一种有效的基于 Web 的双语翻译对获取方法*

郭稷[1]，吕雅娟[2]，刘群[2]

[1]北京大学软件与微电子学院，北京 102600

[2]中国科学院计算技术研究所智能信息处理重点实验室，北京 100190

[1]guoji@ict.ac.cn [2]{lvyajuan, liuqun}@ict.ac.cn

摘要：命名实体和新词、术语的翻译对机器翻译、跨语言检索、自动问答等系统的性能有着重要的影响，但是这些翻译很难从现有的翻译词典中获得。本文提出了一种从中文网页中自动获取高质量双语翻译对的方法。该方法利用网页中双语翻译对的特点，使用统计判别模型，融合多种识别特征自动挖掘网站中存在的双语翻译对。实验结果表明，采用该模型构建的双语翻译词表，TOP1 的正确率达到 82.1%，TOP3 的正确率达到 94.5%。文中还提出了一种利用搜索引擎验证候选翻译的方法，经过验证，TOP1 的正确率可以提高到 84.3%。

关键字：双语翻译对获取，统计判别模型，后验证

An Effective Method to Extract Bilingual Translation Pairs from Web Corpora

Ji Guo[1], Yajuan Lv[2], Qun Liu[2]

[1] School of Software and Microelectronics, Peking University, Beijing

[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing

[1]guoji@ict.ac.cn [2]{lvyajuan, liuqun}@ict.ac.cn

Abstract: The translations of named entities, out of vocabulary words and terminologies are hard to access from traditional bilingual dictionary. This paper proposes a method to automatically extract high quality translation pairs from Chinese web corpora. We analyze the features of bilingual translation pairs in web pages, and use a statistical discriminative model combined with multiple features to extract translation pairs. Experiment results show that the quality of the extracted bilingual translation table has been greatly improved. We also propose an evaluation method after initial extraction with the help of search engines. The Top1 accuracy grows up to 84.3%.

Keywords: bilingual translation pair extraction, statistical discriminative model, post-evaluation

1 引言

随着互联网的普及和发展，互联网已经成为人们获取知识的主要来源。近几年，中文成为了世界上网页数量增长速度最快的语种。据百度数据显示，到 05 年底，中文网页总数达到约 24 亿。互联网上的中文资源越来越丰富。同时，由于国际化需要，越来越多的中文网站成为双语网站。许多网站都加入了双语甚至多语信息。互联网已经成为获取双语或多语翻译资源的巨大来源。

双语翻译词典是重要的翻译资源。由于易实现和翻译词典的可读性，基于词典的方法在很多机器翻译应用，如跨语言检索中仍被广泛采用。但是传统的双语词典通常不包含新词术语以及人名、地名等命名实体的翻译，而这些词的翻译对于机器翻译、跨语言检索、自动问答等系统的性能有着重要的影响。利用互联网丰富的资源，研究大规模、高质量的双语翻译对自动获取方法，

*本课题得到自然科学基金项目的支持，项目批准号 60603095。

已经成为目前的研究热点。前人在双语翻译资源获取方面做了很多尝试。搜索引擎、双语平行语料库和中文网页是获取双语翻译资源的主要来源。本文研究了一种有效的从中文网页中获取高质量双语翻译对的方法。该方法利用网页中双语翻译对的特点，使用统计判别模型，融合多种识别特征自动挖掘中文网页中存在的双语翻译对。实验证明，使用该方法可以获得高质量双语翻译对。

2 相关工作

在获取双语翻译知识方面已经存在一些研究工作。

Zhang[2]，Huang[3]提出利用搜索引擎的返回结果来获取双语翻译知识，他们使用不同的方法构造查询词交给搜索引擎，在返回结果中，利用统计方法获得对应翻译。他们的方法可以获得较好的翻译，但是由于搜索引擎的限制，这种方法不易用于获取大规模双语翻译资源。双语平行语料库已被用于构建大规模双语翻译词典。Huang[4][5]从句子对齐的双语语料库中训练双语命名实体之间的多特征的统计对齐模型，然后利用统计对齐模型进行双语翻译对的抽取。实验证明，他们的方法效果令人满意，然而高质量的双语平行语料库不太容易获取。张永臣[1]利用词间关系矩阵法从特定领域非平行语料中抽取双语词典。其中种子词的选择对抽取结果影响较大，抽取出来的双语词典的质量一般。

