文章编号: 1003-0077(2008)04-0055-06

# 一种命名实体翻译等价对的抽取方法

陈怀兴, 尹存燕, 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 210093)

**摘 要**: 有关命名实体的翻译等价对在多语言处理中有着非常重要的意义。在过去的几年里,双语字典查找,音译模型等方法先后被提出。另一种极具价值的方法是从平行语料库中自动抽取有关命名实体的翻译等价对,现有的方法要求预先对双语语料库的两种语言文本进行命名实体标注。提出了一种只要求对语料库中源语言进行命名实体标注,目标语言不需标注,然后利用训练得到的 HMM 词对齐结果来抽取有关命名实体翻译等价对的方法。在实验中,把中文作为源语言,英文作为目标语言。实验结果表明用该方法,即使在对齐模型只是部分准确的情况下,也得到了较高正确率的命名实体翻译等价对。

**关键词**: 人工智能;机器翻译;命名实体; 翻译等价对; HMM;对齐模型

**中图分类号**: TP391 **文献标识码**: A

## An Approach to Extract Named Entity Translingual Equivalence

CHEN Huai-xing, YIN Cun-yan, CHEN Jia-jun

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China)

**Abstract:** Identification of translingual equivalence of named entities is substantial to multilingual natural language processing. Some approaches to named entity translation, such as bilingual dictionary lookup, word/sub-word translation or transliteration, have been explored in the past years. Another promising approach is to extract named entity translingual equivalence automatically from a parallel corpus, which usually requires the named entities to be annotated manually or automatically for both languages. In this paper, we propose a new approach to extract equivalence of named entities from a parallel corpus with only the source language annotation and the result of HMM alignment. The experiment is carried in a Chinese-English parallel copus, and we treat Chinese as the source language and English as the target language. The result shows that our new approach achieves high quality of named entity pairs with relatively high precision, even though sometimes the word alignment result is partially correct.

**Key words:** artificial intelligence; machine translation; named entity; translingual equivalence; HMM; alignment model

## 1 引言

翻译等价对是指具有语义上等同关系的不同语言的表达。因为命名实体,尤其是人名、地名、机构名,在语言中往往传达着非常有用的信息[1],因此,取得命名实体的翻译等价对,对于多语言信息处理(如机器翻译,跨语言信息检索与自动问答等)有着非常重要的意义。在过去的几年里,一些处理跨语言的命名实体翻译等价对的方法,如双语字典查找,音译模型等先后被提出[2~4]。但是,由于基于双语字典查找方法的字典有限覆盖性,以及基于词或字的音译模型由于缺少考虑上下文有关信息,它们很多时候都不能得到令人满意的结果[5]。

**收稿日期**: 2007-09-07 **定稿日期**: 2008-01-14

**基金项目**: 国家 863 计划资助项目(2006AA01Z143,2006AA01Z139);国家自然科学基金资助项目(60673043);江苏省自然科学基金资助项目(BK2006117)

**作者简介**: 陈怀兴(1982—),男,硕士生,主要研究领域为机器翻译与算法;尹存燕(1976—),女,博士生,主要研究领域为自然语言处理;陈家骏(1963—),男,教授,博士生导师,主要研究领域为自然语言处理、机器翻译。

另一种可行的方法是从平行语料库中自动抽取有关命名实体的翻译等价对。Huang 等在文献[5]中提出了一种基于多特征代价最小的自动抽取命名实体翻译等价对的方法。他们的实验表明,用他们提出的方法抽取命名实体的翻译等价对比以前的方法更加有效,并且,当把抽取得到的翻译等价对加入到统计机器翻译系统时,翻译质量有了显著的提高。但是,他们的抽取方法,一方面要求平行语料库中的源语言与目标语言中的命名实体都需要预先标注,这样增加了工作量;另一方面当标注出现错误,特别是一种语言标对了,而另一种语言标错了,他们的对齐模型无法纠正这个标注错误,这将严重地影响到最后抽取的命名实体翻译等价对的正确率。所以,在本文中,我们提出了一种从只在源语言中标注了命名实体的平行语料库中进行有关命名实体翻译等价对的抽取方法。我们先用 HMM 词对齐模型对平行语料库进行对齐,然后基于对齐模型上的短语抽取技术[9,10,12],先生成与源语言相对应的目标语言的候选命名实体翻译等价单位,然后对它们进行三种置信度估计,最后综合得到与源语言命名实体最为匹配的目标语言命名实体翻译等价单位。在实验中,我们用中文作为源语言,英文作为目标语言。最后的实验结果表明,我们在较小的语料库中得到了 87.75 %的正确率,而在较大语料库中,正确率为 83.46 %。

