文章编号：1003-0077(2010)05-0085-07

Web 平行语料挖掘及其在机器翻译中的应用

林 政，吕雅娟，刘 群，马希荣

（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：双语平行语料库在自然语言处理领域有很多重要应用，但是大规模双语平行语料库的自动获取并不容易。该文提出了一种有效的从 Web 上获取高质量双语平行语料库的方案，研究了候选双语混合网页获取和平行句对抽取等关键技术。运用该文方法共获取了 258 万双语平行句对，平均正确率为 93.75%，其中前 150 万句对的平均正确率达到 96%。该文还提出句对质量排序和领域信息检索两种方法将 Web 数据应用于统计机器翻译的模型训练，在 IWSLT 评测数据上 BLEU 值可以提高 2 到 5 个百分点。

关键词：Web 挖掘；平行语料库；句子对齐；统计机器翻译

中图分类号：TP391　　**文献标识码**：A

Mining Parallel Corpora from Web and Its Application in Machine Translation

LIN Zheng, LV Yajuan, LIU Qun, MA Xirong

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

Abstract: Bilingual parallel corpora can be used in many applications of NLP, but it's not easy to acquire the large-scale corpora automatically. This paper proposes an effective solution to mine high-quality bilingual parallel corpora from web pages and analyses the key technology of obtaining candidate mix-languages web pages and sentence alignment. We have extracted 1.67 million parallel sentences, which average accuracy is 93.75%, and the accuracy of the first 1 million sentences is 96%. This paper also proposes the sentences re-ranking method and domain information retrieval method to apply the web data to the training of SMT model. Experiments conducted on the IWSLT tasks show 2 to 5 BLEU gains over baseline.

Key words: Web mining; parallel corpora; sentence alignment; statistical machine translation

引言

在计算语言学的发展和研究中，双语平行语料库的作用日益突显，双语平行语料库可以用于统计机器翻译的模型训练[1]，双语语料库的建设对于双语词典编纂[2]、跨语言信息检索也有重要价值。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需要[3]。随着互联网的普及和迅速发展，越来越多的信息以多语言的形式发布，这就为双语或多语的语料库建设提供了资源。

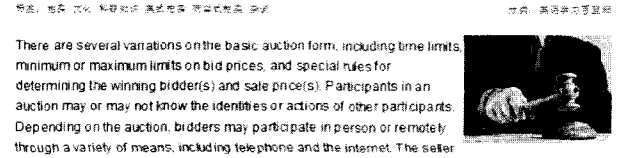
Web 上的双语网页大致可以分为两类，一类是双语信息分布于两个对照的网页间，本文称之为双语平行网页（例如 http://www.gov.hk/tc/residents/和 http://www.gov.hk/en/residents/）；另一类是双语信息包含在同一个网页内，如图 1 所示，本文称之为双语混合网页。已有的研究方法主要处理的是双语平行网页，获取双语平行网页有两种常用方法：一种是基于 URL 相似性[4-5]：首先利用搜索引擎和双语网站中的语言标志作为启发式信息（如网站中的“English Version”、“中文版”等）来获

收稿日期：2009-09-27　**定稿日期**：2010-03-25

基金项目：国家自然科学基金资助项目(60603095)

作者简介：林政(1984—)，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，副研究员，主要研究方向为自然语言处理技术；刘群(1966—)，男，博士，研究员，主要研究方向为自然语言处理技术。

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。另一种是基于网页结构相似性的[6-7]：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。基于双语平行网页的双语平行资源获取方法取得了很好的效果，为平行语料库的自动获取提供了有效的解决方案。


There are several variations on the basic auction form, including time limits, minimum or maximum limits on bid prices, and special rules for determining the winning bidder(s) and sale price(s). Participants in an auction may or may not know the identities or actions of other participants. Depending on the auction, bidders may participate in person or remotely through a variety of means, including telephone and the internet. The seller usually pays a commission to the auctioneer or auction company based on a percentage of the final sale price.

