新闻语料中中日命名实体词汇翻译的自动抽取

尹存燕，黄书剑，戴新宇，陈家骏
（南京大学 计算机软件新技术国家重点实验室，江苏 南京 210023；
南京大学 计算机科学与技术系，江苏 南京 210023）

摘要：本文首先对新闻语料中中日命名实体的翻译特征进行了分析，基于这些特征，本文提出了一种中日双语命名实体词汇翻译自动抽取的方法，该方法融合了中日汉字翻译概率、片假名词汇和中文词汇的拼音矩阵以及双语词汇共现等特征。实验表明本文方法取得较好的效果。
关键词：命名实体；双语语料；对齐模型；拼音矩阵；词汇共现

Automatic Named Entity Translation Extraction From Sino-Japanese Bilingual News Corpus

YIN Cunyan， HUANG Shujian， DAI Xinyu， CHEN Jiajun
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210023,China;
Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu 210023,China)

Abstract: Based on the analysis of Sino-Japanese named entity translation in the news corpus, this paper proposes a novel approach for automatic extraction of named entity translation. We find there are several valuable features that can be used: the translation probability of Japanese kanji and Chinese character, the phonetic matrix between katakana and Chinese word and the source word and target word's co-occurrence count. The experiment shows that our approach can extract the named entity translation with relatively high precision.
Key words: named entity; bilingual corpus; alignment model; phonetic matrix; word's co-occurrence

1 引言

命名实体(Named Entity)主要是指人名、地名、机构名等以名称作为标识的实体。命名实体往往是专有名词，并且在语言表达中包含了重要信息[1]，因此命名实体词汇的翻译对于跨语言信息检索、机器翻译等应用而言是较为重要的翻译知识，此外命名实体词汇的翻译对于双语词典来说，往往属于未登录词，可以对已有的双语词典进行有效的补充。

双语命名实体词汇翻译的自动抽取过程，一般可以分为两个步骤：（1）在源语言

或双语环境中进行命名实体词汇的识别；（2）对识别出的命名实体词汇，利用各种对齐特征，进行翻译对的自动抽取。就目前的研究成果来看，这两个步骤都是具有挑战性的工作。由于命名实体词汇识别本身会出现错误，因此在实体词汇翻译抽取过程中会出现错误延续，甚至扩大的可能。对于中文和日文这类词汇之间没有空格区分的语言而言，在第一个步骤命名实体识别之前，还需要进行分词，同样，分词中的错误也会延续到之后的处理过程中。

在基于平行语料库的命名实体词汇翻译自动抽取的研究中，一般是分别在源语言和目标语言中识别出命名实体词汇，然后利用词典查找、音译模型等实现命名实体的对齐[2-4]。Huang Fei 等人为了避免词典覆盖的局限性以及音译模拟模型缺少上下文有关信息，提出了一种基于最小代价的线性多特征融合模型，以提高命名实体词汇翻译的抽取效果[5]。这些方法基于的前提是双语都进行命名实体识别，对于只在一种语言中进行了命名实体识别来说就不适用了，Donghui Feng 等人针对这样的情况，提出利用最大熵模型集成多种特征实现命名实体翻译词汇的对齐，这些特征包括：翻译模型、音译模型、共现模型以及扭曲模型[6]。现有的命名实体词汇翻译抽取研究主要是基于中英双语，对于中日双语研究较少，而在现实生产生活中，特别是新闻领域，需要中日命名实体翻译的应用仅次于中英双语，因此有必要研究如何实现中日命名实体翻译的自动抽取。已有的研究主要是针对中英双语的命名实体翻译抽取，而中日翻译中有着不同于中英翻译的特点，现有方法照搬到中日双语命名实体抽取中存在着局限性，并且忽略了中日翻译中有助于命名实体词汇翻译抽取的特征。本文在详细分析中日双语新闻中命名实体词汇翻译特点的基础上，提出了一种融合汉字翻译概率、中文词汇和片假名词汇的拼音矩阵以及双语词汇共现特征的中日命名实体词汇翻译抽取方法。