Zhang[6]在研究过程中发现，在中文网页中，如果英文出现在括号中，那么周围的中文很可能是其对应的翻译。她将出现在括号中的英文前面的中文分为两种情况：一种是前面的中文出现在书名号或者引号当中，例如，“东亚奇迹”（East Asia Miracle），《银行保密法》（Bank Secrecy Act），「独立公投」（independence referendum）；另一种是前面的中文不出现在书名号或者引号中，如据考克斯新闻社（Cox news service）。具有这样特征的中文网页是一个获取大量双语翻译对的潜在来源。Cao[7]在大规模的中文网页上做了相应研究。他们训练一个音译对齐判别模型用于音译对的抽取，然后训练一个翻译判别模型用于翻译对的抽取。然而网页内容的复杂性影响了音译判别模型的效果，例如博格斯（Tom Burgis），前面的中文往往只是英文中一个单词的音译。确定用于音译对齐判别的中文和英文，不仅繁琐而且容易出错。此外，日文名也不能用于音译对齐判别。实验结果显示，他们抽取出来的音译对和翻译对正确率较低，质量不能令人满意。

本文与 Cao[7]的研究相似，希望能够在中文网页中抽取双语翻译对。与 Cao[7]的工作不同的是，本文采用统计判别模型 Perceptron 对候选翻译进行训练和识别，其优点是可以有效地融合多种特征。实验证明本文方法有效提高了双语翻译对抽取的正确率。

3 双语翻译对获取

3.1 术语定义

本节开始将要使用的术语有：

候选行，是指中文网页中的一行中文，其中有英文出现在括号中。

固定格式翻译对，是指在候选行中，当英文出现在括号中时，其前面的中文出现在书名号或者引号中的中英文翻译对。

候选翻译单元，是指候选行中抽取出来的不包含非法翻译字符（如。，！等）的中文文字。

候选翻译对，是指从候选翻译单元生成的用于翻译判别的中英文。

下面是一个简单的例子。在候选行 那就是他要承担责任。这也正是丹尼尔·布东（Daniel Bouton）中，候选翻译单元是 这也正是丹尼尔·布东（Daniel Bouton），可能产生的一个候选翻

译对是 布东（Daniel Bouton）。

3.2 候选翻译对生成

从一个候选翻译单元中抽取正确的双语翻译对，可以归结为中文边界划分问题。因为英文已经出现在括号中，要找到正确的翻译对，只需在英文前面的中文中划分出正确的边界，边界之内的中文就认定为英文的翻译。为了找到正确的边界，我们使用中文分词工具对候选翻译单元的中文进行切分，然后组合切分得到的词构成候选翻译对。例如，经过切分后的候选翻译单元是足球 / 教练 / 佐 / 夫（Zolf），那么可以构成下面四个候选翻译对：(1)夫 Zolf (2)佐夫 Zolf (3)教练佐夫 Zolf (4)足球教练佐夫 Zolf。可以看出，一个候选翻译单元可以生成多个候选翻译对。这样，我们就把中文边界划分问题转化为从多个候选翻译对中选择正确的翻译对问题。当前中文分词系统具有很高的精度，因此，候选翻译对中基本上会包含正确的翻译。

3.3 翻译判别模型

翻译判别模型是一个基于多特征的判别式模型。设 S 是 N 组候选翻译对的集合。$s_{i,j}$ 表示第 i 组第 j 个候翻译对，其特征表示为 $f_k(s_{i,j})$。$s_{i,j}$ 的得分如公式(1)所示：

$$Score(s_{i,j}) = \sum_{k=1}^{K} f_k(s_{i,j}) \times \lambda_k \qquad (1)$$

λ_k 是 $f_k(s_{i,j})$ 对应的权值。K 是总的特征个数。翻译判别模型计算各组中每个候选翻译对的得分，然后以得分最高的候选翻译对作为翻译对抽取的结果。

3.4 特征选择

我们对候选翻译单元的中文部分进行分词、词性标注和命名实体识别，并选取了以下特征：

1. 候选翻译共现频率。在生成候选翻译对时，我们将具有相同英文翻译的中文放在一起统计。某候选翻译和英文的共现频率越高，它越可能成文该英文的翻译。

2. 候选翻译的长度。候选翻译的长度是指候选翻译包含的汉字个数，长度过长或过短，其成为英文翻译的可能性就越小。

3. 是否是命名实体。如果某个英文是一个命名实体，那么候选翻译中的命名实体就很可能成为其翻译。

4. 是否包含“·”。在外国人名全称的翻译中，往往会包含·符号。这是外国人名中姓和名的分隔标志。如果候选翻译中包含这个符号，该候选翻译有可能包含了外国人名的全称翻译。这个特征可以保证外国人名全称的翻译不会丢失。

