基于双语混和网页的平行语料挖掘*

林政　吕雅娟　刘群　马希荣

中国科学院计算技术研究所 北京 100080

E-mail: {linzheng, lvyajuan,liuqun}@ict.ac.cn, maxirong@eyou.com

摘　要：双语平行语料是统计机器翻译模型训练必不可少的基础资源，但是大规模双语平行语料库的自动获取并不容易。本文提出了一种从双语混合网页上自动挖掘大规模双语平行语料库的解决方案，研究了候选双语混合网页的获取，网页噪声过滤，双语网页确认以及平行句对抽取等关键技术，最后实现了一个基于双语混合网页的平行句对自动挖掘系统。利用该系统获取了 105 万双语平行句对，平均正确率为 93%， 其中前 20 万获取的双语句对的正确率达到 99%。

关键词：Web 挖掘；双语混合网页；双语平行网页；平行语料库

Mining Parallel Corpora from Mixed-Language Web Pages

Lin Zheng, Lv Yajuan, Liu Qun, Ma Xirong

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080

E-mail: {linzheng, lvyajuan,liuqun}@ict.ac.cn, maxirong@eyou.com

Abstract: Bilingual parallel corpora is the indispensable resource of model training in SMT , but it's not easy to acquire large-scale corpora automatically. This paper proposes a solution to mine large-scale bilingual parallel corpora from mix-languages web pages and analyses the problems of obtaining candidate mix-language web pages, filtering web noises, validating bilingual web pages and extracting parallel sentences. We implement an automatic mining system of parallel corpora from mix-language web pages and have extracted 1.05 million parallel sentences which average accuracy is 93% , and the accuracy of the first 200 thousand sentences is close to 99%.

Keywords: Web Mining ; Mix-Language Web Page; Bilingual Parallel Web Page; Parallel Corpora

1 引言

双语平行语料库在自然语言处理领域有很多重要应用，它为统计机器翻译模型提供不可或缺的训练数据，同时也是词典编纂和跨语言信息检索等应用的重要基础资源。但是大规模双语平行语料库的获取并不容易，现有的平行语料库在规模、时效性和领域的平衡性等方面还不能满足处理真实文本的实际需要。

随着互联网的普及和迅猛发展，越来越多的双语网站被创建，越来越多的信息以多语言的形式发布，这就为双语和多语语料库的建设提供了很大的来源。一些研究者提出了基于 Web 的双语或多语平行语料库自动挖掘方法[2,3,4,5]，为双语或多语平行语料库的自动构建提出了有效的解决途径。

目前已有的研究方法主要是从两个平行的单语网页间抽取双语平行文本，即每种语言的文本分别位于不同的两个网页中，两个网页之间是相互翻译的。一般在双语网站中获取的双语网页都是属于这种类型，我们称之为双语平行网页。这些研究中获取双语平行网页的常用方法是：首先利用搜索引擎和双语网站中的语言标志作为启发式信息（如网站中的“English Version”，“中文

*本文承国家自然科学基金（60603095）的资助。

版”等）来获取候选双语平行网站。然后再利用网页 URL 地址的相似性(如 file_*e*.html 和 file_*c*.html)来获取平行网页[Philip Resnik 2003] 。还有一种基于网页结构相似性的方法[Lei Shi 2006]，通过追踪平行网页上的链接，分析网页之间的 html 标签结构（DOM tree）的相似性，不断迭代发现新的候选平行网页。基于平行网页的双语平行资源获取方法取得了很好的效果，为双语平行语料库的自动获取提供了有效方案。

在研究中我们发现，Web 上还存在着大量的另一类双语网页，这些网页中双语内容同时出现在同一网页内，比如一些双语学习网页、双语新闻网页等，我们称之为双语混合网页。例如网页 http://tr.hjenglish.com/page/73132/，与双语平行网页相比，双语混合网页上的双语资源对照更为工整，翻译质量较高，是非常宝贵的双语资源来源。但是目前国内外学者对这类网页翻译资源的获取研究还不多见。研究基于双语混合网页的双语平行语料库自动获取可以有效地利用这类丰富的双语候选资源，是对前人基于平行网页双语资源获取工作的一个有益补充。

本文提出了一套有效的从混合网页自动获取平行语料的解决方案，包括候选混合网页的发现和获取，网页噪声过滤，双语网页确认以及平行句对抽取。本文方法的主要优点是：获取的双语句对的质量比较高，翻译对应性比较好； 双语平行语料的规模比较大，因为 web 上存在着大量的双语混合网页；系统具有增量更新功能，可以周期性从 web 上采集双语平行语料。

