于排序最前的英汉人名翻译，基准方法的正确率为 65.5%。而采用辅助词后，正确率为 68.6%。再增加规则后，正确率为 74.8%。而采用基于网络搜索的人名翻译，正确率为 81.3%。基于网络搜索的人名翻译同时利用了翻译辅助词、人名翻译规则、基于最小编辑距离的音译相似度和条件概率排序。说明随着处理组合的不断增加，正确率逐步增加。基于规则库的候选词提取及根据音译和统计特征排序候选词都对翻译正确率的提高起到了重要作用。另一方面，从表格的横向来看，随着 Top-N 包含候选词的个数增加，正确率也逐渐增加。基准方法只使用了条件概率对翻译候选进行排序，而基于网络搜索的人名翻译把音译相似度和条件概率结合起来对翻译候选进行排序，翻译正确率进一步增加，说明音译相似度确实对人名翻译的判断确实有帮助。

从实验结果来看，如果待翻译的人名比较有名，从网络上就容易获取其人名翻译。如果待翻译的人名不是很有名，从网络上获取其翻译则比较困难。

4.2 主要错误类型分析

利用网络搜索进行人名翻译的主要错误有以下几类：

（1）从网络语料库中获取的人名翻译与标准不一致。这主要是因为一部分英文人名有多个译文，都是根据发音翻译过来的。比如：Emily 根据网络语料库的中文翻译是“艾米莉”，而用来计算准确率的翻译是“艾米丽”。

（2）英文人名搜索结果里中英文人名共现的信息或语料较少，从而导致无法提取含有正确翻译的候选词。

（3）语料库中出现与英文全名共现的部分中文翻译名的情况，如“巴拉克奥巴马—Barack Obama”在很多网络新闻报道中都是以下列形式出现“……当美国总统奥巴马（Barack Obama）的团队将要拍摄竞选视频时……”这类语料并未将巴拉克这个名字进行翻译，导致提取出“总统奥巴马”这类型的错误候选词。

（4）候选词中包含正确的翻译但排序模型未能将其排在前面。有些英中人名在网络语料库中出现次数很少或者不是根据音译规律来翻译的英中人名，导致对候选词排序时的排序评分比较低。

（5）考虑音译最大长度和最小长度提取翻译候选词，以便缩小候选词范围同时提高系统效率，但这种方法对于不是音译或意译的中英文对并未有效，如“滨崎步—Ayumi Hamasaki”，英文是通过日文发音翻译而成，而中文翻译却是从日文意译而来，因此会出错。

此外，还有切词错误、词性标注错误和人名识别错误等。

为进一步提高人名翻译准确率，需要进一步提高汉语切词、词性标注和人名识别的正确率。判断准确率时，把英中人名翻译的多种可能考虑进来。对于一小部分没有根据音译规律进行翻译的人名需建立人名翻译词典或根据更多的上下文来进行判断。而对于搜索结果较少或者根本没有搜索到的人名对，需利用其他资源来进行人名翻译。比如：利用其他双语对齐语料库或双语可比较语料库来进一步提高人名翻译准确率。

本文把词性标注、规则、上下文、音译和条件概率相结合，使得网络搜索可以根据 ICTCLAS 的人名标注、规则和预估的翻译长度来选择候选集合，这样可以使得搜索空间大大缩小。另一方面，根据音译相似度和条件概率从多个候选结果中选择出正确的翻译可以充分地利用人名翻译统计知识。

5 结论

本文提出的基于网络搜索的中英文人名翻译方法结合了规则、音译及统计等多种资源和特征。首先，为了获取到相关的网络语料和搜索结果，我们利用翻译辅助词和中英文人名共现规则。通过发音音节来预估翻译长度，从而提高了候选词提取和生成的精度。其次，我们结合了基于最小编辑距离的音译相似度和条件概率来对候选词进行综合排序。实验结果表明每一个特征的加入都有效地提高了人名翻译的正确率。

参考文献：

[1] Kevin Knight and and Jonathan Graehl. Machine transliteration[J]. Computational Linguistics. 1998, 24(4):599-612.

