文章编号：1003-0077(2008)05-0067-07

基于 Web 的双语平行句对自动获取

叶莎妮，吕雅娟，黄 赟，刘 群

（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）

摘 要：双语平行句对是机器翻译的重要资源，但是由于获取途径的限制，句子级平行语料库不仅数量有限而且经常集中在特定领域，很难适应真实应用的需求。该文介绍了一个基于 Web 的双语平行句对自动获取系统。该系统融合了现有系统的优点，对其中的关键技术进行了改进。文中提出了一种自动发现双语网站中 URL 命名规律的方法，改进了双语平行句对抽取技术。实验结果表明文中所提出的方法大大提高了候选双语网站发现的召回率，所获取双语平行句对的召回率为 93%，准确率为 96%，证明了该文方法的有效性。此外，该文还对存在于双语对照网页内部的双语平行句对的抽取方法进行了研究，取得了初步成果。

关键词：计算机应用；中文信息处理；双语句对；平行网页；网页挖掘

中图分类号：TP391　　**文献标识码**：A

Automatic Parallel Sentences Extraction from Web

YE Sha-ni, LV Ya-juan, HUANG Yun, LIU Qun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Parallel sentences are valuable resources for machine translation while not readily available in the necessary quantities and often domain limited. This paper constructs a system to automatically obtain parallel sentences of high quality from the Web. This system puts forward a method to find the similarity of URLs in bilingual websites, and also improves parallel sentence extraction technology. Experimental results show that this system gains a recall rate of 93% and a precision rate of 96% when collecting parallel sentences from test set. In addition, this paper makes preliminary research in collecting parallel sentences from bilingual contrast web pages.

Key words: computer application; Chinese information processing; bilingual sentences; parallel corpora; web mining

1 引言

近年来，语料库资源对于自然语言处理研究的巨大价值已经得到越来越多的认可。特别是双语语料库(Bilingual Corpus)，已经成为机器翻译、机器辅助翻译以及翻译知识获取研究不可或缺的重要资源。一方面，双语语料库的出现直接推动了机器翻译新技术的发展，基于统计(Statistic-Based)和基于实例(Example-Based)等基于语料库的翻译方法为机器翻译研究提供了新的思路，有效改善了翻译质量，在机器翻译研究领域掀起了新的高潮。另一方面，双语语料库又是获取翻译知识的重要来源，从中可以挖掘学习各种细粒度的翻译知识，如翻译词典和翻译模板，从而改进传统的机器翻译技术。此外，双语语料库也是跨语言信息检索，翻译词典编纂、双语术语自动提取[1]以及多语言对比研究等的重要基础资源。

然而，大规模双语语料库建设与获取存在着很大的困难。虽然各国都投入了大量的人力、物力和

收稿日期：2008-01-21　**定稿日期**：2008-05-21

基金项目：国家自然科学基金资助项目(60603095,60573188)

作者简介：叶莎妮(1983—)，女，硕士生，主要研究方向为自然语言处理技术；吕雅娟(1972—)，女，博士，副研，主要研究方向为自然语言处理技术；黄赟(1983—)，男，硕士生，主要研究方向为自然语言处理技术。

财力来加强双语语料库的建设，但是现有双语语料库在规模、时效性和领域平衡性等方面还不能满足处理真实文本的需要。

互联网的普及和迅猛发展提供了大量而丰富的电子信息。随着国际化的需要，越来越多的网站成为双语网站，越来越多的网上信息以多语言的方式发布，这就为双语和多语语料库提供了很好的来源。互联网是一个取之不尽、日益增长的信息源，因此是一个潜在的巨大的多语种语料库信息源。这为双语平行语料库的获取提供了潜在的解决途径。研究如何从互联网上自动挖掘这些海量的、真实的双语句对，对于解决双语语料库获取难题，推动相关技术发展和实用化具有重要的意义。

本文的目标就是建设一个高效的、自动化的双语句对获取系统，主要集中在获取中英平行语料。但是除了一些与具体语言相关的配置文件以外，本文采用的方法不依赖具体语言，可以很轻松地移植到其他的语言对上。