5. 候选翻译首词的词性。以名词、形容词等开头的候选翻译成为对应英文翻译的可能性比以介词、连词等开头的候选翻译大。

6. 候选翻译前一个词的词性。在中文里面，尤其是中文网页中，对于特定的词性而言，如介词、连词、助词等，其后面的中文成为相应英文的翻译的概率较大。

7. 候选翻译前一个词。在中文网页中，有些词语带有明显的暗示信息，其后面的中文是对应的英文的翻译。比如基地组织高级指挥官利比（Abu Laith al-Libi），已经与英国航空（British Airlines），其中，“指挥官”、 “与”就带有很好的暗示信息。

3.5 模型训练

我们利用感知机来训练翻译判别模型。

对于频率特征，我们将其离散化，转换成二值特征。将翻译实例出现的频率分为 7 个等级：

等级 1 是出现 1 至 2 次，等级 2 是出现 3 至 5 次，等级 3 是出现 6 至 8 次，以此类推，最后等级 7 是出现 18 次以上。这样，频率的特征函数如公式(2)所示。

$$f_k(x)=\begin{cases}1 & \text{if x 的频率等级是1}\\ 0 & \text{else}\end{cases} \qquad (2)$$

对于长度特征，采用类似的离散化方法，将其转换成二值特征。这样，整个模型使用的全是二值特征。

由 3.2 节的叙述可以知道，一个候选翻译单元会产生多个候选翻译对。由于我们将具有相同英文的候选中文翻译放在一起统计，其产生的候选翻译对的数目可能会很大。而在众多的候选翻译对中，只有一个正确翻译对，因此训练数据是倾斜的，不利于感知机的训练。为了克服训练数据倾斜问题，我们将训练过程看成一个类似于重排序的过程。首先将具有相同英文的训练实例划为一组。在每轮训练中，对于每组训练实例，计算其中每个训练实例的得分，然后选出得分最高的实例。如果得分最高的训练实例是正例（即正确翻译对），则继续进行下一组实例训练，不调整参数；如果是负例，则调整权值参数。整个训练过程迭代进行，直到满足收敛条件。训练过程如图 1 所示：

输入： 训练实例集 S_i (i=1...N)
输出： 权值向量 $\vec{\lambda}$
1. 初始化权值 $\lambda^1_k = 1$ $(k = 1,2...K)$
2. for i = 1 to N
3. 计算第i组每个训练实例得分 $Score(s_{i,j})$
4. $q = \arg\max_j Score(s_{i,j})$
5. if $s_{i,q}$ 是负例
设 $s_{i,p}$ 是正例
6. $\lambda_k^{t+1} = \lambda_k^t + \eta(f_k(s_{i,p}) - f_k(s_{i,q}))$ $k = 1,2...K$ $\eta = 0.001$
7. 重复步骤2直至收敛

图 1 翻译判别模型训练

4 实验

4.1 数据准备

实验使用的网页数据分为训练网页数据和测试网页数据。训练网页数据来自《联合早报》、《欧洲时报》、《华盛顿观察报》三个网站。测试网页数据分为两部分。一部分和训练网页数据一样，来自于《联合早报》网站。为了测试翻译判别模型的性能，另一部分来自于与训练网页数据来源不同的“星岛环球网”。表 1 显示了实验的网页数据情况。

	来源	大小
训练网页	《联合早报》《欧洲时报》《华盛顿观察报》	1.6 G
测试网页	《联合早报》	835 M
	“星岛环球网”	858 M

表 1 实验网页数据

4.2 预处理

对于训练网页数据，首先抽取候选行，然后从中获得固定格式翻译对，接着进行候选翻译单元抽取。在生成候选翻译对前，我们使用 ICTCLAS 对候选翻译单元的中文进行分词、词性标注和命名实体识别。最后生成候选翻译实例。对测试数据采用同样的处理方式。

4.3 翻译判别模型训练和测试

训练实例集包含了人工标注的 29,953 个训练实例。按照图 1 所示的训练方法进行模型训练。在每轮迭代训练之前，我们随机产生 N 组训练实例的训练顺序，这样可以避免训练数据过拟合问题。实验中，迭代收敛条件是前后两次训练的正确率的差值小于阈值 0.0001。

开发实例集一共包含 19,661 个实例。由于在训练过程中，每轮迭代时训练实例的顺序都不同，所以每次训练产生的模型也不同。开发集就是用于从训练得到的多个模型中选取最佳的训练模型。另外，实验的目的是从中文网页中抽取双语翻译对。对抽取出来的双语翻译对质量的评价客观上表现了翻译判别模型的性能，所以实验中不使用测试实例集。

