文章编号：

Web 双语语料挖掘综述

朱泽德[1,2]，李淼[2]，张健[2]，陈雷[2]，曾新华[2]，杨振新[1,2]

（1.中国科学技术大学，安徽 合肥 230026；2.中国科学院合肥智能机械研究所，安徽 合肥 230031）

摘 要：如何从互联网有效地挖掘双语语料是当前研究的热点问题。本文首先说明挖掘 Web 双语语料的具有重要的应用价值，阐述双语语料涉及的平行语料和可比语料的相关概念；接着介绍原始 Web 数据获取的基本概况，分别分析基于特征信息、跨语言信息检索和维基百科挖掘可比语料的研究现状，以及源自双语平行网页、单一双语网页和可比语料挖掘平行语料的研究进展；然后就双语语料质量的评价方法进行讨论；最后探讨该领域仍存在的问题和今后的研究方向。

关键词：互联网；Web 挖掘；双语语料；平行语料；可比语料

中图法分类号：TP391　　**文献标识码**:：A

A Survey of Mining Bilingual Corpora from the Web

Zhu Zede[1,2], Li Miao[2], Zhang Jian[2], Chen Lei[2], Zeng Xinhua[2], Yang Zhenxin[1,2]

(1. University of Science and Technology of China, Hefei, Anhui 230026, China;
2. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China)

Abstract: How to mine bilingual corpora from the web effectively has become a currently hot issue. Firstly, a brief introduction is related to the important value of mining bilingual corpora from web and its relevant concepts including parallel and comparable corpora. Then, this paper introduces the overview on acquiring original data from the web and summarizes respectively the research processes of comparable corpora utilizing feature information, cross-language information retrieval and hyperlink structure, together with parallel corpora mined from bilingual parallel web pages, single bilingual web pages and comparable corpora. Thirdly, the evaluation criteria of comparable corpora are also discussed. Finally, the paper discusses the existing problems and the future directions in the bilingual corpora mining field.

Key words: WWW; Web mining; Bilingual corpora; Parallel corpora; Comparable corpora

1　引言

双语语料库（Bilingual Corpora）极大地促进了跨语言信息处理技术的发展，在双语术语抽取、命名实体识别和翻译字典构建等双语知识获取中显现出重要作用，同时在统计机器翻译、跨语言信息检索、多语言文本处理、多语言类标签描述等领域突显基础资源的巨大价值。近年来双语语料库的自动构建研究十分活跃，一个显著的特征是将 Web 作为挖掘双语语料的潜在来源。

互联网的快速普及和多语言用户的迅速增长引发网络中多语言信息大量传播，为全面改善现有双语语料库在规模性、时效性和领域平衡性等方面的不足提供了重要途径。然而，网络资源有别于一般的文本资源，多具有半结构化、分散和异构的特点，既为挖掘双语语料提供了重要的判断依据，也构成影响语料质量和挖掘效率的不确定因素。

围绕将分散在互联网中不同语言的数据资源整合为可利用的双语语料，计算语言学领域的顶级国际会议 ACL、LREC、WWW 等连续举办了多届 WAC（Web as Corpus）和 BUCC（Building and Using Comparable Corpora）等研讨会，引起了人们对 Web 双语语料挖掘技术

基金项目：国家自然科学基金（61070099）、中国科学院信息化专项（XXH12504-1-10）

作者简介：朱泽德（1985--），男，博士生研究生，主要研究方向为自然语言处理与 WEB 信息挖掘；李淼（1955--），女，研究员，主要研究方向为人工智能与农业知识工程；张健（1954--），男，研究员，主要研究方向为人工智能与农业知识工程；陈雷（1981--），男，助理研究员，主要研究方向自然语言处理；曾新华（1976--），男，副研究员，主要研究方向为数据挖掘；杨振新（1990--），男，博士研究生，主要研究为自然语言处理。

