英语姓名译名的自动辨识

孙茂松
清华大学智能技术与系统
国家重点实验室

张维杰
烟台大学
电子与计算机应用系

摘 要

英语姓名译名的辨识对汉语自动分词研究具有一定意义。本文提出了一种在中文文本中自动辨识英语姓名译名的算法。算法虽然简单、直接，却非常有效。我们从新华通讯社新闻语料库中随机抽取了 1000 个包含英语译名的句子 (共 61017 个中文字符)作为测试样本。实验结果表明，召回率达到了 98%。

Translated English Name Identification in Chinese Texts

Sun Maosong
National Intelligence Technology and System Lab.,
Tsinghua Univ., Beijing, P.R.C.

Zhang Weijie
Dept. of Electronics and Computer Application,
Yantai Univ., Shandong, P.R.C.

ABSTRACT

The processing of translated English names is significant to the approach of Chinese word segmentation. This paper presents a simple,straightforward but effective algorithm for automatically identifying this sort of proper nouns in Chinese texts. The testing sample, involving 1000 sentences each of which contains at least one translated English names, is extracted at random from the Xinhua News Corpus.The preliminary experiment shows that the recall rate of this algorithm reaches 98%.

1. 引言

汉语自动分词系统在中文信息处理领域中占有重要位置。分词过程中，时常会遇到英语姓名译名，如果不予处理，将会影响分词的正确性。例如，输入句子:

埃及总统穆巴拉克访问叙利亚。 (例 1)

若无人名处理功能，例(1)被切成:“埃及 总统 穆 巴 拉 克 访问 叙利亚 。”

有时更对正常的分词形成干扰:

国际田联取消费尔南多.凯勒参加铁人三项赛的资格。 (例 2)

假如从右向左逆向扫描，则导致错误 (注意其中的“消费”):“国际 田联 取 消费 尔 南 多 . 凯 勒 参加 铁人 三 项 赛 的 资格 。”

可见英语姓名译名的辨识是分词中重要一环，不容忽视。关于这方面的研究迄今尚未有报道。

2. 算法设计

2.1. 英语姓名译名用字表

《英语姓名译名手册》[1]收录了英语姓氏、教名约四万个。根据该手册，经计算机统计得到“英语姓名译名用字表”，共包括汉字476个：

啊阿埃艾爱安昂奥巴白柏拜班邦包保堡鲍北贝倍本比彼毕庇辟壁陛边别滨宾玻波博勃伯卜布采蔡藏策查察柴昌彻陈楚垂茨慈次聪存措达大戴代丹当道德得登邓迪狄底地蒂第帝丁东杜敦顿多厄恩耳尔法凡范方菲费芬丰冯佛夫福弗辅富益甘冈高哥戈葛格各根贡古顾瓜圭郭果哈海罕翰汉杭豪赫黑亨洪侯胡华怀惠霍基吉季计嘉佳加贾简姜焦杰捷金津京久居喀卡开凯坎康考柯科可克肯孔扣寇库夸匡奎魁坤昆阔拉腊莱来赖兰朗劳勒乐雷蕾黎理李里礼莉丽历利立莲连廉良列烈琳林霖龄留刘流柳龙隆卢鲁露路吕律略伦萝罗洛玛马麦迈满曼芒茅梅门蒙孟米密敏明名摩莫墨默姆木穆拿娜纳乃奈南内嫩能妮尼年涅宁牛纽农努女诺欧帕派潘庞培佩彭蓬皮匹平泼朴普漆奇齐契恰钱强乔切钦琴青琼丘邱屈让热仁日荣茹儒瑞若撒萨塞赛三缮桑瑟森莎沙珊山尚绍舍申生盛圣施诗石什史士寿舒朔斯思丝松苏孙索所塔泰太坦汤唐陶特藤提惕田铁汀廷亭通透图托脱娃瓦万旺威韦为维伟魏卫温文翁沃乌吴武伍晤西锡希悉席霞夏显香向晓肖歇谢欣辛兴幸姓雄休修许雪逊雅亚延扬阳尧耀耶叶依伊易意因音英永尤雨约宰赞早泽曾扎詹湛章张哲者珍真芝知智治朱卓兹子宗祖佐丕谟葆薇岑弼娅缪珀瑙赍滕斐煦鸠窦艮麟黛

表1　英语姓名译名用字表

利用这张译名用字表，可初步确定译名在句中的大致位置及边界。设任一连续汉字串C1...Ci...Cn (n>1)，若对所有Ci (i=1,...,n)，均有Ci属于表1，则认为该汉字串为译名。此过程称之为“译名粗界定”。

马特乌斯的远射在墨西哥大赛上就已出名；克林茨曼既能插入对方罚球区抢点射门，也能带球从中场快速突破攻门；弗尔勒尔、利特巴尔斯基、布雷默也是个个绝技在身。　(例3)