2 相关工作

现有系统在基于 Web 获取双语语料库资源时主要分为四个步骤：双语候选网站的获取及过滤，双语候选网页的获取，双语候选网页的过滤，双语平行句对的抽取。著名的系统有 PT Miner[2]，PTI[3]，BITS[4]，STRAND[5,6]，WPDE[7]等。

双语候选网站的获取及过滤 候选网站为可能含有双语平行网页的网站，如果一个中文网页中含有以“English”、“English Version”等为锚文本或图片 ALT 信息的链接，或者一个英文网页中含有相应的以“中文”、“中文版”等为锚文本或图片 ATL 信息的链接，则可以认为含有该网页的网站是一个中英候选网站。微软的 WPDE 系统首次提出除了锚文本之外，还可以利用图片的 ALT 信息来搜索双语网站。

双语候选网页的获取 从一个双语网站中获取双语候选网页，可以利用结构与内容两个方面的特征来实现。结构上一般可以利用网页作者在双语网页 URL 命名时的特点。如下例所示的两个中英平行网页的 URL 中只有语言相关的字符串“zh”与“en”以及“c”与“e”不同。

例 1 一对具有命名相似性的中英网页的 URL：

```
www.fao.org/newsroom/zh/field/2005/index_c.html
|                    | |       |  ||   |
www.fao.org/newsroom/en/field/2005/index_e.html
〈----------pathname----------〉〈--basename-〉
```

内容上的特点是显而易见的，如果内容上存在互为翻译关系，那么就很可能是一对平行网页。PT Miner、STRAND、WPDE 等系统都利用了前者，其中 WPDE 系统发现 URL 的 pathname 与 basename 存在不同的命名相似性，需要分别进行处理。BITS 系统只利用后者。而 PTI 则同时利用了这两部分信息，先用一个基 URL 名相似性的抽取器找出一部分双语候选网页，再用基于内容的抽取器对剩下的双语网页进行处理。

双语候选网页的过滤 在取得双语候选网页之后，采用分类器过滤掉伪平行的双语网页，就得到了真正平行的双语网页。STRAND 系统在这个环节采用的一些基于结构和内容的特征，基本上都被其他系统所采用。之后出现的系统不断采用更加合适的基于网页内容的特征与分类器，都取得了很好的效果，达到了预期的目的。

双语平行句对的抽取 最后需要从双语平行网页中抽取出双语平行句对，得到句子级双语语料库，才能满足真实应用的需求。STRAND 中采用双语平行网页之间 html 结构的相似性，对两个网页的 html 标记序列进行对齐[6]，夹在两对对齐的 html 标记之间的句子就够成一个双语平行句对。类似的方法如用 Dom Tree[8] 来描述网页，然后对两棵 Dom Tree 进行对齐从而得到相应的平行句对。

尽管现有系统在基于 Web 获取双语语料时都取得了不错的效果，但仍存有以下不足。首先，所有系统在利用 URL 命名相似性得到双语候选网页都需要依赖预定义的字符串集合。其次，由于网页资源的噪声很大，仅仅依赖双语平行网页 html 结构相似性来获取双语平行句对并不能取得一个理想的效果。最后，双语文本资源来源主要集中于中英平行网页对，但是有相当一部分高质量的中英平行文本存在于双语对照的一个网页中。本文集中解决上述的三个问题，而本文实现的系统中双语候选网站获取与双语候选页过滤部分都类似 WPDE 中的方法，同时融合了已有系统的优点，在此就不再赘述。

3 双语网页 URL 命名模板的自动发现

现有系统中预定义的字符串集合大部分是与特

定语种相关，并且在大量双语网站的URL中出现。中文的有“cn，chinese”等，而英文相应的有“en，english”等。但是，同时存在着大量的双语网站，其URL命名虽然具有语言相似性，但不是通过预先定义可以发现这种规律，例如某网站中英平行网页URL的对应关系为zh-cn.broadcom.com→www.broadcom.com，再比如某网站中存在的URL对应关系则为www.wcbn.com.sg→english.wcbn.com.sg。此外，也存在一些网站在命名时，中文一侧的URL用的是完整的单词，而英文一侧对应的网页URL用的则是该单词的简写。例如某中英平行网页的basename之间的对应关系为cartoon_list.html与cart_list.html。

