英汉人名音译方法研究

邹波，赵军

（中国科学院自动化研究所模式识别国家重点实验室，北京 100080）

摘 要： 本文详细比较了两种机器学习方法和两种统计翻译模型在英汉人名音译上的应用效果。本文首先提出了将英汉人名翻译问题看作序列标注问题的处理思路，并且将两种机器学习方法：最大熵模型和条件随机场模型应用于人名翻译。在使用机器学习算法进行音译时，本文比较了不同特征集对人名音译的影响，实验表明字母串特征和标注之间转移特征对提高音译性能有很大帮助。同时，本文还比较了使用不同的语言模型时，基于短语的机器翻译模型和基于 N-gram 的机器翻译模型在人名音译上的表现。实验表明，好的语言模型能极大地提高音译性能。当使用的训练集相同时，机器学习方法和统计翻译模型取得的音译效果差不多，但是统计翻译方法框架比较灵活，能利用外来的信息，因此更加适合进行英汉人名音译。

关键词： 英汉人名音译；序列标注；统计翻译模型

Comparison of Several English-Chinese Name Transliteration Methods

Bo Zou, Jun ZHAO

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080)

Abstract: This paper applies to the problem of English-Chinese Name Transliteration both the machine learning and the machine translation approaches. Viewing the transliteration problem as a sequence-tagging problem we tested two classic machine learning method: Maximum Entropy model and Conditional Random Fields, and found the latter performed better. Experimental results showed that performance depends on powerful features and transitional information between tags (language model), which is poorly integrated into the sequence tagging approaches. For better LM usage we also tested two machine translation approaches: the phrase-based model and the N-gram model. The influence of the size of the language models on the transliteration performance is also studied, and we draw the conclusion that in translating a English name to Chinese, the target side characters have a strong transitional relationship, which necessitates a better LM. With the same training data, machine learning methods achieved comparable results to that from machine translation models, but machine translation models has advantage in extensibility.

key words: English-Chinese Name Transliteration ;sequence-tagging ; Statistical Translation Model ;

1 引言

人名翻译接收一个源语言表示的人名作为输入，然后输出该人名以目标语言表示的翻译，例如，“Clinton”(英语)->“克林顿”(中文)。在人名翻译过程中，在保持源语言和目

本文受国家 863 计划项目（2006AA01Z144）、国家自然科学基金项目（60673042）的资助

标语言发音基本不变的原则下，调整源语言人名使之符合目标语言的语言习惯。人名自动翻译是很多跨语言应用中一个很重要的组成部分。在机器翻译、跨语言语信息检索系统以及跨语言问答系统中，往往会遇到很多未登录词，这些未登录词的很大的一部分是不同国家的人名，如果有个单独的模块来翻译这些未登录人名，那么将有效提高上述系统的性能。

在以前的研究中，研究人员提出了很多人名音译的方法。Knight 和 Graehl（1997）把日英人名音译分解成好几个步骤，每一步都由有限状态自动机来表示，他的方法是基于发音相似性的方法。Al-Onaizan 和 Knight（1998）提出不经过发音这个中间过程，直接以字素为基本翻译单元，用状态转换机完成翻译。Vigra 和 Khudanpur（2003）将基于词的翻译方法应用到英汉人名音译上。Li haizhou（2004）提出一种联合信源信道模型用于英汉人名音译，这种模型在翻译时同时考虑源语言和目标语言人名的上下文。他的基本翻译单元是双语短语对，他通过 EM 算法来获得双语人名短语对。Sherif (2007)等人将基于字母的自动机模型向前推进了一步，提出了基于子串的自动机模型。本文主要针对如何更好地将英语人名翻译成汉语人名的问题进行研究。本文的音译方法都是基于字素的，即直接把英文字母转换成汉字，而没有经过其他中间过程。本文首先提出英汉人名音译问题可以转化为序列标注问题的处理思路，并将最大熵模型和条件随机场模型应用于英汉人名音译。除了从序列标注问题的视角来看待人名音译问题外，本文还将人名音译问题看作是翻译问题，将基于短语的和基于双语 N-gram 的两种统计翻译方法用于音译，最后本文比较了这几种音译方法。