在本文的以下部分中,我们先简单回顾一下 IBM 对齐模型与 HMM 对齐模型,以及对齐后可以得到的信息。然后在第 3 部分详细讨论了有关在对齐模型上的命名实体翻译等价对的抽取方法。第 4 部分是实验与讨论。最后我们总结了本文所提方法的优点与缺点以及今后可能进行的扩展工作。

## 2 对齐模型

在 IBM 统计翻译模型[6]中包括一个语言模型和一个翻译模型,翻译模型可以表示成 $p(s \mid t) = \sum_{a} p(s, a \mid t)$,其中,隐含的中间变量 $a = a_1 a_2 \ldots a_I$,用来表示源语言句子中的词与目标语言句子中的词之间的对应关系,$a_i$ 表示源语言句中第 $i$ 个词对应的目标语言句子中词的位置。在一对句子的所有对齐方式中,其训练对齐模型中的最大可能的对齐方式通常被称为最大近似对齐(maximum approximation)。在 HMM 对齐模型[7]下,最大近似对齐就是所谓的韦特比(Viterbi)对齐 $\hat{a}_I$。即对齐 $\hat{a}_I$ 满足以下条件

$$\hat{a}_I = \arg\max_{a_I} p(s, a_I \mid t) \quad (1)$$

其中,在 IBM 对齐模型中,$p(s, a \mid t)$ 是由公式(2)表示:

$$p(s, a \mid t) = \prod_{i=1}^{I} p(a_i \mid i, I) \times p(s_i \mid t_{a_i}) \quad (2)$$

而在 HMM 对齐模型中,$p(s, a \mid t)$ 是由公式(3)表示:

$$p(s, a \mid t) = \prod_{i=1}^{I} p(a_i \mid a_{i-1}, I) \times p(s_i \mid t_{a_i}) \quad (3)$$

这里,$p(a_i \mid a_{i-1}, I)$ 表示源语言句子当前词对齐位置 $a_i$ 对前一个词对齐位置 $a_{i-1}$ 的依赖关系,$I$ 表示源语言的句长,$p(s_i \mid t_{a_i})$ 表示词的翻译概率。

在 HMM 对齐模型中,引入了当前词对齐位置 $a_i$ 对前一个词对齐位置 $a_{i-1}$ 的依赖关系,这将有利于对平行语料库中的局部化现象进行有效地建模[7];而在 IBM 词对齐模型中 $a_i$ 只依赖于对齐中的位置 $i$。

在一个句对中,源语言句子中的一些词的集合与目标语言句子中的一些词的集合可以组成一个翻译等价对,而这在对齐图上得到的是一个所谓的划分。图 1 表示的是源语言与目标语言的一对句子的词对齐情况,其中圆圈表示最大近似对齐得到的词对齐的结果(位置)。

| | t1 | t2 | t3 | t4 | t5 | t6 |
|---|---|---|---|---|---|---|
| s1 | ○ | | | | ○ | |
| s2 | | ○ | | | | |
| s3 | | | ○ | ○ | | ○ |
| s4 | | | | ○ | | |
| s5 | | | | | | ○ |

图 1 源语言中的词 $S_i$ 与目标语言中的词 $T_i$ 的对齐图

## 3 基于对齐模型的命名实体翻译等价对的抽取

本文的命名实体翻译对的抽取主要分为三步(总体流程见图 2)。

I 源语言(中文)中的命名实体的识别;

II 基于词对齐结果,生成与源语言中的命名实体相对应的目标语言候选翻译等价单位;