可见，只要URL命名的相似性没有涵盖在预定义的集合中，那么已有的系统就无法处理。本文提出以下这种方法，能够自动发现当前网站在双语网页命名时具有的特点，再进行候选双语平行网页对的获取，完全不需要预定义与语言相关的字符串集合。

3.1　自动发现双语网页URL命名模板

从例1可以看出pathname与basename中语言相关的不同字符串出现位置有一定的相似性，我们将pathname与basename统称为name域。name域都由语言无关的相同部分S与语言相关的不同部分Lang组成。例1中，中英URL的pathname对应的S有“www.fao.org/newsroom/”与“/field/2005/”。Lang部分分别为“zh”与“en”（记为Lang_c与Lang_e），这正是我们关心的中英平行网页之间的命名规律。那么当一个网站中相当一部分中英网页的URL满足如下形式：

$curl=S_1+Lang\_c_1+S_2+Lang\_c_2+\cdots+S_j+Lang\_c_j+S_{j+1}$

$eurl=S_1+Lang\_e_1+S_2+Lang\_e_2+\cdots+S_j+Lang\_e_j+S_{j+1}$

那么我们就有理由相信上述形式中的各个Lang_c与Lang_e就是当前网站中英URL命名时的规律性。S与Lang都由字段组成，其中只有S_1与S_{j+1}可以为空，每一字段看作一个整体，具有操作完整性原则即无论进行任何操作都对整个字段进行。在pathname中，按照/来划分，每一个/str/称为一个字段，而在basename中则按照“_.-”等分隔符分成各个字段。

那么，当存在一些中英网页对的name符合上述形式时，就可以得到我们要抽取的双语网页URL命名模板（具体见定义）。

URL命名模板定义　$Lang\_c_1 \rightarrow Lang\_e_1$，$Lang\_c_2 \rightarrow Lang\_e_2 \ldots Lang\_c_i \rightarrow Lang\_e_i$

其中$Lang_c_i$/ $Lang_e_i$代表中文/英文网页URL的name域中语言相关的不同部分，两者不可同时为空。从例1这对中英网页的URL中可以得到一条pathname的命名模板“zh→en”，一条basename命名模板“c→e”。本文自动发现的URL命名模板可以完全包含已有系统中预定义的URL命名相似性。

每一个模板在应用时都对应着具体的操作动作，比如有一个pathname命名模板“zh→en”，就意味着当存在一个中文URL的pathname中含有“zh”时，“zh”替换为“en”，构造了一个英文URL的pathname，如果这个pathname在英文URL的pathname列表中存在，那么就得到了一对命名具有相似性的中英URL的pathname部分。再配合basename命名模板，就可以一对URL命名有相似性的中英候选网页。根据Lang_c，Lang_e内容不同，模板的动作也有不同的含义详见表1。

表1　应用模板的四种动作

Lang_c	Lang_e	Action
非空	非空	替换操作即用Lang_e替换Lang_c
非空	为空	在中文URL中删除对应的Lang_c
为空	非空	在英文URL中删除对应的Lang_e
为空	为空	说明要查找相同的中文URL，英文URL

在本系统中，首先进行网页语种识别，将一个网站中的所有网页分为中文和英文网页，分别得到中文与英文网页的URL列表。在这两个列表中去自动发现pathname与basename命名相似性，生成相应的命名模板，再进行查找双语候选网页对。生成pathname与basename对应的模板集合算法相同，具体见算法1。

在生成pathname命名模板时，CnameSet与EnameSet是中英URL去掉basename所形成的中英pathname集合。而生成basename命名模板时相应的集合则是pathname具有相似性的所有中英URL所对应的basename形成的集合。