“译名粗界定”可正确给出例(3)中全部译名：“马特乌斯”、“克林茨曼”、“弗尔勒尔”、“利特巴尔斯基”和“布雷默”。但也会出现不少错误。如例(4)、例(5)：

政府总理卢卡诺夫和社会党主席利洛夫等参加了庆祝活动。　(例4)

英国首相撒切尔夫人今天在下议院表示，...　(例5)

运行结果认为“理卢卡诺夫”“席利洛夫”“撒切尔夫”是译名。再如：

他对诺贝尔医学与生理学奖获得者休伯和韦塞尔教授的研究成果作出修正。　(例6)

任命雷乌本.利斯塔为海军新闻发布官。　(例7)

例(6)中“诺贝尔”“生理”“得者休伯”和“韦塞尔”及例(7)中“雷乌本.利斯塔为”被认作译名。

2.2. 称谓表及简单上下文

例(4)、例(5)这一类错误可通过增加一个称谓表予以解决。因为称谓经常与姓名联袂出现，故利用称谓可帮助确定译名的左右边界。以下给出一些称谓的例子：

博士　部长　参赞　大臣　大使　代办　顾问　会长　记者　教授　领袖　律师　明星　上校　将军　首相　书记　公主　先生　刑警　学者　医师　议长　元老　元首　院长　州长　主任　主席　专员　总编　总裁　总监　总理　总统　伯爵　董事长　司令官　参谋长　观察员　众议员　秘书长...

将称谓表嵌入辨识过程的方法非常简单：仅需令“译名粗界定”作用于输入句子中除称谓以外的汉字序列上。

经过这样的处理，例(4)、例(5)得到了正确结果："卢卡诺夫""利洛夫"和"撒切尔"。

实验表明，这一策略对译名辨识极其有效。

此外，某些简单的上下文亦可用于进一步裁决译名的边界，如标点、数字、空格、西文字母以及译者连接符"·"等。值得注意的是，有些动词常紧接在姓名的后面出现，可作为译名的右边界标志。单字动词如：

率 获 说 是 抵 离 谈 偕 称 ...

双字动词如：

报道 率领 会见 反对 强调 表示 接受 指出 主张 认为 发现 主持
介绍 呼吁 出席 电贺 参加 重申 要求 致电 看望 就任 代表 宣誓
批准 逝世 访问 辞去 提出 欢迎 预测 宣布 领导 援引 签署 担任
接到 表现 夺得 会晤 接任 同意 陪同 邀请 出任 前往 拒绝 应邀

对比例(8a)和例(8b)：

德索托她帮助办一下此事。 (例 8a)

德索托主张通过谈判解决双方的冲突。 (例 8b)

"译名粗界定"将认定"德索托"为译名。对例(8a)，导致错误。由于"托"除了可用作译名用字外，还可作为动词独立出现在句子中，故实际上"德索托"存在两个解，"德索托"(译名)和"德索"(译名)+"托"(动词)。例(8b)中，通过动词"主张"排除了后一种可能，使右边界明确下来。

对这类动词，要求其首字不能属于译名用字表。

2.3. 首尾逼近法

对文献[1]中约四万个译名进行计算，可分别得到：只能出现在译名首的字，不能出现在译名首的字，只能出现在译名尾的字，以及不能出现在译名尾的字(见表 2,表 3，表 4 及表 5)。

啊包玻昌陈聪狄帝葛顾郭杭侯计简江井寇魁李刘露麦孟墨牛培朴漆邱屈三圣朔孙所田魏悉许雪阳尧耀有于雨湛张哲珍窦

表 2 字表 HZ-Only-Head(只能出现在译名首的字)

柏堡北倍庇辟壁陛边别滨卜采藏策察垂茨慈次存措大得底地第东敦顿耳尔凡夫辅各果罕翰季佳姜捷京居喀可扣坤阔来乐蕾黎礼历立蓬连廉良烈霖龄留流律略萝满茅密敏名木拿娜乃嫩能年涅女蓬匹泼强琴青仁日荣茹儒若缮珊山上生盛诗石士寿丝太藤提惕汀亭透脱娃为伟卫文吴晤霞显向晓歇兴幸姓雄修逊叶依易意音永宰早者真芝知智治子丕谟薇岑弼娅赘煦鸠艮麟

表 3 字表 HZ-Not-Head(不能出现在译名首的字)

堡庇壁边采藏察垂慈存底敦凡果姜捷京坤来黎礼蓬良霖龄满仁日荣茹上生诗石藤亭透脱娃霞向雄逊意音永早者真芝知智丕弼娅煦鸠艮麟

表 4 字表 HZ-Noly-Tail(只能出现在译名尾的字)