III 对生成的目标语言中的候选翻译等价单位进行置信度估计,最后按所得到的置信度来抽取命名实体翻译等价对。

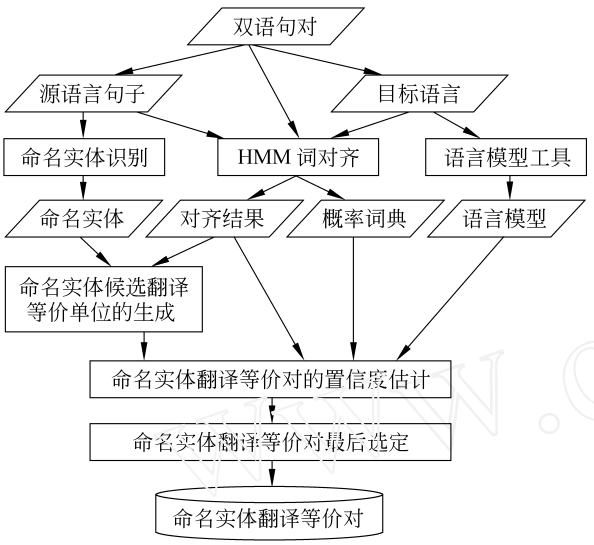


图 2　基于对齐模型的命名实体翻译等价对的抽取方法

对于源语言中的命名实体识别,本文使用了本课题组在文献[8]中介绍的中文命名实体识别方法,该方法在 2006 年 SIGHAN 微软语料库上的命名实体识别封闭测试中获得了最好成绩。

下面重点介绍 II 和 III 的工作。

### 3.1　候选命名实体翻译等价对的生成

在一个句对中,一个翻译等价对可以用一个四元组假设 $H_p(i, l_s, j, l_t)$ 来表示。其中,$i, j$ 分别表示源语言与目标语言中词的起始位置,而 $l_s, l_t$ 分别表示源语言与目标语言中等价单位的长度。例如,在图 1 中,$(2,2,2,3)$ 就是一个翻译等价对假设,即由源语言中$\langle s_2, s_3\rangle$对应于目标语言中的$\langle t_2, t_3, t_4\rangle$。本文的翻译等价对抽取任务就是根据对齐模型选取合适的四元假设。由于最大近似对齐中的位置对齐可能仅仅是部分准确的,而且从整个训练语料库中得到的词的翻译概率也可能存在误差,所以能够在部分准确的对齐模型中正确地抽取翻译等价对就显得非常重要。

对于每个在源语言中被识别的命名实体 $ps_{ne}$,我们需要生成与其相对应的候选的目标语言命名实体翻译等价单位。形式化地说,假设现在要列出与源语言中的命名实体 $ps_{ne}$ 相对应的候选的目标语言命名实体翻译等价单位,其实就是找出满足公式(4)的四元组假设 $H_{ps_{ne}}$。

$$H_{ps_{ne}} s.t.\ \forall H_p(i, l_s, j, l_t) \in H, s_i \cdots s_{i+l_s} = ps_{ne} \quad (4)$$

这里,$H$ 表示句对 $p(s,t)$ 中的对齐模型中的所有可能的划分,$s_i$ 表示句对 $p$ 中的源语言句子 $s$ 中的第 $i$ 个词。$H_{ps_{ne}}$ 为源语言中命名实体 $ps_{ne}$ 在目标语言中所有可能的翻译等价单位的集合。在下文中,我们约定 $h$ 是集合 $H_{ps_{ne}}$ 中的一个特定候选假设。

然后我们用滑动窗口的方法,即对于每个句对 $p(s,t)$,从其对齐图上抽取所有可能为源语言中的命名实体 $ps_{ne}$ 的目标语言翻译等价单位。例如,在图 3 中,假设 $s_2 s_3$ 是被正确识别的源语言命名实体,那么其所有可能的翻译等价单位就是图中的所有方框;其中每个方框表示一个可能的翻译等价单位。即在图 3 中 $s_2 s_3$ 的所有候选翻译等价单位为 $t_2$, $t_2 t_3$, $t_2 t_3 t_4$, $t_2 t_3 t_4 t_5 t_6$。用滑动窗口的方法产生的候选的目标语言的命名实体翻译单位的数量较大,其中有的候选命名实体单位与源语言中的命名实体翻译等价的可能性会很小。但这种能产生较大数量的候选命名实体翻译等价单位,使得我们的方法可以在仅仅是部分准确的对齐模型的情况下依然可能抽取得到正确的命名实体翻译等价对。所以接下来很重要的一步就是对各个候选命名实体翻译单位进行置信度估计,最后选取与源语言中的命名实体最匹配的候选命名实体翻译单位。

| | t1 | t2 | t3 | t4 | t5 | t6 |
|---|---|---|---|---|---|---|
| s1 | ○ | | | | ○ | |
| s2 | | ○ | | | | |
| s3 | | | ○ | ○ | | ○ |
| s4 | | | | ○ | | |
| s5 | | | | | | ○ |