算法1 模板发现算法

```
Require: the pathname or basename vector in chinese and english sides.
Ensure: the map of rules mRule.
1: for cname in CnameSet do
2:     flag←0
3:     for rule in RuleSet do
4:         ename←Apply Rule(rule, cname)
5:         if ename∈ vEname then
6:             AddWeight(rule)
7:             flag←1
8:             break
9:         end if
10:    end for
11:    if (flag==0) then
12:        for ename in vEname do
13:            newrule←CreatNewRule(cname, ename)
14:            Insert newrule to RuleSet
15:        end for
16:    end if
17:    Sort RuleSet by Weight
18:    flag←0
19: end for
```

3.2 获取具有URL命名相似性的双语候选网页

首先根据pathname模板寻找pathname具有相似性的中英URL，再根据basename模板寻找同时还具有basename相似性的中英URL，至此就得到一对可信的双语候选网页。为了自动发现的模板更加准确，限定pathname/basename模板中只要任何一个Lang_c或Lang_e为纯数字字符串，则认为当前模板无效，并不是我们要找的与特定语言相关的字符串。

采取本算法的优点在于可以根据每个网站自身的特点来处理，而不是像预定义那样只能处理URL命名符合特定规律的网站。本文的方法不仅可以发现所有常见的URL命名规律，而且还可以找出不同的网页编辑者带个人特色的URL命名规律，从而可以找出尽可能多的可信的中英候选网页对。

4 双语平行句对的抽取

在实际存在的双语网页中，很少一部分双语平行网页是完全直译的。网页作者很难保证两个双语平行网页之间句子数目相同并且处于对等位置的句子互为翻译。更何况网页资源本身噪声很大，双语平行网页的html结构有相似性但不相同。所以现有系统仅仅依赖平行网页的html结构来挖掘双语平行句对是不可靠的。

本文中把双语网页html结构上的相似性作为一个有力的特征，更加侧重从双语句对的内容上去衡量一对双语平行网页中那些句对是真正互翻译的。这样不仅可以有效地利用平行网页html结构上存在的相似性，而且保证得到的双语句对在内容上确实是互为翻译的。获取互为翻译的双语句对，可以结合Daniel Marcu在不平行文本中抽取平行句对所采用的方法[9]。从两个中英平行网页中得到的中英句子序列进行对齐可以看做一个中英候选句对集合的分类问题，从中英候选句对中抽取一些特征，送入分类器进行分类，判断当前的中英候选句对是否是真正平行的。

采用分类的方法将大大屏蔽由于平行网页内容不完全一致与html结构混乱而带来的抽取句子级平行文本的难度，并且可以最大限度的挖掘出平行网页中的平行句对资源。

4.1 生成双语候选句对集合

先将每一个网页解析成一个由“html开始标记+句子+html结束标记”这种结构组成的句子序列(见图1)。图1中的“句子”实际上就是网页中的一段连续文本，可能是为短语、句子、段落，我们在这里不做仔细地区分。在本环节的最后针对过长的双语句对(即中英句子都是包含多个句子的段落)利用现有的技术进行句子划分并对齐，这里就不再详细叙述。

```
<span>联系我们！ </span>
<h3>学术课程</h3>
<strong>新生：</strong>
<strong>转学学分高于或等于60学分的转学生：</strong>
<strong>英语语言能力：</strong>
<h3>美国语言和文化协会(alc1)</h3>
<strong>强化升学英语预备课程</strong>
<a>联系我们</a>
<td>calfornia state university san marcos是一所发展迅速的现代化学校，
拥有大量新式建筑、尖端的计算机技术和先进的科学实验室。
在这里，您可以感受到时代发展的豪情。</td>
```

图1 带html首尾标记的句子序列

对中文和英文网页进行解析得到图1所示的中文和英文句子序列后进行简单的全组合就可以得到中英候选句对集合。这样得到的中英候选句对集合中含有大量明显不平行的句对，可以采用一些过滤机制，过滤掉明显不平行的句对以节省下一步抽取平行句对的时间。如图1中“〈h3〉学术课程〈/h3〉”在这个中文句子序列中的相对位置为2/9，那么在相应的英文句子序列中，相对位置也为2/9左右的英文句子很可能就与该句子构成平行句对。经过统计可以发现平行的中文句子和英文句子在相应的网页中所处的相对位置比例大量集中在1附近，少数分布在0.5…0.8，1.5…2.5。在本系统中可以用这