啊爱柏包保北倍彼辟陛别滨班蔡柴昌陈楚次聪措代邓狄地帝辅葛各顾瓜郭杭侯计佳江焦井久居喀开柯扣寇匡魁阔蕾历立连留刘流露略律萝麦茅孟密敏名墨木拿乃能年牛女派培匹泼朴漆钱强青邱屈儒三缮史寿朔孙所陶提惕田铁为伟魏吴武伍晤悉席显晓兴幸姓许雪阳尧耀依有于雨湛张哲珍朱卓子祖谟蔡萧斐窦黛

表 5 字表 HZ-Not-Tail(不能出现在译名尾的字)

借助以上四张表，采用"首尾逼近"策略，可以解决例(6)中"生理""得者休伯"及例(7)中

"雷乌本.利斯塔为"这一类错误。所谓"首尾逼近"，描述如下：

步骤 0　设"译名粗界定"给出一候选译名为 Candidate = C1 C2 ... Ci ... Cn-1 Cn
　　t <-- 1;
步骤 1　若 Ct 属于 HZ-Noly-Head 则 Candidate 的左界定，转步骤 3;
步骤 2　若 Ct 属于 HZ-Not-Head
　　则 Candidate <-- Candidate - Ct; t <-- t + 1; 转步骤 1;
步骤 3　t <-- n;
步骤 4　若 Ct 属于 HZ-Noly-Tail　则 Candidate 的右界定，转步骤 6;
步骤 5　若 Ct 属于 HZ-Not-Tail
　　则 Candidate <-- Candidate - Ct; t <-- t - 1; 转步骤 4;
步骤 6　若 Candidate 的长度 < 2 则 Candidate 非译名
　　否则 Candidate 为译名，输出之.

例(6)中，"生理"中的"生"属于 HZ-Not-Head，被步骤 2 滤掉，又因只剩下"理"，长度仅 1，不能成其为译名。"得者休伯"中的"得""者"均属 HZ-Not-Head，经步骤 2 处理后，输出"休伯"为译名。例(7)中，"雷乌本.利斯塔为"中的"为"属于 HZ-Not-Tail，被步骤 5 滤掉，从而得到正确解"雷乌本.利斯塔"。

2.4. 联想表

尽管依据文献[1]得到的表 1 已比较完全，但由于人们翻译译名并非都严格按照某种规范选用适宜的汉字，故表 1 仍未能将译名用字囊括无遗。例如"其""汗""茂"等字均未收入该表中，这会导致在处理译名"伊沙克.汗""奥其尔巴特""马哈茂德"时发生错误。为解决这一问题，我们建立了一张联想表，将这些一般来说不经常出现的译名用字映射到表 1 中同音同调或者同音异调的汉字上，通过联想表联系起来的字对从译名处理的角度来看具有同样的性质。例如：

其--> 奇　　汗 --> 汉　　茂--> 茅

对"伊沙克.汗""奥其尔巴特""马哈茂德"，依靠联想表可将它们分别视作"伊沙克.汉""奥奇尔巴特"和"马哈茅德"予以处理，提高了系统的适应能力，同时避免了把这些字简单地追加到表 1 中所引起的对表 1 完整性及可靠性的破坏。

3. 实验系统与实验结果

系统总体结构见图 1。

实验系统在 486 微机上用 Borland C 实现。为了验证算法的有效性，我们从新华通讯社的新闻语料库中随机抽取了 1000 个包含英语译名的句子作为测试样本。这 1000 个句子共含 61017 个中文字符，1537 个英语译名。系统运行后，辨识出"译名"2376 个，其中 1507 个真正为译名。设：