下面将分别讨论候选双语混合网页获取、网页噪声过滤、双语网页确认和句子对齐问题以及相应的解决方案。

2 候选双语混合网页的获取

相对于候选双语平行网页的获取来说，候选双语混合网页的自动发现更为困难。因为这类网页的分布通常不确定，缺乏一些常见的启发式信息（如双语网站获取中的“中文版”“英文版”等）。这里我们尝试了三种双语混合网页的获取方法：

第一种是限定目标源的方法，事先收集整理若干个相关主题的网站，比如英语学习网站和翻译网站等，然后递归下载。这种做法的优点是候选资源质量较好，缺点是网页数量有限且网站的选择需要人工干预。

第二种是利用搜索引擎的方法，结合搜索引擎和启发式信息可以得到一些网站。通过搜索引擎返回的链接所对应页面大都含有双语对照信息，以这些链接作为种子链接，可以进行递归下载。这种方法的优点是可以自动发现候选网站，缺点是候选资源良莠不齐，会下载到大量不是双语混合的无关网页，所以对于获取的候选双语混合网页还需要进行进一步的判断确认。

第三种方法是 RSS 订阅，给定网站列表 RSS 订阅器会及时返回网站更新，从而实现源源不断的资源获取。这种方法的限制在于并非所有的网站都提供了 RSS 订阅功能，但是它提供了一种很好的增量更新解决方案。

本文主要采用第二种方法，即通过搜索引擎获取候选资源，然后对返回的链接进行后期处理得到候选网站列表。

3 网页噪音过滤

Web 文档不像传统文本那样整齐干净，包含大量的噪声内容，比如链接、广告、图片、导

航条等等。这些噪音通常分布在网页的不同位置，缺乏规律性。噪音的存在会影响句子对齐的正确率，所以必须对网页文件到文本文件进行转换，去掉所有的 html 标记和无用信息。

在候选资源获取环节中，我们是以网站为单位进行递归下载的。查看预处理后的文本，发现每个网站的噪声各异，因为不同网站的编辑规则通常不同，所以很难定义一组通用的规则来处理所有的候选网站。但是仔细观察，发现同一个网站内部的噪声分布和内容是大致相似的，比如

> “Copyright © 1996-2008 SINA Corporation, All Rights Reserved”（新浪）
> “中国国际广播电台国际在线版权所有©1997-2007”（国际在线）

所以，我们想到一种基于模板的网页去噪声方法。首先，我们对所有候选网页进行预处理，把<script>、<img>、<a>等标签标记的内容删除，再把所有的 html 标签删除。

然后，对普通文本构造模板，具体算法如下：

1) 将每个网站内部的所有文本扫描一遍，抽取短行。这里短行的定义是小于文本平均段落长度的行。
2) 统计短行的行频，即统计每个短行出现的次数，从高到低排序。
3) 将高频(频率大于某一阈值)短行记录为噪音集合 C1。
4) 对中频短行进行最大公共子串提取，得到高频子串，记录为噪音集合 C2
5) 对低频子串进行变换统计：空格或制表符 S,中文 C,英文 E,数字 N，得到高频组合，记为噪音 C3。比如网页噪声“编辑时间：2009-04-10”对应着高频组合“CCCC：NNNN-NN-NN”。

最终，我们的噪音集合就是 C=C1+C2+C3。

这种方法运用分而治之的思想，N 个网站就会自动生成N个噪声模板，然后每个网站就可以分别参照自身对应的噪音模板进行过滤了。

4 双语网页确认

获取的候选双语混合网页并不一定是真实的双语内容对应的网页，有很多单语网页或者英语试题等等，因此必须判别哪些是真实的双语混合网页。本文对双语平行网页的确认主要分为两步来完成，分别是基于双语字符数的粗判别和基于词典的细判别。

通常双语平行网页中两种语言的字符数是成比例的，以中英为例，假设中文文件的字符数为 number_zh，英文文件的字符数为 number_en，当“number_zh/number_en > T”或者“number_en / number_zh > T”时，则认为是非双语平行网页。在我们的实验中，T 的取值为 3。

为了进一步提高判别的正确率，本系统在第一步判别过滤的基础上又进行了一步基于词典的细判别。因为有时候网页中中英文字符数尽管符合一定比例，但是内容上不是互为翻译的，所以我们引入词典作为参照，判断网页的中英文是否是互译的。为了避免中文分词带来的错误，我们采用英文到中文的方向进行查词。