个特征过滤掉候选句对中明显不平行的句对。同样的，还可以利用中英句子的词数比例等特征。

4.2 抽取双语平行句对

面对一对简短的中英句子，判断两者是否互为翻译，就变得相对简单。可以利用三组特征：一般文本特征、词语对齐相关特征以及网页结构特征。一般性的特征见表2。

表2 一般文本特征

一般文本特征	特征描述
句子长度比特征	中英句子中的词数之比
词汇化汉译英比例	中文句子在对应的英文句子中有对应翻译的词所占的比例
词汇化英译汉比例	英文句子在对应的中文句子中有对应翻译的词所占的比例

有一些中英句子描述的是同一件事情，虽然从内容上来说词汇化汉译英比例（或英译汉）都比较高，但不是互为翻译的关系。所以仅仅采用双语句对的一般性特征并不能充分刻画互为翻译这个目标。在这里先对每个双语候选句对进行词语对齐，再挖掘出一系列的词语对齐特征来进一步描述互为翻译这个分类目标。词语对齐相关特征详见表3。

表3 词语对齐相关特征

词语对齐特征	特征描述
一对一链接比例	一对一链接数占词语对齐中总链接数的比例
中文词对空比例	中文句子中没有相应对齐的词语所占比例
英文词对空比例	英文句子中没有相应对齐的词语所占比例

针对网页这类特殊的文本，还可以利用网页的html结构信息，来帮助分类过程。首先，对于两个中英网页，得到了图1所示的带html首尾标记的句子序列以后，可以发现在两个中英句子序列中出现在同等位置上的中英句子为平行的可能比较大。其次，两个中英平行的句子其html开始标记（或结束标记）大部分也是相同的。最后，可以用Unix工具diff对两个网页的html标记序列进行对齐，那么也就会得到一个相应的初步句子对齐（具体方法参见文献[6]）如果两个中英网页是平行的并且近乎直译的情况，那么得到初步句子对齐结果里面很多句对应该是平行的。针对这三种情况，我们定义了三个特征，详见表4。

表4 网页结构特征

网页结构特征	特征描述
html对齐特征	当前句对是否在存在于根据html标记对齐得到的句对集合中
html标记特征	中文句子和英文句子在相应的原始网页中紧跟的html开始（或结束）标记是否相同
中英句子相对位置比例	中文句子和英文句子在相应的网页中所处相对位置的比例

例如，一个中英候选句对的相关信息如表5所示。这个候选句对没有出现在初步句子对齐得到平行句对集合中。那么html对齐特征就为false；html标记特征为false；中英句子相对位置比例为(3/10)/(4/11)。实验证实这三个特征对双语候选句对分类有很大的帮助。

表5 示例

中英候选句对	html开始标记	html结束标记	文本块集合中的相对位置
中文句子	〈p〉	〈/p〉	3/10
英文句子	〈span〉	〈/span〉	4/11

在给出的双语候选句对集后，要从中选出真正互为翻译的中英句子。首先要计算每一对候选句子为平行的概率和非平行的概率，根据这两个概率的大小进行分类。基于最大熵理论引入一个判别函数：

$$p(c_i \mid sp) = 1/Z(sp) \prod_{j=1}^{k} \lambda_j^{f_{ij}(c_i, sp)} \quad (1)$$

其中，c_i取值为c_0（不平行）与c_1（平行）；$1/Z(sp)$为归一化因子；$f_{ij}(c_i, sp)$为上述的特征对应的特征函数；λ_j为特征的权重信息，可以通过训练得到。

一个中文句子在中文网页中可能多次出现，英文句子也一样。在进行分类以后，可能一个中文句子与多个英文句子构成互为翻译关系，那么再引入一个平行句对选择函数：

$$sp_{ij} = \max_{j=1\cdots k}\{sp_{ij} \mid p(c_1 \mid sp_{ij}) > p(c_0 \mid sp_{ij})\} \quad (2)$$

