基于 Web 的双语平行语料库自动获取系统[1]

叶莎妮　吕雅娟　刘群

中国科学院计算技术研究所智能信息重点实验室

{yeshani, lvyajuan,liuqun,}ict.ac.cn

摘要：利用互联网上存在的海量多语言文本资源，通过网页的内容分析和链接分析，实现了一个双语语料挖掘的自动获取系统。首先，介绍了系统框架和主要模块；其次，详细描述了各个模块的实现与创新技术；最后，给出下一步工作的展望。本系统为获取真实的中英平行语料库提供了有效的途经。

关键词：双语语料; 网页挖掘; 平行网页

A Bilingual Corpus Automatic Acquisition System Based on Web

Abstract: Implemented a bilingual corpus automatic acquisition system by taking advantage of an abundance of multilingual corpus in the World Wide Web, and analyzing their content and links. First, introduced system framework and main modules; second, described every module and technology innovations in detail. A prospect for the next step was given at last. This system provided an effective way for achieving Chinese-English parallel corpus.

Keywords: Bilingual Text; Parallel Corpora; Web Mining;

1. 引言

语料库的建设是统计学习方法的重要基础，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，像平行语料库为统计机器翻译的模型构建提供了必不可少的训练数据（e.g., Brown et al.1990; Melamed 2000; Och and Ney 2002），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（e.g., Gale and Church 1991; Melamed 1997）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（e.g., Davis and Dunning 1995; Jian-Yun Nie, TREC8;），翻译词典编撰、双语术语自动提取以及多语言对比研究等的重要基础资源。

双语平行语料库建设与获取存在着很大的困难，各国都投入了大量的人力、物力和财力，但是双语平行语料库的来源主要集中在政府报告、新闻法律等特定领域，不适合真实文本应用。同时，互联网上的大规模双语文本并且具有很好的时效性和覆盖性，这为双语平行语料库的获取提供了潜在的解决途径。

研究基于 Web 的大规模双语平行语料库获取技术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语料库自动获取系统。

收稿日期：2007-06-30 **返稿日期：**2007-XX-XX　　**基金项目：**国家自然科学基金（60603095），国家自然科学基金（60573188）

作者简介：叶莎妮，硕士研究生，主要研究方向为自然语言处理技术（yeshani@ict.ac.cn）；吕雅娟，工学博士，主要研究方向为计算语言学与机器翻译；刘群，工学博士，主要研究方向为计算语言学与机器翻译

2. 背景介绍

加拿大蒙特利尔大学的研究者聂建云开发的系统 PT Miner（Parallel Text Miner，1999）：通过搜索引擎查找含有特定锚文本的网站构成双语候选网站，再依赖预先定义的语言的前后缀表，抽取出具有 URL 命名相似性的候选网页即如果某一 URL 含有一种语言的前后缀，则将这些前后缀替换为另一种语言的，构建出一个 URL，如果这样构建出来的 URL 存在。则找到了一对候选网页对，最后再根据文本长度，网页的 HTML 标记结构，网页的语言等特征过滤掉候选网页中不平行的网页对。PT Miner 系统在中英平行网页文本挑出几百对的中英平行网页对，经过人工的评价，有将近 90%的准确率。获取到的英文文本有 137M，中文文本有 117M。

美国马里兰大学的研究者 Resnik 开发的系统 STRAND（Structural Translation Recognition, Acquiring Natural Data，2003）也是利用搜索引擎和定义的挑选候选网站的规则来得到双语候选网站。同 PT Miner 相比，STRAND 再利用 URL 命名相似性来查找一个网站内的候选网页对时，采取在中、英 URL 中删去预先定义与语言相关的字符串的方式，如果去除语言相关的字串后，中、英 URL 相等，则说明当前的中英 URL 是一对候选双语平行网页。此外，STRAND 更加细致深入的研究了平行网页在结构上具有的相似性，采用了更多的基于网页结构的特征来过滤掉候选平行网页中不是互为翻译的网页对。人工评估了大约 400 对的中英平行网页对，取得了 98%的准确率和 61%的召回率。STRAND 系统获取到大约 3，500 对中英平行网页对。

BITS（Bilingual Internet Text Search, Ma and Liberman 1999），下载指定域名下的所有网站作为候选网站，定义了一种计算中英网页内容之间相似度的计算方式即互翻译词占文本总词数的比例，来进行中英平行网页对的确定。