本文按如下方式进行组织：第二节简要介绍一下训练语料的预处理过程，后面的各种方法都是在预处理后语料的基础上进一步加工使之适合相应的翻译方法；第三节将最大熵和条件随机场方法用于音译；第四节将基于短语的和基于双语 N-gram 的两种机器翻译方法用于音译；第五节介绍和分析了实验设计以及实验的结果；第六节中对本文以及已有工作的问题进行了总结，同时展望了下一步的工作的方向。

2 语料预处理

在得到英汉双语人名对训练语料后，如果想将训练语料中的信息用于翻译新的英语人名，需要将训练语料中的双语人名对在更小的颗粒度上进行对齐。本文的音译方法都是基于字素的，即直接把英文字母转换成汉字，所以本文将英文人名的字母与汉语人名的汉字进行对齐。

本文首先将双语人名对中的英语人名表示成英文字母序列，汉语人名表示成汉字序列，然后用 GIZA++将重新表示后的训练语料进行对齐。由于 GIZA++是基于词对齐的，对齐时只考虑一对一的对齐。而音译中有一些多对多的翻译情况，大部分都是由字母 x 后面跟了元音引起的，比如'xa'翻译成'克萨'。为了较好的处理这种情况，在音译中本文采用了 GIZA++中汉语到英文对齐的结果，这样多个汉字能对应到单个字母上。（克萨就会同时对到字母 x 上，后面的元音会对空）然后把那些对空的英文字母依附到前一个有相应汉字对应的字母上组成英文字母串，除了像上面的'xa'这种前面的情况外，通常都是一个英文字母串对应一个汉字。后面的几种音译方法使用的语料都是在 GIZA++预处理产生的语料的基础上，通过进一步加工得到的。

例子：以英文名 yerxa 为例，其中文名为耶克萨，两者通过 GIZA++对齐后的结果为：

y e r x a
耶 克萨

3 基于机器学习的音译模型

本文首先从序列标注的视角来处理英汉人名音译问题。序列标注问题要求输出的标注序列长度和输入序列的长度相等，而音译问题中英语人名字母长度通常与其对应的汉语人名汉字的长度不相等，因此要对训练语料进行处理以满足要求。由于英文人名的字母长度通常都要比它对应的汉语人名长度长，因此本文以英语人名字母的长度作为该序列的长度。我们通过 GIZA++对齐能得到每个汉字对应于哪一个英文字母，然后把该英文字母对应的汉字作为该字母的标注结果，如果一个英语字母对应多个汉字，那么这几个汉字组成的字串作为该字母的标注结果，那些没有相应汉字对应的英文字母我们统一用一个无意义的标注符号‘-’表示（在实际上，这些字母并不是对空，它们实际上与前一个有汉字对应的英文字母组成字母串一起对应哪个汉字或者汉字串）。将训练语料中的英汉人名对用上面的方法进行表示后，就可以用机器学习方法训练出模型，然后对新来的英语人名进行标注（翻译）。本文使用了最大熵模型和条件随机场模型来进行英汉人名音译。

4 基于统计翻译的音译模型

除了从序列标注的视角处理英汉人名音译问题外，本文还从统计翻译的视角处理英汉人名音译问题。在用 GIZA++得到汉字到英文字母对齐的结果后，我们不再将那些对空的英文字母对应到伪标注‘-’上，而是将这些对空的字母与它前面最近的有汉字对应的字母组成字母串，一起对应到汉字上，然后用统计翻译方法来进行音译。

本文将基于短语的和基于双语 N-gram 的两种统计翻译模型应用于英汉人名音译。基于短语的翻译模型是对 $P(T|S)$ 进行建模，最后选择是是使翻译概率 $P(T|S)$ 最大的的翻译。而基于双语 N-gram 的翻译模型是对 $P(T,S)$ 进行建模，最后选择是是使翻译概率 $P(T,S)$ 最大的的翻译。