本文第二节分析了中日双语新闻中命名实体词汇翻译的特点；第三节提出了一种中日命名实体词汇翻译的抽取方法，详细介绍了方法中各种特征的计算方法；第四节给出实验结果，针对结果中的错误分析，提出了改进方法；最后一节对本文的工作进行总结和展望。

2 中日双语新闻中命名实体词汇的翻译特点

在中日双语新闻中，存在着大量的人名、地名以及机构名。由于中文和日文在语言发展历程上的相互影响，因此中日命名实体词汇的翻译存在以下特点：

（1） 中国和日本人名在中日翻译中，通常采用汉字。比如中文新闻中出现的人名：“钟南山”、“丰田章男”，对应的日文为：“鐘南山”、“豊田章男”。在这种情况下，中日翻译之间只是存在字体的变化。但是，由于目前日本政府颁布的“常用汉字”只有 2 千个左右，因此新闻中出现的中文人名汉字常常不在其中，比如：“谭晶”的“谭”，“杨洁篪”的“杨”和“篪”在这种情况下，进行日文翻译时会采用两种方式，一种是直接进行简繁体转换，比如“谭晶”翻译为“譚

晶”；另一种用发音近似的片假名替代，比如“杨洁篪”翻译为“楊潔チ”。不过，在中国人名翻译中也存在例外，比如“成龙”并是不翻译为“成龍”，而是ジャッキー·チェン，这是对应成龙的英文名“Jackie Chan”的音译。

（2） 对于西方人名，在中日翻译通常都是采用音译，区别是：中文使用发音近似的汉字，而日文使用发音近似的片假名。比如：“Clinton”，中文翻译为“克林顿”，日文翻译为“クリントン”；“Bill Gates”，中文翻译为“比尔·盖茨”，日文翻译为“ビル・ゲイツ”。不过，也存在着例外，比如“Kevin Rudd”，中文翻译为“陆克文”，日文翻译为“ケビン・ラッド”。

（3） 机构名在中日翻译中有三种情况：第一种是都是使用汉字，只是字体不同，比如:“铁道部”翻译为“鉄道部”；第二种是中日都有固定词汇，因而采用意译，比如“安理会”翻译为“安保理”；第三种是都采用英文缩略语，比如“WTO”。不过在机构名翻译中，常常会出现一方使用简称的情况，比如：“经保分局”翻译为“経済文化保衛分局”，“哈佛大学”翻译为“ハーバード大”（日文的全称应该是“ハーバード大学”）。

（4） 地名词汇在中日双语新闻的翻译特点和有人名的翻译有相似之处，不过地名中常常出现汉字、片假名混和的情况，比如：“新疆维吾尔自治区”翻译为“新疆ウイグル自治区”，“加利福利亚州”翻译为“カリフォルニア州”。

以上列出的翻译特点在本文的中日命名实体词汇翻译抽取中有着重要的影响，下一节中本文详细说明如何将这些翻译特点作为特征融合在自动抽取过程中。有一点需要说明的是，上文列出的翻译特点，是针对中译日，并且译者母语为中文的双语新闻中总结出来的，如果译者母语是日文，在翻译中可能还有另外的特点，不是本文关注的重点。

3 命名实体词汇翻译的自动抽取

本文的研究工作是基于中日双语新闻语料，在进行中日命名实体词汇翻译抽取之前，首先分别对中日语料进行分词，并在中文一方进行命名实体词汇的识别；然后针对中日双语新闻中的命名实体翻译特点，采用融合多特征的命名实体翻译抽取方法，抽取步骤为：（1）利用中日汉字翻译概率实现全汉字词汇翻译的抽取；（2）利用拼音矩阵特征和共现特征实现片假名词汇翻译的抽取。下面详细介绍每步的实现过程。

3.1 日文全汉字词汇翻译抽取

通过对中日命名实体翻译的特点分析，我们可以看出中文命名实体在翻译成日文时，常常会出现全汉字词汇（词汇中的字符全是汉字）。在本文研究的语料中，日文句子中一共出现了 10217 个全汉字词汇，涉及到 2580 个汉字。对于日文全汉字词汇，由于日文汉字和中文汉字的集合存在差别，并且在全汉字词汇中，有的是意译结果，比如中文词汇“外交部”翻译为日文的“外務省”，因此无法通过字体转换进行命名实体翻译的抽取。针对日文命名实体为全汉字的词汇，本文利用中日汉字翻译概率，来计算中文命