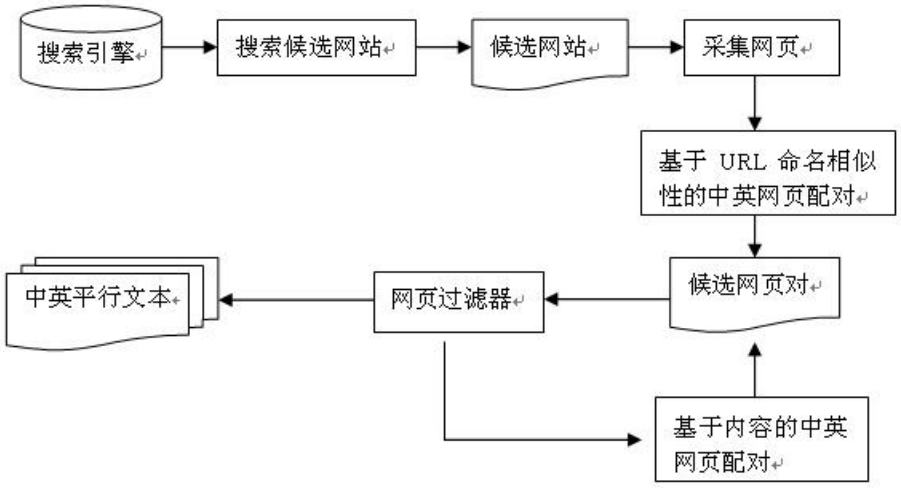
澳大利亚莫纳什大学陈纪淞等人开发的 PTI（The Parallel Text Identification System，2004）通过网页采集器下载了大量的双语网页之后，首先通过了文件名比较模型即根据 URL 命名的相似性来得到双语平行网页对，原理同 PT Miner，在这一过程没有相应对齐链接的网页再通过一个文件内容分析模型，定义了计算网页文本内容之间的相似度计算方式，从而得到双语平行的网页对。PTI 系统总共获取到 193 对的中英平行文本，其中 180 对是正确的，正确率为 93%，召回率为 96%。

亚洲微软研究院的吴克等人开发的 WPDE（Web Parallel Data Extraction，2006）在利用搜索引擎获取候选网站时，不仅利用了锚文本还采用了图片的 ALT 信息。在根据 URL 命名相似性获取候选双语平行网页对时，采用将 URL 分成 pathname 和 basename，pathname 的配对查找上也利用预先定义的启发式字符串，在具体的查找时定义了一些匹配规则；basename 的查找配对不用于前面系统采用的基于预先定义的字符串形式，而是基于改进的最小编辑距离算法，这样的方式经过试验证明取得了更好的效果。候选双语平行网页对的过滤时除采用了文本长度，网页 html 结构等特征，还引入了一个基于网页内容的特征即候选双语平行网页文本句子对齐的好坏。在同 PTI 同样的测试集合上，WPDE 系统取得了 97%的正确率与 94%的召回率。

3. 系统框架

本系统建立了一个基于 Web 的双语语料自动获取平台，能够自动获取文本级双语平行语料库和句子级的双语平行语料库。本系统主要获取中英平行网页文本，但是除了一些与具体语言相关的配置文件以外，采用的方法是不依赖具体语言的，可以很轻松的一致到其他的

语言对上。整个系统如下图所示，由四个关键部分组成。



图表 1　系统结构图

3.1 候选网站的获取

候选网站为可能含有中英平行网页的网站，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”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候选网站。可以通过搜索引擎或指定特定的域来获取候选网站。识别候选网站的目的是要进一步将双语文本的获取限制在可能的网站上。在得到候选网站之后，利用已有的网页采集器 Wget，下载了大量的双语网页

3.2 采集网页与预处理

利用现有的网页采集器 Wget 下载候选网站内部的所有网页，再一系列的预处理操作，例如：进行统一中文网页编码，都转化为 GB 格式；网页语言识别，分为中文网页与英文网页；统一转化为小写格式等。

3.3 基于 URL 命名相似性获取候选中英网页

通过观察可以发现相当一部分网站作者在 URL 命名时有一特点：会引入一些与特定语言相关的字符串，双语平行网页的 URL 只有语言相关的部分字符串不同，例如：“www.fao.org/newsroom/zh/index.html”与“www.fao.org/newsroom/en/index.html”，并且语言相关的部分其中大部分是常见的，可以通过预先定义的，所以已有的系统都通过预先定义与特定语言相关的字符串集合或正则表达式。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。此外，也存在一些网站的在命名时，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。可见，只要 URL 命名的相似性不涵盖在预先定义的集合中，那么已有的系统就无法处理。下面提出一种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，完全不需要预先定义语言相关的字符串集合。