拍卖在基本形式之上存在着某些不同，包括限时、最低或最高竞投价格、以及用于决定竞价胜出者和成交价格的特别规则。参与拍卖的人不一定知道其他参与者的身份或者行为。根据拍卖形式的不同，竞投者可能会亲自出席，或者运用各种远程手段——例如电话和因特网——参与拍卖会。卖家一般会根据最终成交价格的一定比例付给拍卖商或拍卖公司佣金。

http://blog.sina.com.cn/s/blog_486c0f670100didk.html?tj=1

图 1 双语混合网页示例

双语平行网页存在地址或结构上的相似性，处理方法已经很成熟，但这些方法并不适用于双语混合网页。双语混合网页与双语平行网页相比，双语对照更整齐、翻译质量较好、句对长度适中，然而双语混合网页不存在地址和结构上的相似性，很难自动发现和区分，而且页面组织形式多样，很难精确抽取主体内容。目前对于双语混合网页的解决方案仍比较少，一种自适应模式学习的方法[8]最近被提出，该方法首先利用翻译和音译模型找到网页中的翻译词对作为种子，然后利用种子学习泛化的模板，最后利用学习到的模板抽取网页中所有的双语平行数据。这种方法可以获取大量的双语平行句对，但是正确率只有 83.5%。本文提出了另一种从双语混合网页自动获取双语平行语料的方案，不仅可以获得大量双语平行句对，而且正确率比较高，平均正确率有 93.75%，前 150 万的平均正确率可以达到 96%。本文提出的决方案解决了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等难点问题。此外，本文将从 Web 上获取的双语平行句对应用于统计机器翻译的模型训练，提出了句对质量排序和领域信息检索两种不同的应用策略将 Web 平行语料加载到训练集中，实验证明本文提出的两种策略可以提高翻译系统性能，在 IWSLT 评测任务中 BLEU 值可以提高 2 到 5 个百分点。

本文第 1 节主要阐述候选双语混合网页的获取方法，第 2 节描述如何从双语混合网页抽取平行句对，第 3 节研究 Web 平行语料在统计机器翻译中的应用策略，第 4 节是实验结果，第 5 节是对全文的总结和对未来工作的展望。

1 候选双语混合网页获取

1.1 候选双语混合网页获取方法

相对于候选双语平行网页而言，候选双语混合网页的获取更为困难。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)。本文首先介绍两种获取双语混合网页的常用方法：

方法一：限定目标源的方法，预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。

方法二：利用搜索引擎的方法，通过搜索引擎和启发式信息可以获得大量链接，然后以这些链接作为种子链接，进行递归下载。

本文结合以上两种方法，提出了第三种候选资源获取方法——尝试下载策略。首先利用搜索引擎和启发式信息得到一个候选网站列表，比如以“双语新闻 英汉”为启发信息用 Google 进行检索，可以得到不重复的 524 个链接。通过查看，这些页面大致可以分成三类：

(1) 目录型网页：通常是所有双语新闻或双语阅读的标题链接页，追溯链接可以得到大量的双语混合网页。此类网页可以递归下载。

(2) 内容型网页：通常本身是双语混合网页，但是追溯链接得到的都是无关网页，比如某人博客中一篇双语文章。此类网页不可以递归下载。

(3) 无关网页：既不是目录型网页，也不是内容型网页。此类网页不可以递归下载。

真正的候选网站列表应由目录型网页组成，若对内容型网页和无关网页进行递归下载，将得到大量的无关网页，不仅会占用较大的存储空间，还会影响系统的执行效率。由于不同网站的设计风格和组织架构各式各样，所以很难利用规则判断或是特征分类的思想对这三类网页进行区分，所以本文采用了一种尝试下载策略。把通过启发式信息和搜索引擎返回的所有种子链接分别追溯至下一层，即只下载当前页面和当前页面上的链接所对应的页面，不再进行更深层的采集。然后用 2.2 节所提到的方法

进行双语混合网页确认，如果一个种子链接所对应的下一层含有 5 个(经验值)以上的双语混合网页，则认为此种子链接可以进行递归下载，将其放入候选网站列表中，否则将其舍弃。

1.2 方法比较

方法一的优点是候选资源质量较好，避免了大量非双语混合网站的下载，缺点是网页数量有限且网站的选择需要人工干预。

方法二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，会下载到大量非双语混合的无关网页，需要对大量无关网页进行过滤，空间和时间开销都很大。