的极大关注，有力推动了双语语料库构建的发展。

当前，自动挖掘 Web 双语语料的方法研究已取得了巨大的进步，但仍存在很多值得深入研究的问题。本文对 Web 双语语料挖掘技术作了系统归纳和总结，以便于后续研究的展开。文中其他部分安排如下：第二节阐述平行语料和可比语料的概念；第三节分析基于 Web 挖掘平行语料和可比语料的研究方法；第四节探讨如何评价双语语料库质量高低；最后就目前 Web 双语语料挖掘研究中存在的问题进行总结和未来的发展进行展望。

2 双语语料的概念

根据互译程度的差异，双语语料库可分为平行语料库（Parallel Corpora）和可比语料库（Comparable Corpora）。平行语料库是由源语言文本和翻译的目标语言文本构成文本对集合，两种语言文本间存在严格的互译关系[1]。由于互译粒度不同，平行语料库有篇章、段落、句子、短语和单词等多级别，本文讨论的平行语料特指可直接从互联网获取的篇章对或句对，不涉及其他粒度的探讨。网络中分布的双语资源多呈现形式复杂和内容不平衡的特点，提高了平行语料库获取的难度。

可比语料库是语言不同、内容相似但非互译的文本对集合，可比语料蕴含了三层意义：源语言和目标语言文本独立产生于各自语言环境[2][3]；源语言和目标语言文本具有一定的相似性，主要体现在采集标准[4]、结构标准（文档结构、文档长度）、语言标准（隶属领域[5]、涵盖主题[6]、文本类型）、语料功能[3]等方面；源语言和目标语言文本不具备严格的互译关系[1][5][7]。相似性是可比语料最显著的特点，可比语料因相似程度不同分如下五个等级[8]。

等级 1 Same Story：两篇文本描述同一个话题；

等级 2 Related Story：两篇文本从不同角度描述相同事件或存在包含关系；

等级 3 Shared Aspect：两篇文本报道相关或类似事件；

等级 4 Common Terminology：两篇文本描述不相关事件，但存在大量共同术语；

等级 5 Unrelated：两篇文本有较低或不存在相似性。

可比语料仅要求内容相似性，降低了源语言和目标语言文本匹配条件，相对平行语料具有来源广阔、领域全面、内容新颖[4]和易于获取[3][6]的优势。

由上述分析可知，平行语料与可比语料虽同属于双语语料的范畴，但存在以下三方面区别：

1）语言真实性，平行语料包含原文和译文，译文质量受限于原文质量和翻译水平；可比语料的文本均为独立产生的原文，语言的真实性强于平行语料。

2）互译等价性，平行语料约束源语言与目标语言文本严格互译；可比语料仅要求源语言与目标语言文本具有相似性，较平行语料表现的互译等价性弱。

3）获取难易性，Web 平行语料多源自双语网站，双语文本数量和语种受限；可比语料的源语言和目标语言文本来自不同网站，双语文本数量大且更新速度快，较平行语料更易获取。

3 Web 双语语料挖掘

本节先介绍双语语料挖掘的前提——原始 Web 数据获取；再阐述双语语料挖掘方法，因可比语料可作为挖掘平行语料的资源，为使内容前后照应，可比语料挖掘方法先平行语料阐述。

3.1 Web 数据获取

Web 数据获取主要涉及网页采集、编码规范、语种识别和正文抽取，此部分仅简要介绍通用工具，不作为讨论的重点。

1）网页采集

为采集互联网原始网页，Heritrix[9]是一种开源可扩展的网络爬虫；BootCat[10]获取网络中特定语言文本，通过预定义种子词表查询 Yahoo!搜索引擎，下载网页并移除模板；

Combine[11]结合了语言识别和话题自动分类实现开源网络爬虫；HTTrack[12]实现了跟踪HTML 页面的链接依次下载远程网站，并重构原站点目录结构；Linux/Unix 命令 Wget 实现的功能与 HTTrack 基本相同。

2）编码规范

为解决网页文档编码不统一问题，Linux/Unix 平台提供了以下命令进行操作：File 可检测文档编码；Iconv 在已知编码下转换文档编码；Enca 可检测并转换文档编码。

3）语种识别

为判定网页语言的种类，LibtextCat①利用分类技术识别能 69 种语言；Lingua:Identify②可进行参数设定并能判定 33 种语言；Lextek③识别语种数达 260 种。