3.3.1 根据 URL 命名相似性获取候选中英网页

我们通过观察统计发现在那些具有 URL 命名相似性的双语网站中，URL 的 pathname 与 basename 的相似性是不同的，应该将一个 URL 分为 pathname 和 basename 两部分，分别寻找当前网站中 URL 这两部分命名时的相似性，再进行查找候选平行网页对。

例如：www.fao.org/newsroom/zh/field/2005/ index.html

其中：“www.fao.org/newsroom/zh/field/2005/” 为 pathname，“index.html”为 basename

分别找出 pathname 和 basename 中语言相关的部分，分别生成 pathname 部分和 basename 部分的两组替换规则，依靠生成的 pathname 产生式与 basename 产生式来将中英 URL 配对，得到候选双语平行网页对。具体思想描述如下：

1. pathname替换规则的生成

a) 基本思想

将pathname看作由/str/组成字符串，每一个/str/是处理单元（字段），假设一个pathname中不存在重复的处理单元。获取pathname中符合以下规则的不同部分(可能是我们要的语言相关的部分）

简记为：cpath:f/lang_c/l/ ,epath: f/lang_e/l

f表示两个URL相同的前端部分

l表示两个URL相同的后端部分

lang_e，lang_c表示两个pathname中不同部分，为中间部分，可能含有多个字段。两个pathname可能存在多个这样的中间部分。

b) 替换规则定义

符合上述规则的两部分字段（lang_c,lang_e）表示一处替换,两个pathname中的所有的替换组成一个产生式，多处替换以分号分隔。

c) Pathname产生式的应用

cpart非空，epart非空：替换操作

cpart非空，epart为空：在中文URL中删除对应的cpart

cpart为空，epart非空：在英文URL中删除对应的epart

cpart为空，epart为空：则说明要查找相同的中文URL，英文URL

d) 生成替换规则与相应权重的计算（算法描述）

从中英文的URL列表得到相应的pathname列表，从中文pathname列表方向开始与英文pathname列表中的pathname比较，应用已有的替换规则与当前中文的pathname，构造一个英文的pathname，如果构造出来的英文pathname在英文的pathname列表中存在，则说明当前的替换规则有效，增大其权重，否则当前的中文pathname与英文列表中的每一个pathname进行比较寻找符合上述定义的新的替换规则。当当前生成的替换规则中cpart1，epart1；cpart2，epart2；任意一个部分是一个数字串时，当前的替换规则无效。

2. Basename产生式的生成

首先应用pathname产生式，找到具有相应的pathname的两个basename集合，在这两个集合中找替换的部分。例如：

根据pathname的替换规则（zh→en），中文pathname：www.fao.org/newsroom/zh/，对应的英文pathname：www.fao.org/newsroom/en/，分别得到具有该中文pathname的URL集合，再得到相应的中文basename集合（index.html，rss.xml），同样的可以得到英文的baseneme集合（index.html）。

a) basename字段划分

pathname具有很规整的格式可以由 / 来分成各个字段，而basename则根据”_.-“等分隔符

分成各个字段。

b) 替换规则定义

寻找两个basename中替换的部分生成潜在的产生式思想同生成pathname产生式。当找到符合规则的lang_c，lang_e后，当lang_c或lang_e中存在数字串，并且在lang_c与lang_e中同等位置的数字串不相等时，我们认为当前找到的替换规则不是语言相关的，是无效的