图 3　对齐图上的候选翻译等价单位的滑动窗口

### 3.2　候选命名实体翻译等价对的置信度估计

我们需要对候选命名实体翻译等价对用某些指标进行置信度估计。每个候选命名实体翻译等价对都在对齐图上定义了一个划分,我们可以用一些特征与指标对这个划分进行置信度估计。下面,我们从三个特征来估计一个四元假设的翻译置信度: 最大近似对齐置信度估计,基于词翻译的置信度估计和语言模型置信度估计。得到这三个特征的估计后,我们再对它们做线性回归。最后选取得到与源语言命名实体最匹配的候选的目标语言命名实体翻译等价单位。

#### 3.2.1　最大近似对齐置信度估计

源语言中的任何一个命名实体 $ps_{ne}$ 所对应的任何一个候选的翻译等价单位都在最大近似对齐图中定义了一个划分。我们用最大近似对齐中的对齐点是否与这个划分一致作为一个特征来估计这对命名

实体翻译等价对的置信度。对齐点 $A_p(x,y)$ (源语言,目标语言)与某个划分一致是指这个对齐点是否出现在由 $H_p(i,l_s,j,l_t(t))$ 定义的划分里。即对齐点 $A_p(x,y)$ 被认为是不一致的时候,当且仅当满足下面两个条件:

$$i \le x \le i + l_s \text{ and } (y < j \text{ or } y > j + l_t) \quad (5)$$

$$j \le y \le j + l_t \text{ and } (x < i \text{ or } y > i + l_s) \quad (6)$$

对于每个 $H_{ps_{ne}}(i,l_s,j,l_t)$,在其最大近似对齐图上都决定了一个一致对齐点的集合与一个不一致对齐点的集合。例如在图 3 中,对四元假设(2,2,2,3)来说,对齐点 $A(s_2,t_2)$,$A(s_3,t_3)$,$A(s_3,t_4)$ 为与其一致的对齐点;而对齐点 $A(s_3,t_6)$,$A(s_4,t_4)$ 则为与其不一致的对齐点。于是,我们可以用如下的公式来进行置信度估计:

$$Conf\ Est_{ma}(H_p(i,l_s,j,l_t)) = \frac{\#\ cons}{\#\ cons + \#\ incons} \quad (7)$$

公式(7)表示了四元假设 $H_p(i,l_s,j,l_t)$ 在句对 $p$ 中的最大近似对齐中的置信度估计,$\#\ cons$ 表示与四元假设一致的对齐点数,$\#\ incons$ 是与四元假设不一致的对齐点数。显然,这一个划分里,如果一致的对齐点越多,不一致的对齐点越少,那么这个四元假设的置信度就越高。

3.2.2 基于词翻译的置信度估计

在对齐模型中,最大近似对齐的置信度估计为有关命名实体的抽取提供了在一个句对中的上下文信息,而利用词的翻译概率 $p(s_i|t_{a_i})$,我们可以得到有关整个训练语料库的基于词的翻译等价对的概率估计。如果能把这个基于词对的翻译概率也作为置信度估计的一部分,那么将是对最大近似估计那种只考虑一个句对 $p$ 的局部信息的置信度估计的一种有效的平衡与补充。

基于词翻译的概率提供的信息有对于每个对齐点 $A_p(x,y)$ 的条件概率 $p(s_x|t_y)$ ($s_x$ 表示在句对 $p$ $\langle s|t\rangle$ 中源语言句子 $s$ 中的第 $x$ 个词)。所以我们用公式(8)来表示基于词翻译的置信度估计。

$$Conf\ Est_{lex}(H_p(i,l_s,j,l_t)) = \sum_{i \le x \le i+l_s} \sum_{j \le y \le j+l_t} p(s_x|t_y) \quad (8)$$

公式(8)表征了一个四元假设 $H_p(i,l_s,j,l_t)$,即一个候选的命名实体翻译等价对中源语言中的词与目标语言中的词互为翻译的概率。由于最后结果是它们相互加起来,所以这个特征倾向于给较长的命名实体翻译单位较高的置信度。

3.2.3 语言模型置信度估计

为了使与源语言中的命名实体相对应的目标语言的翻译等价单位尽可能地符合目标语言模型,我们又从目标语言的语料库中训练得到了一个语言模型 $LM(t)$,来对候选的目标语言翻译等价单位进行语言模型的置信度估计。即

$$\begin{aligned} &Conf\ Est_{lm}(H_p(i,l_s,j,l_t)) \\ &= LM(t_j t_{j+1} \cdots t_{j+l_t}) \\ &= P(t_j)\,P(t_{j+1}|t_j) \prod_{i=3}^{j+l_t} P(t_i|t_{i-1}t_{i-2}) \end{aligned} \quad (9)$$