方法三同时具备以上两种方法各自的优点，即实现了高质量候选网站的全自动筛选，克服了方法一和方法二各自的不足。

为了衡量以上三种方法各自的特点，本文进行了一组实验，即在相同时间下，考察分别使用三种方法得到的候选网站的正确率和候选网站的数量。

表 1　候选双语混合网站获取方法比较

方法	候选网站质量(正确率)	候选网站数量
方法一	100%	87
方法二	41%	524
方法三	100%	215

综合考虑，在相同的时间开销下，方法三是最优的候选资源获取方法，候选网站的质量比方法一持平，候选网站的数量是方法一的 2.47 倍。与方法二相比，候选网站的正确率提高了 59%，与此同时空间开销也降低了一半。

2 双语平行句对抽取

通过第一节的方法可以获取大量的候选双语混合网页，然后需要从候选的双语混合网页中区分出真正的双语混合网页，并从真正的双语混合网页中抽取双语平行句对，主要任务可以分成三部分：网页噪声过滤、双语混合网页确认和句子对齐。

2.1 网页噪声过滤

Web 文档包含了大量的噪声内容，比如广告链接、导航条和图片等，这些噪音通常分布在网页的不同位置，缺乏规律性。大量的网页噪声不仅会影响双语混和网页确认的准确率，也会影响句子对齐的准确率，所以首先需要对所有候选网页进行噪声过滤。

本文采用一种专用的基于模板的网页噪声过滤方法。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，所以可以在每个网站内部自动学习噪声模板，N 个网站就会自动生成 N 个噪声模板，然后每个网站分别参照自身对应的噪音模板进行过滤，具体算法见文献[9]。

2.2 双语混合网页确认

候选双语混合网页并不一定是真实的双语内容对照的网页，有很多单语网页或者英语试题等等，因此必须区分真正的双语混合网页和非双语混合网页。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，以中英文为例，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。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，比如英文试题的解析，所以引入词典作为参照，判断网页的中英文是否互译。为了避免中文分词带来的错误，所以采用从英文向中文方向查词[10]，定义互译率 $transratio = \frac{count(hit\_ch\_word)}{count(total\_en\_word)}$，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。

2.3 句子对齐

经过上一步的双语混和网页确认，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。

Brown 和 Gale 最早提出了基于长度的句子对

齐方法[11]。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[12]。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[13-15]。

本文的主要工作是在长度加词典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。基于混合特征的句子对齐方法主要考虑了3类特征，分别是：

(1) 长度特征：这是最广泛被采用的特征，因为互为翻译的句子长度符合一定比率。

(2) 翻译特征：利用翻译特征来进行句子对齐可以大幅度提高对齐的精度。

(3) 符号特征：句子中的符号主要包括标点符号、数字、缩略词等。互为翻译的句子通常会使用对应的标点符号。

一些出现频率较低的符号具有很高的参考价值，比如?、!、*、$。句子中的数字和缩略词一般不会出现在双语词典中，在互译文本中却经常采用相同的形式，比如表示日期、数量、专有名词、机构名等。因此，考虑符号特征对句子对齐是有意义的，可以作为长度特征和翻译特征之外的一个很好的补充。

3 Web平行语料在统计机器翻译中的应用

这一节主要研究 Web 平行语料的特点，以及根据 Web 平行语料的特点提出的两种将 Web 语料应用于统计机器翻译的方法。

3.1 Web平行语料特点

从 Web 上获取的双语平行语料库主要有三个特点：

领域分布广泛：Web 平行语料是从互联网上随机采集的，可能来自于政府的官方新闻网站，可能来自于英语学习网站，可能来自于某人的博客等等，所以具有领域分布广泛的特点。现有的双语平行语料库通常都是限定领域的，比如官方的双语法律文档，而 Web 平行语料库的多领域性可以克服现有平行语料库领域局限的不足，也为领域性课题应用提供了很好的基础资源。

实时数据更新：由于互联网上的数据及时更新瞬息万变，所以从 Web 上获取的双语平行语料具有一定的实时性，可以捕捉到最新颖的词汇和翻译，新词发现是计算语言学中的一个重要课题。比如“我被雷到了。”对应英文翻译“I am startled.”，把这些实时数据加到统计机器翻译系统的训练集中可以让系统学习到更多的知识从而提高性能。

存在噪音干扰：Web2.0 时代的最大特点就是用户的参与性，从 Web 上获取的双语平行数据很多来自互联网用户的个人发布，比如论坛中的翻译擂台，所以可能存在一些拼写和语法上的错误，这些噪音的存在使得 Web 平行语料不可能具有百分之百的正确率，所以需要去粗取精提取真正有价值的信息。