c) Basename产生式的应用

不采用替换的方式，采用删除当前产生式中出现的所有字符串的方式（因为pathname中字段不会重复，basename中字段(即字符)会重复）

例如，index_e.html,index_c.html

当前产生式：c,e；则delset{e,c},删除出现在delset中的字符串，

变成：indx_.html,indx_.html

相等则认为当前的cbasename，ebasename符合当前产生式的要求

采取本算法的明显优点在于可以根据每个网站自身的特点来找出URL pairs，不受限制，基本上不会遗漏正确的候选双语平行网页对。

3.4 双语平行网页的确认

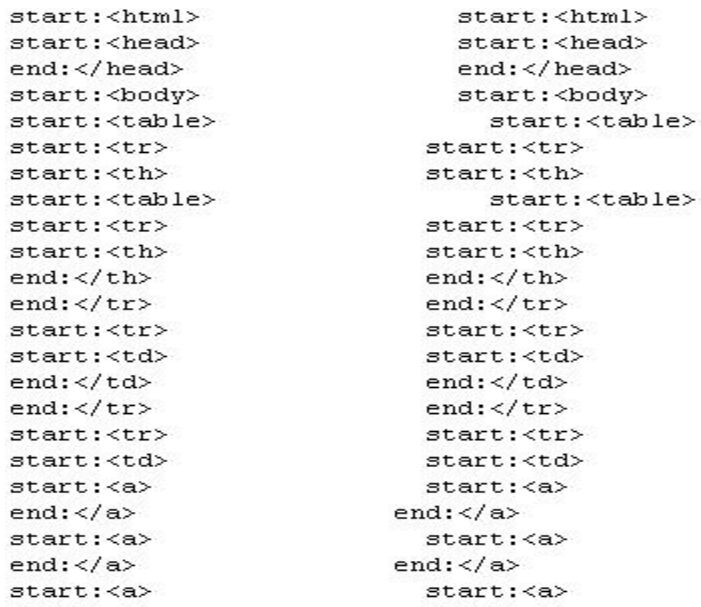
获得的候选中英网页中，存在着一些实际上不平行的网页，我们需要根据一些特征或判定准则过滤掉不平行的网页，得到真正平行的中英网页。本系统中采用文本长度、网页 HTML 结构、一对网页文本中的互翻译词比例以及词语对齐等特征，训练了一个最大熵分类器来进行候选双语平行网页对的验证，过滤掉实际上不平行的网页对。

3.4.1 文本长度特征

双语平行网页在文本长度是具有规律的，去除网页中的HTML标记、空白、空行等噪声得到文本，在进行切词的基础上定义文本长度为词数。基于文本长度的特征定义为F=length（ctext）/ length（etext）。

3.4.2 网页结构特征

抽取出网页的HTML标记，组成一个标记序列，然后再利用UNIX工具sdiff将中英网页对应的两个标记序列进行对齐，基于网页HTML结构的特征定义为：F=N（diff）/ N（all）。如下图所示：



图表 2 对齐 HTML 标记示例

3.4.3 内容互翻译词特征

双语平行网页在内容上应该是相关的，互为翻译的。我们首先在560万句中英平行语料中训练了两部带翻译概率的词典，中英词典与英中词典。然后我们统计中文网页文本中的有多少词在对应的英文网页文本中存在对应着翻译词，定义一个衡量内容互翻译性的特征：F_c2e=N(c2e) / N(call)即在英文网页中存在相应翻译词的词数与中文网页文本总词数的比值。相应的可以统计英文网页文本中那些词在中文网页文本中存在着翻译词，相应的特征表示为：F_e2c=N(e2c)/N(eall)。如果F_c2e，F_e2c都大于2，说明中英网页中都至少存在一般的词在另一侧网页中存着相应的翻译词。

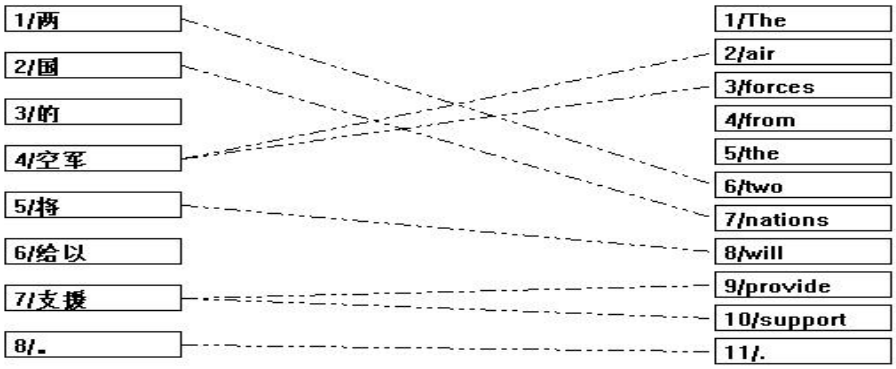
3.4.4 词语对齐特征

假如某一对中英网页描述的是同一新闻事件，但不是一一对应互为翻译的，虽然它们在文本长度，结构，内容互翻译词上都有很大的相似性，这一类网页我们称之为双语可比网页。我们引入词语对齐特征就是为了区分双语平行网页与双语可比网页。我们需要一个高效快速的词语对齐算法，因而我们采用基于词典的方式，经过两边的添加链接来完成。

a) 将每个中文词c_j链接到具有最大中译英或英译中翻译概率的英文词e_i上，即选择$\max_i\{p(e_i/c_c),p(c_j/e_i)\}$的英文词$e_i$,如果e在英文句子中只存在一个，就直接添加链接，否则在记录下所有e_i的位置，在第二遍添加链接时在决定c_j具体链接到哪一个e_i

b) 对于对应的英文词e_i在英文句子中存在多个的中文词c_j，不采取任意选择一个e_i的方式，我们在添加一条c，e链接时，使得当前存在的交叉连接最少。

经过上述两遍添加链接以后，仍然没有相应链接的中文词与英文词都当链接到两个空节点上，即添加空链接，这样就得到了一个词语对齐，例如（图片）



图表 3 词语对齐示例

定义词典对齐特征为：F=N(bilink) / N(alllink)，即两个词的链接数与总链接数的比值。对于例子给出的词语对齐，相应的词语对齐特征值为：8/（8+5）。