语言模型的置信度估计的增加,使得与源语言的命名实体相对应的候选目标语言的命名实体翻译等价单位尽可能地符合目标语言的语法。同时也避免了因在基于词翻译的置信度估计偏向于获取较长的目标语言命名实体翻译单位而引入的一些多余词,因为这些多余词的存在使得它们在语言模型中的置信度估计中将获得很低的置信度。比如,当只用前两者的置信度估计时,我们获得了"特别行政区 - The Special Administrative Region the"的对应关系。但当我们加上语言模型的置信度估计时,我们就得到了准确的对应关系"特别行政区 / The Special Administrative Region"。

3.2.4 线性插值

由于训练得到的对齐模型可能含有错误的对齐点,因此,仅采用最大近似的置信度估计,结果的正确率会受到影响。这时,如果再加入基于词翻译的置信度估计和语言模型的置信度估计,最后的正确率就会有提高。基于上述的考虑,我们对以上三个特征所得到的置信度估计进行了线性插值。即给定一个四元组假设,其最后的置信度估计为:

$$\begin{aligned} &Final\ Est(H_p(i,l_s,j,l_t)) \\ &= \lambda_1 Conf\ Est_{ma}(H_p(i,l_s,j,l_t)) \times 1 \\ &\quad + \lambda_2 Conf\ Est_{lex}(H_p(i,l_s,j,l_t)) \times 10e8 \\ &\quad + \lambda_3 Conf\ Est_{lm}(H_p(i,l_s,j,l_t)) \times 10e6 \end{aligned} \quad (10)$$

其中 $0 \le \lambda_i \le 1$ 并且 $\sum_{i=1}^{3} \lambda_i = 1$。由于这三个置信度的值相差很大,所以在三个置信度前各自有一个规范化系数,使得在规范化后三个置信度的值大致在同一个数量级。初始,这些权值都设为 1/3,而后我们利用启发式搜索,最后设定的这三个参数的值分别为 $\lambda_1 = 0.153\,413$,$\lambda_2 = 0.364\,538$,$\lambda_3 = 0.482\,049$。线性插值所获得的置信度估计值综合了以上三个特征的优点,使得正确的目标语言中的命名实体翻译单位在候选的目标语言的命名实体翻译

单位中具有了相对较高的置信度估计值;这样,也使得我们的有关命名实体翻译等价单位对的抽取方法具有了较强的健壮性,即使在最大近似对齐模型只是部分正确或翻译词的概率并不很正确的情形下,最后我们也能得到较高正确率的命名实体翻译等价对。

### 3.3　命名实体翻译等价对的最后选定

对于一个在源语言中已经被识别的命名实体 $ps_{ne} = s_i s_{i+1} \cdots s_{i+l_s}$,我们选取的目标语言中的翻译等价单位 $pt_{ne} = t_j t_{j+1} \cdots s_{j+l_t}$,其中 $(j, l_t)$ 满足公式(11):

$$(j, l_t) = \arg\max_{j, l_t}(FinalEst(H_p(i, l_s, j, l_t))) \tag{11}$$

即在与源语言的命名实体 $ps_{ne}$ 相对应的所有候选目标语言翻译等价单位中,选取其置信度估计值最高的作为最后的目标语言命名实体翻译等价单位。

## 4　实验结果与讨论

我们先用 GIZA ++ 从源语言中文(未经分词,而是简单地把每个汉字当成一个词)到目标语言英文训练得到最大近似对齐结果。再把英文作为源语言,中文为目标语言,重新训练一遍,得到了如图 1 中的圆圈所示的结果。然后用 CMU-Cam Language Model Toolkit 生成一个英文的语言模型,再用中文的命名实体识别程序[8]给源语言中文进行命名实体的自动标注。最后,用本文中所提的方法进行有关命名实体翻译等价对的抽取。

表 1 我们给出了实验所用的语料库的基本情况。其中,语料库是我们自己收集整理而成的。

**表 1　实验用的语料库的基本情况**

| 语料库名 | 句对数 | 中　文 | 英　文 |
|---|---|---|---|
| 较小 | 1 500 | 66 K | 104 K |
| 较大 | 202 174 | 5.626M | 6.137M |