其中sp_{ij}指的是中文句子i与英文句子j构成的双语候选句对。

5 获取单一网页内部的平行语料

根据双语平行语料的存在形式可将 Web 资源分为两大类即中英平行文本分别存在于两个中英平行的网页中和同一页面内的情形。以往的系统都主要集中于从前者这类 Web 资源中获取双语平行语料库。本文不仅要充分挖掘前者的潜在双语资源，而且尝试了如何获取并利用后者这类资源。

5.1 双语对照页面的获取

在本系统中对于获取这类平行文本，作了初步研究：通过观察，这类网页有一些共性。可以以一些关键词为锚文本，利用搜索引擎来获取包含这些锚文本的网页。常见的关键字例如“双语新闻”、“双语学习”、“双语阅读”等。这些网页很大一部分本身就有是双语对照的或者含有指向双语对照页面的链接。那么以这些网页作为种子，追踪其中包含的链接，进行深度下载，就可以获取相当大数量的这类网页，这就初步解决了这类双语对照页面的获取问题。

5.2 双语平行句对的抽取

中英文本在中英对照网页中主要有三种对照方式：上下对照、左右对照以及段落或句子之间的相邻对照。在本系统中采用了统一的处理方式，不受中英文本在网页中出现格式所限制。在这个环节中，先将一个含有中英对照文本的网页解析成图 1 所示的句子序列，去掉相应的 html 标记后，进行句子的语种识别，得到中文与英文的句子序列。这两者序列之间肯定存在着互为翻译的关系，并且由于这种类型的网页中英之间对照关系是非常严格的，所以得到的中英句子序列之间为直译的比例很大在本系统中，可以利用 4.2 节中介绍的抽取平行句对的方法，抽取出中英平行句对。

6 实验结果与分析

6.1 自动发现双语网页 URL 命名模板的实验结果

我们随机选出具有 URL 命名相似性的 18 个网站进行测试，比较本文的方法与 WPDE 系统中采用的方法。其中采用 WPDE 系统中的方法可以抽取出 2 110 对候选中英平行网页，而我们的方法可以找出 3 013 对候选中英平行网页，多找出 903 对中英候选平行网页，经过后续的双语候选网页过滤步骤，发现这 903 对中英网页确实为平行网页。这是因为采用自动发现网站内部双语网页 URL 命名特点这个方法，不仅可以避免预定义带来的缺失，还可以避免网站建设者采用大小写、省略词等问题造成的缺失。

6.2 双语平行句对抽取的实验结果

在人工标注后的 270 对中英平行网页中人工找出其平行的句对集合，总共 1 520 对，组成训练语料中平行句对部分。在剩下的所有中英句子进行简单的全组合，过滤掉中文句子和英文句子相对位置的比例小于 0.3，大于 3 这些明显不平行的句对，再从过滤后的集合中随机选出 1 520 对组成训练语料中不平行句对部分。

先从随机挑选出的 20 对中英平行网页中抽取出的中英候选句对集合，再从中随机选出 2 173 对中英句对组成测试集。在这个测试集合上，抽取中英平行句对模块最终取得了 93% 的召回率与 96% 的准确率，具体结果见表 6。

表 6 特征组合实验

特征类型	召回率	准确率
一般文本特征	89%	88%
一般文本特征＋词语对齐相关特征	90.5%	94%
一般文本特征＋词语对齐相关特征＋网页结构特征	93%	96%

这个实验结果来源于两方面的贡献，一个是双语候选网页过滤环节，本文中采用最大熵分类器对中英候选网页进行分类，融合了几个系统的长处，采用了一些对分类非常有帮助的特征。在随机挑选出的 450 对中英候选网页上取得了召回率为 99%，准确率为 97.8% 的结果，为中英平行句对的抽取做了很好的除噪音工作。另一个更为重要的环节就是本系统中采用的双语平行句对抽取技术。

可见在经过双语平行网页过滤以后，得到的中英平行网页中确实存在互为翻译的句对，而本文采用的双语平行句对抽取方法能够很好地利用网页本身的结构优点，并且对内容上互为翻译这个条件采用了合适且高效的特征，从而大大提高了召回率与准确率。