3.4.5 最大熵分类器

最大熵模型是最大熵分类器的理论基础，其基本思想是为所有已知的因素建立模型，而把所有未知的因素排除在外。也就是说，要找到这样一个概率分布，它满足所有已知的事实，且不受任何未知因素的影响。

最大熵模型的一个最显著的特点是其不要求具有条件独立的特征，因此，人们可以相对任意地加入对最终分类有用的特征，而不用顾及它们之间的相互影响。另外，相对 SVM 等基于空间距离的分类方法，最大熵模型能够较为容易地对多类分类问题进行建模，并且给各个类别输出一个相对客观的概率值结果，便于后续推理步骤使用。同时，最大熵的训练效率相对较高。上述优点使其成功应用于信息抽取、句法分析等多个自然语言处理领域。

手工选择训练语料，采取上述的特征，训练了一个最大熵分类器。

4. 实验结果与分析

1. 候选网站数量

我们定义中英锚文本列表，例如“Chinese”，“chinese version”，“简体”，“English”等等，然后通过搜索引擎 Google 检索含有以这些关键字为锚文本的链接的网页，经过候选网站的定义规则过滤后就得到了候选网站。

在我们实验过程，由于空间的限制，仅选用了以“简体”，“简体中文”，“简体版”，“中文”，“chinese_simplify”，“Chinese”，“chinese version”，“in Chinese”，“chinese_tradition”等作为锚文本。通过检索引擎获得的候选网站数量为 975。

2. Wget下载网页

用网页采集器 Wget 下载上一步得到的所有候选网站中的可用的文件，不下载“.gif”、“.jpeg”等类型的文件。

3. 自动获取具有URL命名相似性的候选双语平行网页

我们随机选出具有 URL 命名相似性的 18 个网站进行测试，比较我们的方法与 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 2110 对候选中英平行网页，而我们的方法可以找出 3013 对候选中英平行网页，多找出 903 对中英候选平行网页，经过过滤掉不平行的网页，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 URL 命名的特点，不仅可以避免预先人为定义带来的缺失，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。

5. 下一步工作

目前从网页这种类型的语料获取句子级对齐的语料，难点在于，网页本身的噪声，网页之间即使表述的是同一件事情，作者不同，文章内容就不同，无法做到真正意义上的平行，那么直观上理解：先做到文本块对齐（例如 … 可以看做一个文本块，带有开始标记 ,结束标记 ），块内再进行句子对齐。

在解析网页结构之后，我们可以得到所有带首尾 HTML 标记的文本块序列，如下图所示，

```
start:<a>
2005千年/@发展/@目标/@报告/@
end:</a>
start:<a>
2005年/@世界/@首脑/@会议/@9月/@14-16日/@, /@纽约/@
end:</a>
start:<a>
千年/@计划/@
end:</a>
start:<a>
秘书长/@报告/@“/@大/@自由/@”/@
end:</a>
```

图表 4 带首尾 HTML 标记的文本块序列

在这基础上进行文本块的对齐。换一个角度思考：文本块对齐可以看做一个分类问题，将候选的中英文本块也即可能是对齐一对的文本块，抽取一些特征，送入分类器进行分类，得到当前的候选文本块对是否是平行的文本块。我们同样可以采取文本块长度，互翻译词个数等互为翻译的文本块所具有的特征。

参考文献

[1] Chen, J., R. Chau, and C.-H. Yeh. 1991. Discovering Parallel Text from the World Wide Web. In Proceedings of the second workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalization.

[2] Nie, J. Y., M. S. P. Isabelle, and R. Durand. 1999.Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development.

[3] Resnik, P. and N. A. Smith. 2003. The Web as a Parallel Corpus. Computational Linguistics, 29(3)

[4] Zhang, Y., K. Wu, J. Gao, and Phil Vines. 2006.Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In Proceedings of 28th European Conference on Information Retrieval.

[5] Melamed, I. Dan. 1997. Automatic discovery of non-compositional compounds in parallel data. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), pages 97–108, Providence, RI, August

[6] FUNG, PASCALE ~ KENNETH W. CHURCH. 1994.K-vec: A new approach for aligning parallel texts. In Proceedings of the Fifteenth International Conference on Computational Linguistics, Kyoto. To appear.