4）正文抽取

为获得网页正文内容，过滤广告、导航等噪声，Victor④采用条件随机场标记网页为正文内容或模板序列；Boilerpipe⑤根据文字数量和链接密度分类网页元素，检测并清除网页模板；NCleaner⑥利用分类器依据字符级的 N-gram 模型识别正文。

3.2 可比语料挖掘

Web 可比语料挖掘的核心是基于特征信息、跨语言信息检索和维基百科建立源语言与目标语言文本相似性的映射。

3.2.1 基于特征信息

特征信息是伴随内容相似所呈现出的互译词汇或共现信息（Co-occurrence Information），研究者利用特征信息筛选可比语料促进了可比语料库建设初期的发展。最早的可比语料库是由 Sheridan 等[13]利用描述性字段和新闻发布日期构建的德文-意大利文集合；Aker 等[14]利用标题信息实现简便地收集可比语料库，主要依据网页搜索和 RSS 新闻订阅获取新闻标题内容、标题长度、出版时间和日期计算双语文档相似性。此类方法选择相似文本严重依赖共现信息，可比语料的匹配速度快但质量较低。

Tao 等[15]依据相同时间段内不同语言描述的事件应基本一致的假设，提出不同语言词汇在一定时间段内词频分布越相似则互译或描述同一主题概率越大来发现源语言与目标语言词汇的相关性，并通过皮尔森相关系数（Pearson Correlation Coefficient）实现词汇相关性计算，再结合词汇频率和反文档频率计算源语言与目标语言文本的相似性。Thuy 等[16]根据新闻发表时间窗与标题&内容翻译词过滤部分文档对；再提取标题与内容的互译词（Title-n-Content，TNC）、语言无关单元（Linguistic Independent Unit，LIU）和术语词频分布（Monolingual Term Distribution，MTD）计算文本相似性（见图 1），其中术语词频分布采用离散傅里叶变换（Discrete Fourier Transform）获得。

利用内容互译提高了可比语料的质量，但降低了算法的执行效率。词频分布计算词汇相关性避免对双语词典和双语句对等外部资源的依赖，因需要计算每个词在一定时间段内的频度分布，构建过程中时间消耗巨大，不适合大规模可比较语料库的发展。研究多将共现信息和内容互译结合：共现信息过滤部分非相似文档，内容分析筛选出可比语料。