定义互译率 $transratio = \dfrac{\text{count}(hit\_ch\_word)}{\text{count}(total\_en\_word)}$，即查词命中的中文词数比上英文总词数，如果 $transratio$ 大于 1/2 认为是双语混合网页，通过查词典验证可以进一步过滤掉那些内

容上不是互为翻译的候选双语混合网页。

5 平行句对抽取

在进行混合双语网页确认过后，我们得到的是篇章级或段落级对齐的双语文本，而统计机器翻译模型训练需要的是句子级对齐的双语平行语料库，所以还需要在两个单语文本之间抽取双语平行句对。设 S 和 T 是互为翻译的两段文本，S 中包含 m 个句子，T 中包含 n 个句子，用 p 表示 S 和 T 中的一个最小对齐，称 p 为一个句珠。则 S 和 T 两段文本的对齐关系可以表示成一个句珠序列 P：$P = p_1 p_2 ... p_k$，其中 p_i 包含 S 中的 x 个句子和 T 中的 y 个句子。设 h 为每个句珠 p_i 的评价函数，则 $h(p_i)$ 为句珠 p_i 的评价分值，我们要找出匹配最好的句珠序列，则需要对各种可能的句珠序列进行评估。因此，句子对齐问题就可以表示成一个最优化求解问题，即找到一个句珠序列 P 使得该序列的总分值最优。这实际上是一个标准的动态规划问题，句子对齐问题的关键是评价函数的选择。。

基于长度的句子对齐的基本思想是互译的句子通常句子长度符合一定的比率，设两种语言的句子长度分别为 l_1 和 l_2，c 是互译句子平均长度比的期望，σ 是互译句子平均长度比的标准差，则变量 $\delta = (l_2 - l_1 c) / \sqrt{l_1 \sigma^2}$ 服从标准正态分布，任意句子 S_i 和 T_j 对齐的可能性可以表示为一个条件概率：

$$P(Mathe(s_i, t_j) \mid \delta) = \frac{P(\delta \mid Match) \times P(Match)}{P(\delta)}$$

这种基于长度的方法利用了统计学的原理，是一种被广泛使用的双语句子对齐方法，它的优点主要在于对其效率比较高，执行速度快，因为句子长度的计算非常简单。

Stanley F.Chen 通过建立词到词的翻译模型，实现了另一种基于词典的句子对齐方法。基于词典的方法是通过计算源语言和目标语言句子之间的互译信息来估计句子互为翻译的概率。设源语言句子共有 a 个单词，目标语言句子共有 b 个单词，其中互为翻译的词数为 c 对，则句子间的互译概率为：$r = \frac{2c}{a+b}$。

Wu、Utsuro 将长度方法和词典方法相结合，分别进行了汉英和日英句子的对齐试验，得出了混合方法好于单纯的长度方法或者词汇方法。

本文的主要工作是在以上工作的基础上又考虑了标点符号和数字、缩略词等其他混合信息，实现了一个汉语和英语的句子对齐方法。经过大量统计，互为翻译的句子间会使用对应的标点符号，尤其是出现频率较低的符号，比如“%、￥、&”，如果一个句对中的句子同时出现上述符号则很有可能是互译的；在从英文向中文查词典时，有些缩略词是查不到的，比如 Google、Baidu，所以考虑这些专有名词在句对中的对应，也可以进一步提高句子对齐的准确率。

6 实验与评价

利用本文方法共获得候选双语混合网页 119 万个，其中能够成功转化编码执行预处理的网页有 71 万个。因为不同网站的网页编码是不同的，常见编码有 UTF-8、GB2312、GBK 等等，常见编码之间是比较容易转换的，有些网页使用的编码是不常见的，有些网页同时使用多种编码，这都为编码转换造成一定困难。

表 1 给出了双语网页确认的实验结果。可见，基于双语字符数量的判别已经可以过滤掉大量的非双语平行网页了，但是正确率不是特别高，所以还需要更精细的判别。在第一轮过滤结果的基础上，再通过基于词典的判别方法进行过滤，得到双语平行文本 13 万个，经过抽样统计，正确率为 80.72%。这个结果对于双语平行句对抽取程序而言基本是可以接受的，因为平行句子对齐程序也考虑到了一定的文本噪声情况，可以通过调整阈值来控制双语平行句对的准确率和召回率。

方法	输入	输出	正确率
基于字符数比的双语网页确认	71 万	42 万	67.28%
基于词典的双语网页确认	42 万	13 万	80.72%

表 1 双语网页确认实验结果

双语平行句对	正确率
1 – 200000	99%
200001 – 400000	94%
400001 - 600000	91%
600001 - 800000	83%