名实体词汇和它的翻译概率。

假定，中文命名实体词汇 ne_c 由 n 个汉字组成，表示为：$ne_c=\{c_1, c_2, \ldots, c_n\}$，日文全汉字词汇 ne_j 由 m 个汉字组成，表示为：$ne_j=\{j_1, j_2, \ldots, j_m\}$，这两个词汇的翻译概率 $P_{kanji}(ne_j|ne_c)$ 表示为下面的公式：

$$P_{kanji}(ne_j|ne_c) = \prod_{y=1}^{m}\sum_{x=1}^{n} p(j_y|c_x) \qquad (3.1)$$

在公式 3.1 中，$p(j_y|c_x)$ 表示中文汉字翻译 c_x 为日文汉字 j_y 的概率，$p(j_y|c_x)$ 可以利用 IBM Model 1[7]从中日双语平行语料中获得。

在一个中文句子中，针对每一个中文命名实体词汇，计算平行的日文句子中所有全汉字词汇和它的翻译概率，取概率值最高的全汉字词汇作为该中文命名实体的翻译。

3.2 片假名词汇翻译抽取

完成 3.1 节说明的抽取步骤后，对于语料中没有找到翻译结果的中文命名实体词汇，本文认为其可能翻译为片假名词汇，下面介绍片假名词汇翻译抽取的步骤。

对于外来语的翻译，中文和日文中都有音译方式，并且在人名、地名中采用音译方式的比例很高。在中文中，音译的方式是使用发音近似的汉字来表示；在日文中，音译的方式是用发音近似的片假名来表示。比如：“Boston”，中文翻译为“波士顿”，汉字对应的拼音序列为“boshidun”，日文翻译为“ボストン”，片假名对应的罗马音序列为“bositon”。

经典的音译模型是基于统计的音译模型[8,9]，也就是利用统计方法从标注好的数据中学习音译规律。中文字符和日文字符直接利用统计的音译模型并不合适，因为日文片假名字符（清音、浊音、半浊音）不到 80 个，而中文外来语词汇中的汉字则可以达到上千。Donghui Feng 等人在研究中计算中文词汇的拼音字符串和英文字符串的 XDice’s 系数，以此作为中文词汇和英文词汇的互为翻译的特征之一[6]。本文在研究中文词汇和对应的片假名词汇时发现，利用中文词汇拼音序列和片假名罗马音序列构成的拼音矩阵，计算中文词汇和片假名词汇的翻译概率比利用 XDice’s 系数更有效。

拼音矩阵的构造方法是：假设中文命名实体词汇 ne_c 的汉字拼音序列表示为：$ne_c=\{p_1, p_2, \ldots, p_n\}$，日文片假名词汇 ne_j 的罗马拼音序列表示为：$ne_j=\{r_1, r_2, \ldots, r_m\}$，则这两个词汇的拼音矩阵 $M(ne_c, ne_j)$ 可以表示为：$M(ne_c, ne_j)=[m_{ij}]$，其中 $i\in[1,n]$，$j\in[1,m]$，

$$m_{ij} = \begin{cases} 1, & p_i\text{和}r_j\text{相同} \\ 0, & p_i\text{和}r_j\text{不相同相同} \end{cases} \qquad (3.2)$$

比如：“波士顿”和“ボストン”对应的拼音矩阵如表 1 所示：

表 1：拼音矩阵

	b	o	s	i	t	o	n
b	1	0	0	0	0	0	0
o	0	1	0	0	0	1	0
s	0	0	1	0	0	0	0
h	0	0	0	0	0	0	0
i	0	0	0	1	0	0	0
d	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0
n	0	0	0	0	0	0	1

拼音矩阵特征$Transli(ne_c, ne_j)$的计算公式为：

$$Transli(ne_c, ne_j) = \frac{\sum_{i=j} m_{ij} + \sum_{j=i+1} m_{ij} + \sum_{j=i-1} m_{ij}}{min(n,m)} \quad (3.3)$$