实验的结果如表 2 所示,NER 列表示用中文命名实体识别程序在中文(源语言)中识别得到的命名实体数目;HMM 列表示的是直接在韦特比对齐的结果上抽取[6,9]得到的翻译等价对的正确率;E1 列表示只用最大近似的置信度估计时候的正确率;E1 + E2列表示只用最大近似的置信度估计与基于词翻译的置信度估计时候的正确率;E1 + E2 + E3 列表示用最大近似的置信度估计、基于词翻译的置信度估计以及基于语言模型的置信度估计三者的线性插值时候的正确率。其中,E1 表示用最大近似的置信度估计,E2 表示基于词的置信度估计,E3 表示语言模型的置信度估计。

**表 2　用不同的抽取方法在不同语料库上所得命名实体等价翻译对的正确率**

| | NER | HMM | E1 | E1 + E2 | E1 + E2 + E3 |
|---|---|---|---|---|---|
| 较小语料库 | 369 | 72.47 % | 72.81 % | 80.78 % | **87.75 %** |
| 较大语料库 | 51 325 | 71.38 % | 72.39 % | 79.54 % | **83.46 %** |

其中,NER 所对应的列表示在源语言中用命名实体抽取工具抽取得到的命名实体单位数目。对于小语料上正确率的计算,我们先对源语言中被正确识别出的命名实体进行了人工翻译。然后把实验得到的命名实体翻译等价对与人工翻译的进行比较而统计得到的结果(如果发现源语言中命名实体识别错误,我们就简单地忽略这个翻译等价对)。对于大语料库中所得的命名实体翻译等价对的正确率是通过从 51 325 个翻译对中随机选择部分翻译对(200 个)进行统计得到的。

从实验可以看出,当我们只用最大近似的置信度估计的时候,我们发现正确率不是那么理想,主要原因可能是我们训练得到的对齐模型可能含有错误的对齐点而引起的。但当我们加入基于词翻译的置信度估计后,最后的正确率有了明显的提高;而语言模型的置信度估计的加入使得我们进一步提高了命名实体翻译等价对的正确率。

## 5　结束语

本文给出了一种从只要求对源语言进行过命名实体标注的平行语料库中抽取有关命名实体翻译等价对的方法。我们用从对齐模型上抽取短语等价对的有效技术,使得我们这种有关命名实体翻译等价对的抽取方法可以在对齐模型只有部分准确及语料库含有噪音的情况下,也能够取得高质量的命名实

体翻译等价对。实验表明我们的抽取方法取得了较好的正确率,比基于 HMM 的短语抽取方法所得的正确率有了相当明显地提高。我们下一步工作为把句对与句对间的关系、源语言的命名实体短语长度与目标语言的命名实体短语长度之间的关系等作为置信度估计的新特征加入到命名实体翻译等价对的抽取方法中来,希望得到更好的结果。

## 参考文献:


[1] D. Bikel, S. Milker, R. Schward, etc. A High-performance Learning Name-finder [C]// Proceedings of Applied Natural Language Processing, Washington DC: 1997.

[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. F, Vogel. S, and Waibel. A. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-Feature Cost Minimization [C]// Proceedings of Association of Computational Linguistics, Sapporo, Japan: 2003.

[6] 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.

[7] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation [C]// The 16th International Conference on Computational Linguistics, Copenhagen, Denmark: 1996.

[8] Zhou Jun-sheng, Dai Xin-yu, Ni Rui-yu, Chen Jia-jun. A Hybrid Approach to Chinese Word Segmentation around CRFs [C]// Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea: 2005.

[9] Ashish Venugopal, Stephan Vogel, and Alex Waibel. Effective Phrase Translation Extraction from Alignment Models [C]// Proceedings of 41st Annual Meeting of ACL, Sopporo, Japan: July, 2003.

[10] Bing Zhao, and Stephan Vogel. Word Alignment Based on Bilingual Bracketing [C]// HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Alberta, Canada: May 2003.

[11] Al'Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Prudy, Noah H. Smith and David Yarowsky. Statistical Machine Translation, Final Report [D]. Johns Hopkins University, JHU Summer Workshop, 1999.

[12] 刘冬明,赵军,杨尔弘. 汉英双语语料库中名词短语的自动对应[J]. 中文信息学报,2003,17(5): 6-12.