为了确认 Web 平行语料的领域分布情况，我们进行了如下实验，对从 Web 上获取的双语平行文本进行分类，分类器采用的是中国科学院计算技术研究所的 DRAP 分类系统，这种分类器的效果要优于支持向量机、朴素贝叶斯和 K 近邻等分类技术（详情参照 http://www.searchforum.org.cn/tansongbo/software.htm），分类结果如图2所示。

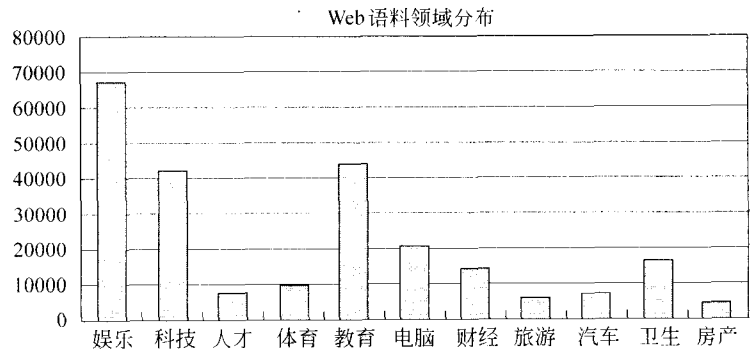


图2 Web双语平行语料领域分类

从分类结果图可以看出，Web 双语平行语料的领域分布比较随机，其中娱乐、科技、教育和电脑四个领域的分布比例较高，其他领域所占比例较小，可见这些领域的双语平行语料比较稀缺，而 Web 双语

平行语料本身规模巨大，所以对这些稀缺领域的平行语料获取是非常有意义的。

由于 Web 双语平行语料存在一定的噪音干扰，且领域分布非常随机，所以若将其直接加载到统计机器翻译的模型训练中效果并不理想，因此根据 Web 双语平行语料的特点，我们提出了两种应用策略，使其更好的适应实际应用的需要。

3.2 Web 平行语料应用于 SMT 的两种方法

(1) 句对质量排序方法

统计机器翻译系统的性能通常和双语平行句对的质量成正比，所以本文提出一种平行句对打分重排序的方法，以挑选质量较好的双语平行句对。这里定义一个评价函数为每一个平行句对打分，然后将平行句对按得分由高到低排序。

定义评价函数：F=Len_Ratio_Score (S,T)+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：

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

$$-\infty<\delta<+\infty$$

双语平行句对的句子长度比值服从正态分布，则变量 $\delta=(l_2-l_1c)/\sqrt{l_1\sigma^2}$ 服从标准正态分布，$|\delta|$ 越小则平行句对的相似性越高，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：

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

平行句对互翻译程度的得分定义为：按照从英文向中文的方向查词典，中文解释在中文句子中命中的次数比上英文句子的总单词数。平行句对的互翻译程度越高则 Trans_Rate_Score(S,T)得分越高。

(2) 领域信息检索方法

基于统计的机器翻译方法使用双语平行语料库作为翻译知识的来源，翻译知识的获取在翻译之前完成。基于统计的方法需要大规模双语平行语料，其翻译模型、语言模型参数的准确性直接依赖于语料的规模，其翻译质量主要取决于概率模型的好坏和语料库的覆盖能力。在已知测试集领域的情况下，可以挑选与测试集领域相关的双语平行语料进行模型训练，使机器翻译系统学习到的翻译知识尽可能的与测试集一致，从而提高翻译质量。因此，为了更好的利用 Web 双语平行语料库，本文提出了一种领域信息检索的方法从 Web 双语平行语料库种检索与测试集相似的句子用于模型训练，具体分为三步：

(1) 在测试集上建立索引，本文使用开源的信息检索工具 Lemur 进行索引的建立和查询。

(2) 把 Web 平行语料库中的句对逐一作为查询字符串，检索测试集中与之相似的句子，然后把返回的所有句子的相似度得分相加，即得到 Web 平行语料库中每个句对与整个测试集的相似程度。

(3) 按相似程度分值对整个 Web 双语平行语料库进行排序。

4 实验

4.1 双语平行句对获取实验

目前针对双语混合网页的研究还比较少，微软提出的自适应模式学习的方法[8]有效的解决了从双语混合网页抽取平行句对的问题，取得了良好的效果。为了证明本文提出的方法同样有效并且具有更高的句对正确率和网页召回率，特将两种方法进行对比，结果如表 2 所示。