表 2 显示了不同特征集合的训练结果。我们发现，在候选翻译识别中，首词词性特征(FT)是一个重要的特征，去掉首词词性，正确率下降了 6.1 个点。长度特征(LEN)也起到了重要的作用。去掉长度特征，正确率下降了 4.1 个点。接下来，去掉前一个词的特征(PW)，正确率下降了 3.2 个点。去掉是否是命名实体的特征(NE)带来了 3 个点的正确率下降。频率特征(FQ)起到了一定的作用，去掉该特征，带来了 2.7 个点的正确率下降。前一个词词性(PT)的特征所起的作用较小，去掉该特征使得正确率下降了 1.5 个点。最后，是否包含·的特征(DOT)起到的作用最小，去掉该特征，正确率只下降了 0.3 个点。我们认为，该特征的主要作用在于获取外国名字的全称翻译，在候选翻译对方面的判别能力不如其他特征。

特征	开发集正确率(%)	特征	开发集正确率(%)
All	76.0	**All - DOT**	75.7
All - FQ	73.3	**All - FT**	69.9
All - LEN	71.9	**All - PT**	74.5
All - NE	73.0	**All - PW**	72.8

表 2 不同特征集合的训练结果

4.4 翻译对抽取

利用上面训练得到的翻译判别模型，我们在测试网页数据上进行了两组实验，测试抽取的翻译对的正确率。

第一组实验的网页数据的来源和训练数据相同，来自于《联合早报》网站。经过预处理，得到了 1,181 个固定格式翻译对。经验证，准确率为 98.4%。使用翻译判别模型，我们从剩下候选翻译对中抽取了 5,118 个翻译对。随机选择了 600 个翻译对进行人工验证，实验结果如表 3 所示。

为了验证翻译判别模型的健壮性，在第二组实验中，我们使用了与训练网页数据来源完全不同的网页数据。该组实验的网页来自于“星岛环球网”。预处理后，得到 666 个固定格式翻译对，经验证，准确率为 96.9%。使用翻译判别模型，我们从剩下的候选翻译对中共抽取了 4,267 个翻译对。随机选择 600 个翻译对进行人工验证，实验结果如表 4 所示。

表 3 和表 4 中，翻译对正确率是指从候选翻译对中抽取出来的翻译对的正确率。对于多译的英文，如果抽取出来的中文是其中的一个翻译，就认为该翻译对正确。TOPN 是指每个翻译对

取得分最高的前 N 个的中文翻译作为翻译抽取结果。在表 3 中，TOP1 的翻译对正确率达到了 78.5%。TOP3 的翻译对正确率达到了 93.7%。表 4 中 TOP1 的翻译对正确率达到 79.8%，TOP3 的翻译对正确率达到了 89.7%。翻译对的质量令人满意。两组不同的实验表明在不同来源的中文网页中我们的方法可以抽取高质量的双语翻译对。附录 A 是我们获取的双语翻译对的一些实例。

	翻译对正确率（%）
TOP 1	78.5
TOP 2	89.7
TOP 3	93.7

表 3 《联合早报》网站翻译对质量

	翻译对正确率（%）
TOP 1	79.8
TOP 2	86.7
TOP 3	89.7

表 4 “星岛环球网”网站翻译对质

另外，抽取的双语翻译词表的正确率是指加上固定格式翻译对后总的翻译对的正确率。第一组实验中，TOP1 的总正确率为 82.1%，TOP3 的总正确率为 94.5%。第二组实验中，TOP1 的总正确率是 82.1%，TOP3 的总正确率是 90.6%。

5 翻译对后验证

从表 3 和表 4 可以看到，TOP3 的翻译对正确率比 TOP1 的正确率高出很多。这意味着，使用第 3 节介绍的翻译判别模型从中文网页中抽取翻译对，绝大部分的正确结果都在得分最高的前 3 个候选翻译中。相对而言，TOP1 的正确率比较低。那么对翻译判别模型抽取出来的候选翻译结果进行验证，选出正确的结果，提高 TOP1 的正确率，会是非常有意义的事情。本节介绍一种借助搜索引擎进行后验证的方法，其思路来源于跨语言检索领域中解决 OOV 翻译问题的方法[2]。

我们使用第一组实验中随机选择的 600 个翻译对作为初始抽取结果，其中每个英文包含得分最高的 3 个中文翻译。后验证的过程就是利用搜索引擎的返回结果，从这 3 个候选中文翻译中选择正确的翻译。验证过程如图 2 所示。