---

① http://software.wise-guys.nl/libtextcat/

② http://search.cpan.org/~ambs/Lingua-Identify-0.26/

③ http://www.lextek.com/langid/

④ http://ufal.mff.cuni.cz/victor/

⑤ http://code.google.com/p/boilerpipe/

⑥ http://webascorpus.sourceforge.net/

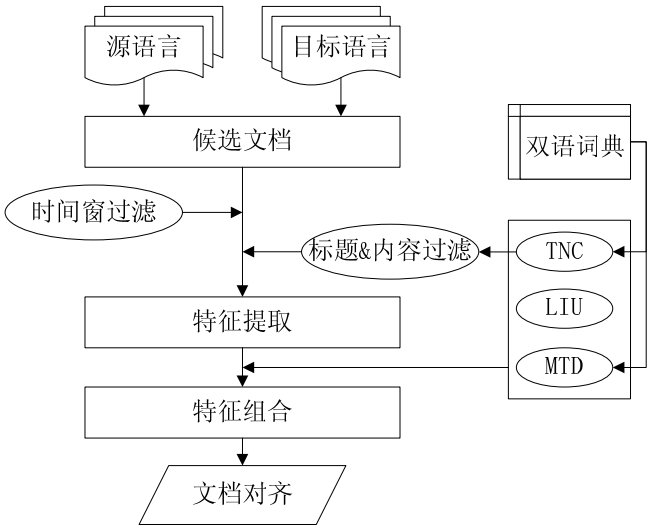

图 1 基于特征的文档对齐

3.2.2 基于跨语言信息检索

跨语言信息检索被广泛应用于可比语料挖掘，其中提问式翻译能高效地实现源语言与目标语言相似文档的映射[17]，如图 2 流程所示。

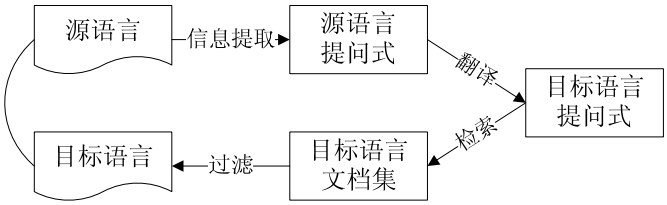

图 2 基于跨语言信息检索创建可比语料

Braschler 等[8]将英文中频词汇翻译成德文词汇，检索德文文档集后根据专有名词、数字和发表日期等信息匹配主题覆盖重叠（Coverage Overlap）较高的英文-德文相似文本。Munteanu 等[18][19]取源语言文档中每个词汇最可能的 5 个翻译作为目标语言查询词检索目标语言文档，选择与源语言文档发布日期相差 5 天的文档返回。为提高查询词对源语言文档的代表性，Huang 等[20]抽取源语言文档的关键词翻译成目标语言查询词。Fiser 等[21]在保持原始语料基础上引入外部资源扩充语料，具体是先从医疗卫生杂志中采集斯洛文尼亚文和英文作为初始可比语料，再计算网络文档与初始文档集合的相似度，逐步加入较高相似度的网络语料。

可比语料库是篇章对应的文本对集合，因文档篇幅有限，2 种语言文档的多数词汇不存在互译等价对（Translational Equivalence）。为避免挖掘更细粒度的互译等价对出现数据稀疏问题，跨语言信息检索被用于聚类可比语料。Talvensaari[22]构建了英语-西班牙语-德语可比语料库，通过 Google 获取每种语言的领域词表，利用主题爬虫（Topic Crawler）采集可比语料，跨语言信息检索匹配源语言和目标语言文本的共现信息。Leturia 等[23]通过两种方法获取领域词表，其一，收集不同语种的领域语料，从中抽取关键词作为查询词；其二，搜集一种语言的领域语料抽取关键词，利用词典翻译获得目标语言的查询词。

跨语言信息检索方法提高大规模可比语料库构建速度，其中，查询词的选择至关重要，要求源语言查询词能全面覆盖文档主题又不发生主题漂移，同时要求源语言查询词准确翻译成目标语言查询词，保持目标语言查询词与源查询词具有相同的主题信息。

3.2.3 基于维基百科

维基百科是一种自由、免费、开放的多语言百科全书，作为可比语料重要的新型来源，受到越来越多研究者的关注[24]。

挖掘维基百科可比语料主要有两种方式，第一，先从维基百科中下载不同语种的数据，然后使用语言间链接进行双语匹配，即“先下载，再匹配”。Yu 等[25]采用此方案获取可比语料。Otero 等[26]以维基百科类别信息作为主题约束，以语言链接进行双语映射，实现考古领域可比语料挖掘。第二，先收集词表，获取页面标题中含有词表中词语的单一语种页面，再使用语间链接采集其他语种维基百科页面，即“先匹配，再下载”。Ion 等[27]从 WordNet 中抽取命名实体，下载含有这些命名实体的英语维基百科页面；然后挖掘语言链接结构，采集对应的罗马尼亚语和德语页面构成可比语料。Li 等[28]使用跨语言评测论坛（Cross-Language Evaluation Forum）的语料 GH95 和 SDA95 作为初始语料，选择 LAT94、MON94、SDA94 和 Wikipedia 数据进行扩展。

从维基百科多语言资源中挖掘可比语料主要利用网页间的语言链接结构。链接的相关性简化了内容的相似性分析，降低了算法执行的复杂度并提高了可比语料质量。然而维基百科的语言种类和信息量有限，无法满足构建多语种、大规模可比语料库的需求。