6.3 同一网页内部平行句对抽取的初步结果

对于这类资源的挖掘工作，本文目前已经解决

了从互联网上获取潜在的候选资源这个难题。下面仅以“双语阅读”这个关键字为例，来评估本文采取的获取方案是否可以获取大量的有用资源。通过Google检索得到不重复的732个页面。其中，确实为中英对照或者含有指向其他中英对照页面链接的网页有353个，无效或不存在页面有127个。剩下的252个页面中很大一部分也为双语对照的，包括中文与法语、韩语、日语以及俄语等。说明这一类型的网页是大量存在的，并且可以通过特定关键字的方式来获取。

通过“双语阅读”这个关键字获取得到的353个有用的候选资源中，仅以腾讯网2007年教育频道中的网页为例，就抽取出150个中英对照页面。从中可以得到450k中英平行句对。这说明在这种类型的网页中双语句对资源很丰富，并且这些网页所处的网站会不断地推出新的双语对照页面，为持续获取双语句对资源提供了一个便利的途径。这类页面中英对照部分基本上都是互为翻译的并且非常工整。

6.4 双语平行语料的挖掘结果

本文所实现的原型系统目前已经获取中英平行网页1.8万对，中英平行句子27.5万句对。本系统可以持续地获取更多的双语平行资源，进一步的结果请关注本项目的主页 http://mitel.ict.ac.cn/webmine/index.htm。

7 总结及下一步工作

本文构建的双语平行语料自动获取系统在融合现有系统优点的基础上，对关键技术进行了改进，取得了较好的效果。首先，利用URL命名相似性获取双语候选网页，打破了以往预定义前后缀词表的限制，根据每个网站URL命名的特点，自动发现命名规律从而获取更多可靠的双语候选网页。经过后续的双语平行网页过滤环节取得了高质量的文本级双语平行语料库。其次，采用了一种新方法进行双语平行句对的抽取，不仅利用了平行网页之间html结构的相似性，并且更加侧重双语句对之间的互翻译性，有效地提高了双语平行句对抽取的召回率和准确率。最后，本文还尝试从同一网页内部的双语网页资源中获取双语平行句对，取得了初步结果。

下一步工作我们将继续研究有效的双语网页分析和过滤技术，进一步提高双语句对获取的质量，去除重复句对，使得我们的系统可以持续、稳定、实时地获取大规模的双语平行句对。此外，我们也将深入探索同一页面内部的双语平行句对自动获取方法，使得该类双语平行句对的获取可以不受具体双语对照格式的限制。

参考文献：

[1] 孙乐，金友兵，杜林，等. 平行语料库中双语术语词典的自动抽取[J]. 中文信息学报，2000，14(6)：33-39.

[2] JiangChen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval[C]//Proceedings of the International Conference on Chinese Language Computing. San Francisco：2000：21-28.

[3] Jisong Chen，Rowena Chau，and Chung-Hsing Yeh. Discovering parallel text from the World Wide Web[C]//CRPIT'32：Proceedings of the second workshop on Australasian information security，Data Mining and Web Intelligence，and Software Internationalization. Australia：2004：157-161.

[4] Xiaoyi Ma and Mark Y. Liberman. Bits：A method for bilingual text search over the web[C]//Proceedings of the Machine Translation Summit VII，1999.

[5] Philip Resnik. Parallel strands：a preliminary investigation into mining the web for bilingual text[C]//Proceeding of the Third Conference of the Association for Machine Translation. America：1998：72-82.

[6] Philip Resnik and Noah A. Smith. The web as a parallel corpus[J]. Computational Linguistics，volume 29，349-380.

[7] Ying Zhang，Ke. Wu，Jianfeng Gao，and P. Vines. Automatic acquisition of chinese-english parallel corpus from the web[C]//Proceedings of ECIR-06，28th European Conference on Information Retrieval. ACL，2006.

[8] Lei Shi，Cheng Niu，Ming Zhou，and Jianfeng Gao. A dom tree alignment model for mining parallel data from the web[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Sydney：489-496，2006.

[9] Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora[J]. Computational Linguistics，volume 31，477-504.