表 2　基于双语混合网页的平行语料挖掘方法对比

方法	候选网页数量	双语混合网页数量	网页召回率	平行句对数量	平行句对正确率
Microsoft	3 500 000 000	20 000 000	0.6%	7 522 803	83.5%
ICT	1 884 496	472 641	25.1%	2 580 117	93.75%

从表 2 可以看出，对比两种基于双语混合网页的平行句对挖掘方法，Microsoft 的方法在获取的平行句对总的数量上占优势，而本文提出的方法却在双语混合网页召回率和平行句对正确率上占优势。之所以前者能获取大量的平行句对是因为具有 35 亿的候选网页可从中筛选，但其混合网页的召回率只有 0.6%，这无疑会消耗大量的空间和时间成本，而本文提出的方法具有更高的存储利用率。此外，

本文提出的方法所抽取的平行句对正确率几乎比前者高出10%，显然我们的研究是有价值的。

4.2 平行句对排序实验

将258万双语平行句对按照评价函数 F 进行重排序，取前150万句对平均分成五组进行随机抽样，每组随机抽样500个句对，一共抽样2 500个句对，然后通过人工查验的方式统计正确率，前150万双语平行句对的平均正确率是96%，分组统计结果如表3所示。

表3 句对重排序后正确率统计

双语平行句对	正确率
1-300 000	98.6%
300 001-600 000	96.2%
600 001-900 000	95.8%
900 001-1 200 000	94.8%
1 200 001-1 500 000	94.6%

从分组统计结果可以看出，评价函数 F 的设置是合理的，经过打分重排序，可以将高质量的平行句对排在前面，将低质量的平行句对过滤掉，从而获取正确率更高的双语平行语料。

4.2 Web数据应用于SMT的实验

Web双语平行语料应用于统计机器翻译系的实验环境设置如下：解码器为著名的开源解码器moses(摩西)(http://www.statmt.org/moses/)，对齐工具使用的GIZA++(http://www.fjoch.com/GIZA++.html)，语言模型为四元，参数训练方法使用的是最小错误率训练，系统实现采用对数线性模型，机器翻译性能的评测标准为国际评测的通用标准BLEU[16]。

第一组实验，将用国际IWSLT评测的公用语料BTEC语料训练的翻译系统作为基准系统，按照平行句对打分重排序的方法将排序靠前的Web平行句对加入到训练集，以混合后的数据重新训练翻译系统。为了观察BLEU值的变化，我们按照指数级增加的方式分别加入2 000、4 000、8 000、16 000个Web双语平行句对到训练集，实验结果如表4中数据所示。

从表4可以看出，随着添加到训练集数据规模的增加，翻译系统的性能也随之提升。添加2 000个双语平行句对时效果提升最明显，主要因为这2 000个双语平行句对的翻译质量是最好的，打分排序最靠前，翻译正确率可以达到99%。随着添加数据规模的增大，系统性能的提升速度越来越缓慢，因为后面添加的数据的翻译质量要低于前面添加的数据，但是BLEU值都在稳步提升，实验结果证明按照打分重排序挑选出的Web双语平行句对是可以应用于统计机器翻译系统的模型训练的，效果比较理想。

表4 打分重排序方法挑选添加数据的实验

训练数据	05测试集上BLEU值	06测试集上BLEU值	07测试集上BLEU值
Baseline	0.400 9	0.146 2	0.216 9
Add2000	0.431 8	0.161 4	0.243 6
Add4000	0.432 1	0.167 4	0.246 9
Add8000	0.440 3	0.169 1	0.247 5
Add16000	0.453 4	0.172 7	0.248 8

第二组实验，将用BTEC语料训练的翻译系统作为基准系统，按照平行句对信息检索的方法将查询返回的与测试集相似的Web平行句对加入到训练集，以混合后的数据重新训练翻译系统。与第一组实验相同，也按照指数级增加的方式分别加入2 000、4 000、8 000、16 000个Web双语平行句对到训练集，实验结果如表5中数据所示。

表5 信息检索方法挑选添加数据的实验

训练数据	05测试集上BLEU值	06测试集上BLEU值	07测试集上BLEU值
Baseline	0.400 9	0.146 2	0.216 9
Add2000	0.412 3	0.155 5	0.219 6
Add4000	0.427 7	0.168 3	0.243 4
Add8000	0.453 0	0.171 0	0.251 7
Add16000	0.457 1	0.172 1	0.253 8