从公式 3.3 可以看出，公式中的分母$min(n,m)$表示汉字拼音序列和片假名罗马拼音序列长度值的较小者，公式中的分子：$\sum_{i=j} m_{ij} + \sum_{j=i+1} m_{ij} + \sum_{j=i-1} m_{ij}$，实际上是在计算矩阵对角线上以及临近的元素的和，这些元素不仅可以体现 Donghui Feng 等人提出的 XDice's 系数计算中涉及的连续 2 元子串以及间接 2 元子串，还可以体现相同的 1 元子串，因此本文的拼音矩阵特征要比 XDice's 系数更有效地反映互译词汇之间的音译特征。

除了拼音矩阵特征之外，我们还加入了词汇共现特征，共现特征$Cooc(ne_c, ne_j)$的计算公式如下：

$$Cooc(ne_c, ne_j) = \frac{count(ne_c, ne_j)}{\sum count(ne_c)} + \frac{count(ne_c, ne_j)}{\sum count(ne_j)} \quad (3.4)$$

其中$count(ne_c, ne_j)$，表示中文命名实体 ne_c 和日文平假名词汇 ne_j 在平行句对中同时出现的次数，$count(ne_c)$表示中文命名实体 ne_c 在整个语料中出现的次数，$count(ne_j)$表示日文平假名词汇 ne_j 在整个语料中出现的次数。

对于中文命名实体词汇 ne_c 和日文片假名词汇 ne_j 的翻译概率 $P_{kata}(ne_j|ne_c)$ 计算公式如下：

$$P_{kata}(ne_j|ne_c) = \alpha Transli(ne_c, ne_j) + \beta Cooc(ne_c, ne_j) \quad (3.5)$$

其中$Transli(ne_c, ne_j)$，表示 ne_c 和 ne_j 的拼音矩阵特征，$Cooc(ne_c, ne_j)$表示 ne_c 和 ne_j 的共现特征，α表示拼音矩阵特征的权重，β表示共现特征的权重，因为这两个特征都是重要的特征，在实验中我们将这两个权重值均设为经验值 0.5。对于一个中文命名实体词汇的候选片假名词汇翻译中，取翻译概率最高的作为结果输出。

4 实验结果及分析

本文研究的语料是新闻网站中的中日双语新闻，包含 6300 条平行句对。中文句对中包含命名实体词汇的句子占 98.6%，平均每个中文句子包含 1.7 个命名实体。在进行

命名实体词汇翻译抽取前，本文利用中科院分词工具 ICTCLAS(http://ictclas.org.cn)进行分词、词性标注，在词性标注结果中以“/nr”作为人名词汇的识别结果，“/ns”作为地名词汇的识别结果，“/nt”作为机构名词汇的识别结果。日文采用分词工具 MeCab(http://sourceforge.net/projects/mecab/)对日文进行分词以及词性标注。实验结果评价采用准确率（Precision,P)、召回率(Recall,R)和 F 值（F-score，F)。准确率 P 的计算方法如公式 4.1 所示，召回率 R 的计算方法如公式 4.2 所示，F-sorce 的计算方法如公式 4.3 所示[10]。

$$\mathrm{P}=\frac{\text{正确翻译对个数}}{\text{抽取出的所有命名实体词汇翻译对}}\times 100\% \qquad (4.1)$$

$$\mathrm{R}=\frac{\text{正确翻译对个数}}{\text{语料中的所有命名实体词汇翻译对}}\times 100\% \qquad (4.2)$$

$$\mathrm{F}=\frac{2*P*R}{P+R}\times 100\% \qquad (4.3)$$

命名实体翻译中，由于中日分词的标准不同，因此命名实体词汇存在一对多和多对一的情况。我们在统计词汇翻译抽取结果时，如果一方是另一方实际翻译结果的子串，也认为该结果正确。比如：抽取出“马朝旭--馬”，统计中这个结果属于正确。