[2] Bonnie Glover Stalls and Kevin Knight. Translating names and technical terms in Arabic text[C]. Semitic '98 Proceedings of the Workshop on Computational Approaches to Semitic Languages.1998:34-41.

[3] Helen M.meng,Wai-Kit Lo, Berlin Chen and Karen Tang. Generating phonetic cognates to handle named entities in english-chinese cross-language spoken document retrieval[C]. Automatic Speech Recognition and Understanding. 2001:311-314.

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

[5] Yuqing Guo, Wang Haifeng. Chinese-to-English Backward Machine Transliteration[C]. International Joint Conferences on Artificial Intelligence on Nature Language Processing. 2004.

[6] Chun-Jen Lee, Jason S. Chang, Jyh-Shing Roger Jang. Extraction of transliteration pairs from parallel corpora using a statistical transliteration model[J]. Information Sciences. 2006, 176(1): 67-90

[7] Li Haizhou, Zhang Min, Su Jian. A Joint Source-Channel Model for Machine Transliteration[C]. Proceedings of the 42nd Annual

Meeting of the Association for Computational Linguistics. 2004: 21-26.

[8]Asif Ekbal, Sudip Kumar Naskar, Sivaji Bandyopadhyay. A Modified Joint Source-Channel Model for Transliteration[C]. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions .2006:191-198

[9] Fei Huang, Stephan Vogel, Alex Waibel. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[C]. Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition. 2003, 15:9-16.

[10] Fei Huang, Stephan Vogel, Alex Waibel. Improving Named Entity Translation Combining Phonetic and Semantic Similarities[C]. Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics. 2004: 281-288.

[11]陈怀兴，尹存燕，陈家骏.一种命名实体翻译等价对的抽取方法[J].中文信息学报,2008,22(4):55-60.

[12]陈钰枫，宗成庆，苏克毅.汉英双语命名实体识别与对齐的交互式方法[J].计算机学报，2011，34(9)：1688-1696.

[13] Jinhan Kim,Long Jiang, Seung-Won Hwang et al. mining entity translations from comparable corpora: a holistic graph mapping approach[C]. Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 1295-1304

[14]Jinhan Kim, Seung-won Hwang, Long Jiang, Young-In Song, Ming Zhou. Entity Translation Mining from Comparable Corpora: Combining Graph Mapping with Corpus Latent Features[J]. IEEE Trans. Knowl. Data Eng. 2012,25(8): 1787-1800.

[15]Taesung Lee and Seung-won Hwang. Bootstrapping Entity Translation on Weakly Comparable Corpora[C]. The 51st Annual Meeting of the Association for Computational Linguistic. 2013:4-9.

[16] You Gae-won, Hwang Seung-won, Song Young-in, Jiang Long. Nie Zaiqing. Efficient Entity Translation Mining-A Parallelized Graph Alignment Approach[J]. ACM Transactions on Information Systems. 2012, 30(4):1-23.

[17]张永臣，孙乐，李飞等.基于 Web 数据的特定领域双语词典抽取[J].中文信息学报，2006,20(2)：16-23.

[18]Lu W. H., Lee H. J., Chien L. F.. Anchor Text Mining for Translation Extraction of Query Terms[C]. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001:388-389.

[19]Zhang Y. and 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. 2004:162-169.

[20]Wang J. H., Teng J. W., Lu W. H., and Chien L. F., Exploiting the Web as the multilingual corpus for unknown query translation[J]. Journal of the American Society for Information Science and Technology.2006, 57(5): 660-670.

[21] Zhang Y., Huang F. & Vogel S. Mining translations of oov terms from the web through cross-lingual query expansion[C]. Proceedings of the 28th Annual International ACM SIGIR.2005.

[22]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

[23] 郭稷，吕雅娟，刘群. 一种有效的基于 Web 的双语翻译对获取方法[J].中文信息学报,2008,22(6)：103-109

[24] 赵明明，洪宇，姚建民，朱巧明. 基于音译和网络的命名实体翻译方法研究[C].第六届全国信息检索学术会议论文集. 2010，357-366

[25] Ricardo Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison- Wesley & ACM Press, Harlow, UK, 1999.