从表5可以看出，随着添加到训练集数据规模的增加，翻译系统的性能也随之提升，但BLEU值提升的趋势与实验一有所不同。第二组实验中，翻译系统性能的提升速度是比较平稳的，而第一组实验呈现先快后慢的趋势。因为第三组实验添加的句对翻译质量比较平均，而第二组实验添加的句对翻译质量是由高到低排序的。从实验二可以看出，用信息检索的方法对Web双语平行句对加以利用是有效的，因为IWSTL评测是旅游会话领域的语料，

所以相比第一组实验的 NIST 语料更能体现根据特定领域选取的语料对翻译效果的影响。

上述实验证明，我们提出的两种对于 Web 双语平行语料的利用方案是都是有效的，按照两种方案挑选出的数据加入统计机器翻译系统是可以提高翻译性能的。

5 小结与展望

双语平行语料库在自然语言处理领域有很多重要应用，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。而互联网作为广泛使用的信息载体，为我们提供了大量的双语候选资源。因此，本文提出一种基于双语混合网页的双语平行语料库自动获取方案，解决了候选资源获取、平行句对抽取等难点问题，运用该解决方案实际获取了百余万双语平行句对。为了有效利用 Web 数据，我们提出了两种应用策略，将从 Web 双语平行语料中挑选出的数据加入到统计机器翻译的模型训练，实验证明，我们提出的两种方案都可以提高翻译质量，可以使 Web 数据更好的服务于统计机器翻译的应用。

在以后的研究中，我们希望解决以下几个方面的工作：

第一，继续探索候选资源获取的解决方案，以期望能够快速、自动获取双语候选网站列表。

第二，构建更大规模更高对齐正确率的双语平行语料库，以供实际应用。

参考文献

[1] Peter F. Brown, John Cocke, Stephen A, et al.. A Statistical Approach to Machine Translation: Parameter Estimation[J]. Computational Linguistics, 1990, volume 16: 79-85.

[2] 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.

[3] 冯志伟. 中国语料库研究的历史与现状[J]. Journal of Chinese Language and Computing, 2002, 11(2): 127-136.

[4] Resnik, p. and N. A. Smith.. The web as a Parallel Corpus[J]. Comoutational Linguistics, 2003, volume 29: 349-380.

[5] 叶莎妮，吕雅娟，黄赟，等. 基于 Web 的双语平行句对自动获取[J]. 中文信息学报，2008，22(5)：67-73.

[6] Lei Shi, Cheng Niu, Ming Zhou,, et al. A DOM Tree Alignment Model for Mining Parallel Data from the Web[C]//Joint Pro-ceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics, Sydney, Australia, 2006: 489-496.

[7] Lei Shi, Ming Zhou: Improved Sentence Alignment on Parallel Web Pages Using a Stochastic Tree Alignment Model[C]//EMNLP, 2008: 505-513.

[8] Long Jiang, Shiquan Yang, Ming Zhou, et al. Mining Bilingual Data from the Web with Adaptively Learnt Patterns[C]//Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, 2009: 870-878.

[9] 林政，吕雅娟，刘群，等. 基于双语混和网页的平行语料挖掘[C]//全国第十届计算语言学会，烟台，2009：352-357.

[10] 刘非凡，赵军，徐波. 大规模非限定领域汉英双语语料库建设及句子对齐研究[C]//全国第七届计算语言学联合学术会议，哈尔滨，2003：339-345.

[11] Gale, William A. Kenneth W. Church. A program for aligning sentences in Bilingual corpora[J]. Computational Linguistics, 1993, 19 : 75-102.

[12] Stanley F. Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information[C]//Proceedings of the 31st Annual Meeting of the Association for Computational Linguaistics, 1993: 9-16.

[13] DeKai Wu. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria[C]//Proceedings of the 32nd Annual Conference of the Association for Computational Linguaistics, 1994: 80-87.

[14] T. Utsuro, H. Ikeda. Bilingual Text Matching using Bilingual Dictionary and Statistics[C]//15th COLING, 1994: 1076-1082.

[15] 张艳，柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报，2005，19(5)：31-36.

[16] Kishore Papineni, Salim Roukos, Todd Ward, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 311-318.