输入： 600组实验数据，每组包含英文及其3个中文候选翻译
输出： 600组实验数据，每组包含英文及其中文翻译

1. 对于每组实验数据T，c_{ij}表示第j个候选翻译 （i=1...600， j=1, 2, 3）
2. for j=1 to 3
3. 使用$c_{i,j}$和英文e构成查询词，交给搜索引擎，取返回的前100个结果
4. 统计$c_{i,j}$在返回结果中的出现频率$f_{c_{i,j}}$
6. $S(c_{i,j}) = score(c_{i,j}) + \beta \cdot \mu \cdot f_{c_{i,j}}$
7. $l = \arg\max_j S(c_{i,j})$, 取$c_{i,l}$为英文e的中文翻译

图 2 后验证过程图

图 2 中，$score(c_{ij})$表示候选翻译 c_{ij} 在翻译判别模型下的得分。μ 是出现频率的调整因子，目的是将出现频率的值调整到合适的数量级。实验中取 μ =0.001。β 是出现频率的加权系数，它的值使用类似于最小错误率训练的方法进行调整。实验中 β =0.4。经过后验证，对这 600 个翻译对重新进行人工验证，TOP1 的翻译对正确率为 81.7%。可见，后验证提高了 TOP1 的正确率。

6 未来工作展望

未来的工作可从以下两个方面开展：

1、在 3.2 节中，我们提到，从候选翻译单元中抽取正确的翻译对，就是中文边界划分问题。本文采用的方法是将英文前面的中文进行分词，然后逐个组合，形成多个候选翻译对，再从候选翻译对中选出正确的翻译对。然而，中文分词的错误，会造成候选翻译对中不包含正确的候选翻译对。例如 传奇/魔/幻/师大/卫/考/柏/菲(David Copperfield)，中文分词的错误造成不能抽取正确翻译对。研究不使用中文分词的边界划分方法，将留给下一步的工作。

2、本文对利用搜索引擎进行后验证的方法进行了初步尝试，实验证明，这种方法能够提高双语翻译对抽取的正确率。但是后验证方法的效果还有待改进。搜索引擎返回结果的处理，包括词频统计，噪声去除问题等，将对后验证方法效果产生重要的影响。

7 结论

本文提出了从中文网页中抽取双语翻译对的有效方法。我们目前只处理有英文出现在括号中的中文网页，暂不考虑其他情况。首先，对具有上述特点的中文网页进行预处理，然后对候选翻译单元的中文进行分词、词性标注和命名实体别，最后得到候选翻译对。我们训练一个翻译判别模型，并使用这个模型从候选翻译对中抽取翻译对。实验结果表明，翻译判别模型对于不同来源的中文网页是健壮的。抽取的翻译对正确率高，质量令人满意。此外，我们还提出了利用搜索引擎进行翻译对后验证的方法。经过后验证，抽取的翻译对的正确率得到了进一步的提高。

参 考 文 献

[1] 张永臣, 孙乐等. 基于 Web 数据的特定领域双语词典抽取[J]. 中文信息学报, 2006 年第 2 期: 16-23.

[2] Y. Zhang and P. Vines. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In the Proceedings of SIGIR 2004.

[3] F. Huang, Y. Zhang and S. Vogel. Mining Key Phrase Translations from Web Corpora. HLT-EMNLP 2005.

[4] F. Huang, S. Vogel and A. Waibel. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In the Proceedings of ACL 2003.

[5] F. Huang and S. Vogel. Improved Named Entity Translation and Bilingual Named Entity Extraction. In the Proceedings of ICMI 2002.

[6] Y. Zhang and P. Vines. Detection and Translation of OOV Terms Prior to Query Time. In the Proceedings of SIGIR 2004.

[7] G. H. Cao, J. F. Gao and J. Y. Nie. A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web Pages. In MT Summit XI, pp. 57-64.

附录 A 获取的双语翻译对实例

类型	中文	英文	类型	中文	英文
人名	阿当·阿济兹	Aadam Aziz	地名	罗尼湾	Rodney Bay
	中川昭一	Shoichi Nakagawa		阿巴拉契亚中心山脉	Central Appalachian
组织机构名	统一俄罗斯党	United Russia Party	新词术语	新加坡式英语	Singlish
电影、报刊、书籍	《得克萨斯城太阳报》	Texas City Sun		就业入息补助	Workfare Income Supplement
	《爱上大姐大》	Marrying the Mafia		血管成形术	angioplasty
	《广角镜》	Panorama		"菠菜人"	Spinach Man