已有命名实体词汇翻译抽取的研究中，双语语言大多是以中英双语作为研究对象[2,5,6]，这些研究中命名实体词汇翻译抽取方法，一般是在 IBM Model 词对齐的基础上，加入中英翻译的音译模型以及共现特征抽取词汇翻译。考虑到命名实体词汇翻译中一对多和多对一的情况，本文命名实体词汇翻译抽取实验的 baseline 系统抽取步骤为：（1）分别采用 IBM Model 4 和 HMM（Hidden Markov Model）进行双向（中文到日文，日文到中文）对齐；（2）在 IBM Model 4 和 HMM 的双向对齐结果中，对中文命名实体词汇对应的日文翻译中概率最高的前三名求交集，以此作为两个模型的抽取结果；（3）将两个模型的抽取结果求交集，以此作为 baseline 系统的结果。

本文的命名实体词汇翻译抽取方法对于人名、地名以及机构名来说没有区别，但是日文全汉字词汇和片假名词汇的抽取结果有所不同，因此 baseline 的实验结果和本文方法的实验结果对日文全汉字词汇翻译和片假名词汇翻译的抽取分别进行评价，并取两者的平均值作为综合评价的数值。实验结果如表 2 所示：

表 2： 命名实体词汇的抽取结果评价

	P/%	R/%	F/%
Baseline 日文全汉字词汇	60.58	69.52	64.74
Baseline 片假名词汇	67.74	74.87	71.13
Baseline 综合	64.16	72.19	67.94
日文全汉字词汇	96.45	98.17	97.30
片假名词汇	74.73	60.83	67.06
综合	85.59	79.50	82.43

在表 2 中的结果中可以看到，如果中文命名实体翻译成日文全汉字词汇，则利用汉

字翻译概率特征可以取得很好的效果。本文对日文全汉字词汇抽取结果错误分析中发现，分词中的错误是影响精确率和召回率的主要原因。命名实体往往属于未登录词汇，并且在语料中出现的次数较少，现有的中文分词工具对于命名实体的分词容易产生错误。比如下面的例句中，分词结果就出现了错误。

例 1：

原文：*该党干事长冈田克也将出任外务大臣。*

分词结果：*该党 干事长 冈田克 也 将 出任 外务 大 臣 。*

命名实体识别结果：*该党/r 干事长/n 冈田克/nr 也/c 将/d 出任/v 外务/n 大/a 臣/n 。/w*

说明：“/nr”是人名的标注

日文分词结果：*外務/n 大臣/n に/u 岡田/n 克也/n 幹事/n 長/n を/u 起用/n する/v 。/w*

在例 1 中，人名“冈田克也”被错误的分成了“冈田克”和“也”，并且将“也”的词性标注为“/c”（连词标识）。

例 2：

原文：*他们考察了宫城县石卷市的两家水产品加工厂。*

分词结果：*他们 考察 了宫城 县 石 卷 市 的 两 家 水产品 加工厂 。*

命名实体识别结果：*他们/r 考察/n 了宫城/ns 县/n 石/j 卷/ng 市/n 的/b 两/m 家/k 水产品/n 加工厂/n 。/w*

说明：*“/ns”是地名的标注*

日文分词结果：*担当/n 者/n ら/n は/u 宮城/n 県/n 石巻/n 市/n の/u 水産/n 品/n 加工/n 場/n 2/n 社/n を/u 視察/n し/v た/uv 。/w*

在例 2 中，地名“宫城县”在分词中，将前面的“了”和“宫城”合并在一起成为“了宫城”，并且识别为“/ns”（地名标识）。地名“石卷市”分成了“石”、“卷”、“市”，并且词性分别被标注成“/j”（简称标识）、“/ng”（名词性语素标识）以及“/n”（名词标识）。

上面两个例子中的分词错误，会延续到命名实体词汇翻译的抽取，从而产生错误：本文利用汉字翻译概率抽取的日文全汉字词汇翻译时，只计算中文命名实体词汇和日文全汉字词汇的翻译概率。例 1 中“*也/c*”，由于它没有命名实体的标识，因此不会被抽取出来，类似的，例 2 中的“*石/j 卷/ng 市/n*”，也不在抽取结果之中。而“了宫城”虽然是一个错误的地名，但是由于有地名的标注，因而会被抽取出来。