基于短语的翻译模型表达式如下：

$$\hat{T} = \arg\max_{T} P(T)P(S|T) \tag{4.1}$$

基于双语 N-gram 的翻译模型的表达式如下：

$$P(T,S) \approx \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, (t,s)_{k-2}, ..., (t,s)_{k-n+1}) \tag{4.2}$$

基于短语的翻译模型在翻译各个源短语片段时，不需要考虑它前面的源语言短语片段是什么，各个源语言短语是独立翻译的，各个源语言短语的目标语言短语翻译之间仅仅靠目标语言的语言模型进行联系。而基于双语 N-gram 的模型将源语言短语和候选的目标语言短语作为一个整体进行翻译，研究人员通常把这种双语短语对称之为双语元组。在翻译时，选择哪一个双语元组作为最后的翻译不仅要考虑该双语元组在训练语料中出现的概率情况，而且还要考虑该双语元组与在它前面的双语元组的同现概率。

5 实验结果及分析

（1）实验准备

本文将 15730 个英汉双语人名对（其中含中科院计算语言学课程上的 11406 对英汉双语人名对和 LDC 语料上有英美人名标记的 4324 对双语人名对）为语料集，采用上述几种方法进行英汉人名音译。本文将其中 14436 对作为训练语料，1294 对作为测试语料。

在进行完预处理后，本文的双语训练语料中英语人名的平均长度为 6.74，与其对应的汉语人名的平均汉字个数为 3.14 ，测试语料英文人名的平均长度为 7.64，与其对应的汉语人名的平均汉字长度为 3.75 个，在训练语料中出现汉字个数 421 个，测试语料中出现的汉字的个数为 284 个，在测试语料中有 31 个汉字没有在训练语料中出现。在测试语料中，有 46 个样例的中文人名包含没有在训练语料出现的汉字。

在利用统计翻译方法进行音译时，本文不仅使用双语训练语料中的 11406 个汉语人名来训练语言模型，本文还使用了更大的汉语人名语料集来训练语言模型。我们将 LDC 上的英汉人名对中（LDC2005T34）的日韩人名去掉后，得到了 565370 对英汉人名语料，我们利用其中的汉语人名来训练语言模型。

（2）评测方法

本文的实验以系统给出的最优结果的正确个数来评价系统的性能。除了给出各个系统的最优结果的正确个数外，本文还考察了系统输出的前 500 个结果。输入一个测试英文人名，得到系统输出的概率最大的前 500 个结果后，观察这 500 个结果是否含有正确答案，以及正确答案在这 500 个结果中首先出现的位置。本文统计了正确答案的在结果集的 top1、2-5、6-10、11-50、51-100、101-500 这六个区间段内个数以及正确答案在前 500 结果中出现的总数。

（3）基于机器学习的音译实验结果

本文将最大熵模型和条件随机场模型应用于英汉人名音译。在本文的训练语料中，有 425 个翻译标注（421 个汉字和 4 个双汉字）。本文对不同的特征集合进行了实验。下面以一个具体的例子来说明实验中采用的几种特征。对于训练语料中的人名对（abbadessa ，阿巴--德-萨--），在标注其中英文人名中的'd'时，采用了以下四种特征集。

1. 只使用周围的字母特征，这时抽取的特征(用观察，标注对的形式出现)为：<b，德>，<b，德>，<a，德>，<d，德>，<e，德>，<s，德>，<s，德>，假设待标注序列为 x，在标注位置 i 时，使用的特征可以表示成：$\phi(x_i, y_i)$，$\phi(x_{i-1}, y_i)$，$\phi(x_{i-2}, y_i)$，$\phi(x_{i-3}, y_i)$，$\phi(x_{i+1}, y_i)$，$\phi(x_{i+2}, y_i)$，$\phi(x_{i+3}, y_i)$。
2. 在 1 的基础上加入标签之间的联系。在标注 'd' 时，还加入<- ，德>即加入 $\phi(y_{i-1}, y_i)$。
3. 除了加入 1 中的字母特征外，还加入字母串的特征。例如：<bbad，德>，<bad，德>，<ad，德>，<de，德>，<des，德>，<dess，德>这些特征，用数学符可以表