召回率 = 文本中的译名被辨识出的比例;
精确率 = 辨识为译名者真正为译名的比例

则本系统的召回率和精确率分别为：

召回率 = 1507 / 1537 = 98%
精确率 = 1507 / 2376 = 63%

以下给出若干运行实例：

【输入】
　　赫德和美国副总统奎尔昨天在伦敦会晤时一致表示，坚持不向劫持者妥协。　　(例 9)
【输出】
　　赫德　奎尔

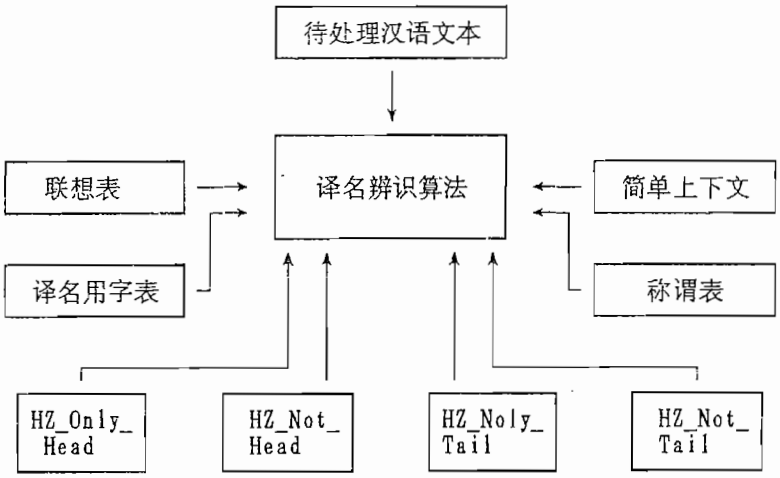

图 1　系统总体结构

【输入】

当大会主持人逐一介绍美国队员维尔·斯蒂格，法国队员让·路易斯·艾地安，苏联队员维克多·巴雅夫斯基，英国队员杰夫·萨莫斯，日本队员舟津圭三，中国队员秦大河时，场内一次又一次响起热烈的掌声。　(例 10)

【输出】

维尔·斯蒂格　让·路易斯·艾地安　维克多·巴雅夫斯基
杰夫·萨莫斯

【输入】

在这些被捕的人中包括“麦德林卡特尔”负责安全的头目、埃斯科瓦尔的内弟埃尔南·达里奥等　(例 11)

【输出】

麦德林卡特尔　埃斯科瓦尔　埃尔南·达里奥

【输入】

随同穆巴拉克总统来访的有副总理兼外交部长阿斯马特·阿卜杜勒·马吉德，副总理兼计划部长卡马勒·艾哈迈德·甘祖里，总统府办公厅主任扎克里亚·侯赛因·阿兹米和总统政治事务办公室主任兼外交部第一国务秘书乌萨马·巴兹等。　(例 12)

【输出】

穆巴拉克　阿斯马特·阿卜杜勒·马吉德　卡马勒·艾哈迈德·甘祖里
扎克里亚·侯赛因·阿兹米　乌萨马·巴兹

【输入】

人们向保加利亚共产党的创始人布拉戈耶夫·季米特洛夫和基尔科夫的故居

博物馆献了花圈和花束 (例 13)

〖输出〗

保加利亚　布拉戈耶夫·季米特洛夫　基尔科夫

〖输入〗

不过，半决赛英格兰队莱因克射入西德队的那个球却增添了比拉尔多的几分信心，而给贝肯鲍尔增加了几丝担忧。 (例 14)

〖输出〗

赛英格兰　莱因克　西德　比拉尔多　贝肯鲍尔

〖输入〗

比赛开始仅七分钟，中国奥林匹克队守门员和后卫处理门前冲吊球失当，被9号纽厄尔门前从容射入一球；下半场开赛后25分钟，范志毅受伤离场，埃弗顿队抓住中国奥林匹克队阵脚未稳之机，由10号科蒂再下一城 (例 15)

〖输出〗

奥林匹克　纽厄尔门　埃弗顿　奥林匹克　科蒂

〖输入〗

联邦德国代表团团长蒂特迈尔说：... (例 16)

〖输出〗

邦德　蒂特迈尔

在召回率和精确率之间，显然前者比后者更为重要。首先追求比较高的召回率，是我们设计本算法的基本出发点。实验结果表明，算法在这方面的表现令人满意，输入文本中仅有2%左右的英语译名成为"漏网之鱼"，且对不同国别的译名具有较强的适应性，同时对中、日姓名亦可起到过滤作用。精确率虽然相对较低，但对实验结果的分析显示，要提高这个指标并不困难。绝大多数辨识错误源自：或者将某些外国国名、地名误判为译名，如例(13)、例(14)中的"保加利亚""西德"，或者从两个以上邻接汉语词中各取一部分，拼成译名，如例(14)、例(16)中的"(半决)赛英格兰"、"(联)邦德(国)"。如果把译名的辨识过程与自动分词结合在一起(实际上也应如此)，则上述问题就会"迎刃而解"。就测试的1000句而言，如果扣除这类轻易可以解决的辨识错误，精确率即能提升至90%以上。

当然，有些辨识错误的纠正是较困难的。如例(15)中"(被9号)纽厄尔门(前从容射入一球)"，要将"纽厄尔门"中的"门"滤掉，上面讲过的几种手段都束手无策。进一步改进的办法是引入译名的音节信息，发现译名内在结构上的联系，藉以提高区别译名边界的能力。但我们的经验显示，这样做也有负作用，会把一些译名排除掉。此类问题，我们正在继续研究。

参考文献

〔1〕 新华通讯社译名资料组，《英语姓名译名手册》，商务印书馆，北京，1989年

〔2〕 张俊盛，陈舜德，郑荣，刘显仲，柯淑津，"多语料库作法之中文姓名辨识"，《中文信息学报》，第6卷，第3期，1992年