在片假名词汇抽取的实验中，召回率相对于日文全汉字词汇有较大差距，其主要原因是本文利用的拼音矩阵特征是针对仅由片假名字符组成的词汇，而对于片假名和中文混合的日文词汇则并不适用。针对这一情况，本文对日文分词工具 MeCab 产生的结果进行了调整：将日文分词结果中片假名和中文混和的词汇进行再切分，分成仅有片假名字符组成的词汇以及全汉字词汇。在日文分词调整后的双语语料中，切分产生的全汉字词汇对于日文全汉字词汇的命名实体翻译抽取实验结果几乎没有影响，而对于片假名词汇的命名实体翻译抽取的实验结果是有影响的，分词调整前后的实验结果对比如表 3 所示：

表 3： 日文分词调整对片假名命名实体词汇翻译抽取的影响

	P/%	R/%	F/%
调整前	74.73	60.83	67.06
调整后	73.52	86.17	79.34

从表 3 中我们可以看出，日文分词的调整在片假名命名实体词汇翻译抽取的召回率上有明显的提高。

5 结束语

本文提出了一种中日双语命名实体词汇翻译的自动抽取方法，该方法融合了中日汉字翻译概率、片假名词汇和中文词汇的拼音矩阵特征以及双语词汇共现等特征，实验结果表明本文提出的方法能有效的从中日双语新闻语料中抽取出命名实体词汇翻译。我们下一步的工作将考虑把日文命名实体的识别作为新特征融合进来，以提高命名实体词汇翻译抽取效果，并且在命名实体词汇翻译抽取的基础上进一步实现命名实体短语翻译的抽取。

参考文献

[1] Hobbs, J. et al. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural Language Text, MIT Press. Cambridge, MA. 1996.

[2] Y. Al-Onaizan , and K. Knight . Translating Named Entity Using Bilingual and Monolingual Resources [C]// Proceedings of Association of Computational Linguistics , Philadelphia PA : 2002.

[3] H. Meng , W. K. Lo , B. Chen , and K. Tang. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval [C]// Proceedings of the Automatic Speech Recognition and Understanding Workshop , Trento ,Italy : 2001.

[4] B. Stalls , and K. Knight . Translating Names and Technical Terms in Arabic Text [C]// Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, Philadelphia , Pennsylvania : 1998.

[5] Huang Fei, Vogel S , Waibel A. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-Feature Cost Minimization [C]//Proceedings of the ACL. Sapporo, Japan, 2003. 184-192

[6] DongHui Feng, Yanjuan Lv, Ming Zhou. A New Approach for English-Chinese Named Entity Alignment. [C]//Proceedings of EMNLP 2004.

[7] Peter F. Brown, Stephen A. Della Pietra, Vincent J .Della Pietra , and Robert L. Mercer. The mathematics of Statistical Machine Translation: Parameter Estimation [J] . Computational Linguistics , 1993 , 19(2):263-311.

[8] Gao Wei. Phoneme based Statistical Transliteration of Foreign Names for OOV Problem [D] . The Chinese University of Hong Kong . 2004.

[9] Wei Hao Lin and Hsin Hsi Chen. Backward Machine Transliteration by Learning Phonetic Similarity[A]. the 6th Conference on Natural Language Learning [C] 2002:1-7
[10] Daniel Jurafsky， James H. Martin 著. 冯志伟,孙乐 译 自然语言处理综论[M] 北京：电子工业出版社. 2005

感谢
国家社会科学基金重点项目（编号：11AZD121）
国家自然科学基金（课题编号：61003112）
优秀国家重点实验室研究项目（课题编号：61223003）

作者简介：
尹存燕（1976-），女，讲师，主要研究方向为自然语言处理；
黄书剑（1984-），男，助理研究员，主要研究方向为统计机器翻译和自然语言处理；
戴新宇（1979-），男，副教授，主要研究方向为自然语言处理；
陈家骏（1963-），男，教授，博士生导师，主要研究方向为自然语言处理；