示成：$\phi(x_{i-N},\ldots,x_i,y_i)$，$\phi(x_i,\ldots,x_{i+N},y_i)$（N 可以取 1，2，3）。

4. 在 3 的基础上加入标签之间的联系。即在特征集 3 的基础上，还加入<-，德>，即加入 $\phi(y_{i-1},y_i)$。

当使用特征集 1 时，训练语料上的特征总数为 21,120 个，使用特征集 2 时，特征总数为 57,573 个，使用特征集 3 时，特征总数为 192,968 个，使用特征集 4 时，特征总数为 229,421 个。下面是最大熵模型和条件随机场模型在各个特征集上的最优结果。

表 1 最大熵模型在各个特征集上的音译结果

	Top1	2-5	6-10	11-50	51-100	101-500	总数
特征 1	180	190	87	190	54	152	853
特征 2	228	202	73	170	65	151	889
特征 3	292	199	90	153	73	146	953
特征 4	322	205	69	168	69	132	965

表 2 条件随机场模型在各个特征集上的音译结果

	Top1	2-5	6-10	11-50	51-100	101-500	总数
特征 1	189	199	87	174	64	137	850
特征 2	243	185	71	164	67	117	847
特征 3	311	176	89	155	70	120	921
特征 4	348	201	67	153	67	107	943

通过对比机器学习方法在各个特征集上的效果，我们发现下面两种特征在英汉人名音译过程中是非常重要的。

1. 字母串特征。当人们进行音译时，总是根据上下文把一串字母翻译成对应的汉字，所以使用字母串特征时能为翻译提供更准确的信息。因此加入字母串特征后，系统的性能要比不加字母串的性能要好不少。

2. 标注之间转移的特征。为了使用机器学习方法，虽然人为地加入了很多无意义的标记‘-’，但是标记间的状态转移信息依然非常有用，当考虑标注间的转移关系时，系统的效果比不考虑转移关系的效果要好很多，不仅最优结果正确率要多，系统的前 500 个结果包含正确个数也要多。

在本文的实验结果中，基于条件随机场方法的音译结果在各个特征集上都比基于最大熵方法的效果要略好一些。但是总的来说，基于条件随机场的方法在各个特征集上的表现与最大熵方法很相似，这是因为本文采用的一阶线性链条件随机场模型与最大熵模型比较类似，采用的特征也完全一样，两种方法仅仅是归一化不一样，而音译的偏置问题不是很严重，因此效果也差不多。

通过对比各种特征集上的结果可以看出，机器学习方法的优势在于能在特征集中选择出有用的特征，但是并不是完全随便的给定一个特征集就想取得好的效果，对于不同的问题，不能套用一个特征模板。要针对不同问题选择不同的特征集，扩大特征集的选择范围，

只有把有代表性的特征放入候选特征集，机器学习方法才能取得比较好的效果。

在使用上面两种机器学习方法进行英汉人名音译时，如果加入标记间的转移概率，系统的效果比不加入标记间转移的效果要好不少。这说明了在英文人名翻译成汉语人名时，标注之间的信息非常重要（即英语人名对应的汉语人名的字和字之间有着很强的联系）。在机器学习方法中，为了将音译问题转化成序列标注问题，我们将汉字仅仅依附到它对应的英文字母串的第一个字母上，并且人为地将其他的字母对应到伪翻译'-'上。通过这种处理后标注间的转移关系有些失真，因此对于标注特征（标注之间的转移特征）利用得不是很充分。为了克服序列标注方法在音译时对标注转移信息利用不充分的弱点，下面将采用机器翻译的方法来进行翻译。

（4）基于统计翻译的音译模型

本文将基于短语和基于双语 N-gram 的翻译模型应用于英汉人名音译。本文使用了与机器学习算法中一样的训练集来获取翻译模型。本文在不同的语言模型上实验了两种统计翻译模型的性能。下面给出了没有使用语言模型时（这时是各个模型的简化版本）、使用在语料 1 上训练的语言模型时（语料 1 包含 14,436 个汉语单语人名）和使用在语料 2 上训练的语言模型时（语料 2 包含 LDC 上 56 万多个汉语单语人名）两种翻译模型的性能。