表 2 句对重排序后正确率统计

对 13 万双语平行文本进行句子对齐处理得到双语平行句对 158 万对，去重后得到 105 万对。对已经得到的 105 万个双语平行句对进行平均随机抽样，每 1000 句中随机抽样 1 句，共抽样 1050 句对，然后通过人工查看的方式检验正确率，如果互译程度大 50%则认为是平行句对是正确的，否则是错误的，统计得到平均正确率为 93%。为了更好地利用获取的双语平行句对，我们定义了一个评价函数给每一个平行句对打分，然后进行重排序。

定义评价函数 $\mathrm{F} = \mathrm{Len\_Ratio\_Score}(S,T) + \mathrm{Trans\_Rate\_Score}(S,T)$

Len_Ratio_Score(S,T)是源语言句子和目标语言句子的长度比得分：

$$\mathrm{Len\_Ratio\_Score}(S,T) = 2\left(1 - \int_{-\infty}^{\delta} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} dt\right) \qquad -\infty < \delta < +\infty$$

双语平行句对的句子长度比值服从正态分布，则变量 $\delta = (l_2 - l_1 c)/\sqrt{l_1 \sigma^2}$ 服从标准正态分布，$|\delta|$越小则平行句对的相似性越高，则 Len_Ratio_Score(S,T)的分值越高。

Trans_Rate_Score(S,T)为源语言句子和目标语言句子的互翻译率得分：

$$\mathrm{Trans\_Rate\_Score}(S,T) = \frac{\mathrm{count}(hit\_ch\_word)}{\mathrm{count}(total\_en\_word)}$$

平行句对的互翻译程度越高则 Trans_Rate_Score(*S,T*)得分越高。

对已经获取的105万双语平行句对按照评价函数F打分排序，取前80万句对进行评价，然后分成四个等级，每1000句随机抽样1句，每一组抽样200句，然后通过人工查验的方式统计正确率，结果如表2所示。

由此可见，前60万平行句对的质量基本还是比较好的，评价函数F的设定也基本是合理的，参照评价函数F对双语平行句对进行重排序后，不仅可以过滤掉很多垃圾对齐，还能将双语平行句对按照等级加以利用。

7 总结与展望

本文的主要工作是提出了一种从双语混合网页挖掘平行语料的方法，而以往的工作主要是从双语平行网页间进行语料获取，并且实现了一个自动从Web上获取双语平行语料的系统。解决了候选资源获取、网页噪声过滤、混合双语网页确认等问题，可以自动获取大规模双语平行句对，以缓解目前大规模双语平行语料库建设的困难。本文研究表明，双语混合网页是双语平行语料库获取的重要候选资源，具有资源丰富，双语对照整齐，获取的双语平行句对翻译质量高等优点。本文研究是对前人工作的一个很好的补充。在以后的研究中，我们希望解决以下几个方面的工作：

第一，网页去重，即把内容相同的网页过滤掉，避免相同语料的重复获取。

第二，进行更精细的噪音过滤，以提高句子对齐的质量。

第三，将从Web上获取的双语平行句对应用于统计机器翻译的训练，看是否可以提高统计机器翻译系统的性能。

参 考 文 献

[1] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics

[2] Resnik,p. and N.A.Smith.2003.The web as a Parallel Corpus.Comoutational Linguistics

[3] Chen Jiang and Jian-Yun Nie. 2000. Web parallel text mining for chinese english cross-language in-formation retrieval.In International Conference on Chinese Language Computing, Chicago, Illinois

[4] Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao,2006.A DOM Tree Alignment Model for Mining Parallel Data from the Web In Joint Pro-ceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics, Sydney, Australia

[5] Zhang,Y.,k.Wu,J.Gao,and Phi Vines.2006.Automatic Acquisition of Chinese-English Parallel Corpus from the web.In Proceedings of 28th European Conference on Informtion Retrieval

[6] GALE, WILLIAM A. & KENNETH W. CHURCH.1993. A program for aligning sentences in Blingual c orpora.Computational Linguistics

[7] DeKai Wu.Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria.Proceedings of the 32nd Annual Conference of the Association for Computational Linguaistics

[8] Stanley F.Chen.Aligning Sentences in Bilingual Corpora Using Lexical Information Proceedings of the 31st Annual Meeting of the Association for Computational Linguaistics

[9] T.Utsuro,H.Ikeda.Bilingual Text Matching using Bilingual Dictionary and Statistics.In 15th COLING,1994