在基于短语的翻译模型中，在进行完短语抽取后，得到了 4,149 个长度为 1 的短语对，16,805 个长度为 2 的短语对，14,388 个长度为 3 的短语对（在短语抽取过程中，本文以汉字的长度为短语长度）。本文中的语言模型均采用三元语言模型。在本文的实验中，当短语长度为 1 时，系统的效果最好。下面给出了短语长度为 1 时在各个语言模型上的基于短语的音译系统的性能。

表 3 基于短语的翻译方法在不同语言模型上的音译结果

	Top1	2-5	6-10	11-50	51-100	101-500	总数
不使用语言模型	12	31	29	105	68	147	392
在语料 1 上训练的语言模型	285	202	76	136	37	58	794
在语料 2 上训练的语言模型	447	249	67	64	13	5	845

通过表 3 可以看出基于短语的翻译系统非常依赖于语言模型的作用。当没有语言模型时，系统往往会选择那些出现频率比较低的汉字组合，因为当汉字出现频率低时，它所对应的英文短语少，因而短语间的相对概率大，从而翻译概率也大。当使用在语料 2 上训练的语言模型时，系统效果比使用在语料 1 上训练的语言模型要好很多，这说明把英文人名翻译成汉语时，用字是有一定规律的，好的语言模型对结果有很强的重排序作用，从而提高系统的性能。

在基于双语 N-gram 的翻译模型中，在进行完元组抽取后，从训练语料中得到了 4,138 个基本元组，21,195 个二元元组，4,828 个三元元组。基于双语 N-gram 模型使用的双语语言模型和目标语言的单语语言模型均采用三元语言模型。表 4 给出了在各个语言模型下系统的最优性能：

表 4 基于双语 N-gram 的翻译方法在不同语言模型上的音译结果

	Top1	2-5	6-10	11-50	51-100	101-500	总数

无语言模型	363	208	75	111	25	32	814
在语料 1 上训练的语言模型	377	193	69	122	24	32	817
在语料 2 上训练的语言模型	419	230	67	92	19	12	839

从表 4 可以看出，基于双语 N-gram 的音译模型由于有双语语言模型的作用，因而受目标语言的语言模型的影响较小。当使用的目标语言模型是在语料 1 上训练时，这时使用语言模型对系统的性能影响不大。但是当使用更好的目标语言的语言模型时，系统的性能得到了较大的提升，但是性能提升没有基于短语的模型那么明显。

表 5 对比了当语言模型和翻译模型都从语料 1 上训练得到时，上面四种方法的性能。最大熵模型和条件随机场模型给出的均是在最好的特征集上取得的效果。

表 5 四种音译模型在同一训练语料上的音译性能对比

	Top1	2-5	6-10	11-50	51-100	101-500	总数
最大熵模型	322	205	69	168	69	132	965
条件随机场模型	348	201	67	153	67	107	943
基于短语的模型	285	202	76	136	37	58	794
基于双语 N-gram 的模型	377	193	69	122	24	32	817

表 5 表明使用本文的实验语料时，基于双语 N-gram 的模型的性能最好，条件随机场模型次之。本文对比了这四种翻译模型的音译结果，发现对于大部分人名，四种翻译模型的区别仅仅是正确答案在结果集中的排序位置不同。但是对于一小部分人名，用序列标注的方法的翻译结果同统计翻译方法的翻译结果有较大区别。这是因为两类方法进行音译时的原理不同引起的，序列标注算法在进行音译时是综合各种特征进行推断，因而有一定的泛化能力，而统计翻译方法的原理是对训练语料进行组合，形成一个最优的翻译，能较好的处理语料中出现过的情况。

当所有信息来自于同一语料时，利用序列标注算法进行音译取得了与机器翻译方法差不多的效果。但是基于机器学习的方法也有其局限性。首先，如果采用大规模语料集的话，这时候可供选择的汉字（标注）会更多，那么训练时间会更长，也更加消耗内存，需要在大内存上的机器上才能完成训练。另一方面，最大熵和条件随机场模型均是在双语训练语料上获得的，它们的所有的信息均来自双语训练语料。而采用翻译方法进行音译时，可以从其他的资源中获取信息，比如在本文的实验中，当使用的语言模型来自更大的语料时，基于统计翻译的音译模型的性能得到了很大的提升，远远超过了基于序列标注的方法，所以基于翻译的方法的框架更加灵活，获得的信息的渠道更广，因而更加适合人名翻译。

6 结语

本文提出了将英汉人名音译问题转化为序列标注问题的处理思路，使用了两种机器学习方法：最大熵模型和条件随机场模型来进行英汉人名音译。本文首先比较了最大熵模型在不同的特征集上的效果，实验结果表明最大熵模型能结合各种特征进行最优选择，但是这并不代表不需要设计好的特征集，只有把与问题相关的特征加入到特征集中，最大熵模

型才有可能有好的效果。最大熵模型在很多 NLP 问题中有偏置问题，而条件随机场模型能避免偏置问题，为了防止可能存在的偏置问题，本文还使用了一阶线性链条件随机场方法来进行音译。同最大熵模型一样，条件随机场模型也需要针对问题构建合适的特征集。实验结果表明，基于条件随机场模型的音译效果要比最大熵的音译效果要稍好一些。

除了从序列标注的视角处理英汉人名音译问题外，本文还从翻译的角度来处理音译问题，并将基于短语和基于双语 N-gram 的两种机器翻译模型用于音译。实验结果表明统计翻译的方法比机器学习方法的要更适合于音译问题。基于翻译的方法框架比较灵活，能有效利用更多的外来信息，因而效果比机器学习方法要好。

尽管本文使用的几种方法均取得了不错的效果，但是这些方法的 top1 的正确率仍然不太理想，每种方法均有大量的正确翻译排在后面，如果能把这些正确翻译的排序提前，那么能大大提高系统的性能。上面的几种统计方法的实验结果在各个区间的数据分布上有一定的相似性，这一方面说明了统计方法的局限性，统计方法只能反映数据的普遍规律，对于特殊情况，是比较难以处理的。同时也表明音译的确是一个比较复杂的问题，人们在翻译时往往随意性比较大，因此通过统计方法不能完全解决问题，需要采用更好的方法来解决统计方法存在的问题。

近年来，随着网络资源的爆炸式增长，研究人员开始尝试从网络中获得各种各样有用的资源。已经有研究人员开始将网络挖掘应用到音译上，并取得了一定的效果。我们下一步可以尝试将统计方法和网络挖掘结合起来，用更有效的方法进行音译。

参考文献：

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

[2] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML-01*, pages 282.289, 2001.

[3] Philipp Koehn, Franz Josef Och, and Daniel Marcu, 2003. Statistical Phrase-based Translation, In Proceedings of HLT-NAACL, Edmonton, Canada

[4] J.B. Mari˜no, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-juss`a. 2006. N-gram based machine translation. Computational Linguistics, 32(4):527–549.

[5] Virga Paola and Sanjeev Khudanpur. 2003.Transliteration of Proper Names in Cross lingual Information Retrieval. Proceedings of the ACL2003 Workshop on Multilingual and Mixed language Named Entity Recognition, Sapporo , Japan, 57-60.

[6] S. Chen and R. Rosenfeld. 2000. A Survey of Smoothing Techniques for ME Models. IEEE Trans. Speech and Audio Processing, 8(1), pp. 37.50. January 2000.

[7]Zhang, L., A Maximum Entropy Modeling Toolkit for Python and C++, http://www.nlplab.cn/zhangle/

[8] Taku Kudo CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net/

[9] 何中军 Camel Version1.0 使用手册及系统描述

[10] MARIE A N-gram-based Statistical Machine Translation decoder http://gps-tsc.upc.es/veu/soft/soft/marie/