3.3 平行语料挖掘

平行语料对双语文本互译性的严格要求，导致仅能从特定的来源获取，本节分别阐述从双语平行网页、单一双语网页[29]和可比语料挖掘平行语料的研究进展。

3.3.1 源自双语平行网页

双语平行网页识别是平行语料挖掘的核心，基于结构信息和内容信息是两个重要判断依据。

3.3.1.1 基于结构信息

网页结构信息是平行网页有别于其他双语语料来源的显著特征，其包含网页 URL 命名语言相关性（如 File_e.html 和 file_c.html）和网页 HTML 结构对称性。

Resnik 等[30][31]最早利用平行网页的结构信息开发了 Web 平行语料挖掘系统 STRAND（Structural Translation Recognition for Acquiring Natural Data）。实现过程是先利用搜索引擎和“中文版”等启发式信息获取双语候选网站；再采用 URL 命名相关性匹配查找候选平行网页；最后通过文本长度、HTML 结构（见图 3）和网页语言等特征过滤伪平行网页；并在 HTML 对齐的基础上抽取 HTML 标记间的对应文本获取双语句对齐语料。

[START:HTML]	[START:HTML]
[START:TITLE]	[START:TITLE]
[Chunk:13]	[Chunk:15]
[END:TITLE]	[END:TITLE]
[START:BODY]	[START:BODY]
[START:H1]	
[Chunk:13]	
[END:H1]	
[Chunk:112]	[Chunk:112]

图 3 HTML 标签序列结构对齐

STRAND 的方法被后来研究者广泛采用，Chen 等[32]开发的系统 PTMiner（Parallel Text Miner）与 STRAND 十分类似，差异在处理 URL 命名的语言相关性时，前者是替换预定义字符串，后者是删除预定义字符串。不同于 HTML 标记序列匹配相似网页，Shi 等[33][34]提出 DOM 树模式对齐 HTML 结构（见图 4），再以链接中相似网页为种子迭代发现新的候选平行网页。

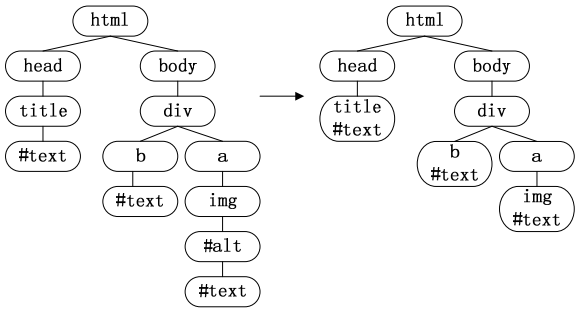

图 4 Dom 树模式转换

研究者针对网页 URL 命名的语言相关性开展了进一步分析。Zhang 等[35]在系统 WPDE（Web Parallel Data Extraction）中增加了图片的 ALT 信息搜索候选网站，重点将 URL 划分为 Pathname 和 Basename 分别匹配语言的相关性，Pathname 匹配是利用预先定义的启发式字符串和匹配规则；Basename 匹配则采用最小编辑距离。为避免人工预定义 URL 语言标识发生遗漏，叶莎妮等[36]通过 URL 中语言无关的相同部分 S 与语言相关的不同部分 Lang 划分（见图 5），提出自动发现双语网站中 URL 命名规律的方法。

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

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

图 5 URL 划分 S 与 Lang

基于结构识别平行网页无需双语知识，可快速挖掘任意语言对的平行网页，促进了 Web 平行语料的挖掘效率。然而，结构对齐信息仅存在于特定双语网站，也非平行语料的内在特征，不能作为判断网页平行的决定性因素。

3.3.1.2 基于内容信息

源语言和目标语言文档的词汇互译是平行网页最根本和最直观的标准，Ma[37]提出以互译词汇占文本总词汇比例作为平行网页相似度判定依据，然而文档集中不同语言两两文档计算词汇互译率导致算法运行效率过低，同时受限于双语词典的规模。

针对运行效率过低的不足，一种方法是选择特定词汇代替所有词汇降低文档对比的代价，Enright 等[38]选择低频词互译率作为平行文档判断依据。另一种方法是通过预处理措施降低文档词汇比对搜索次数，Chen 等[39]开发的系统 PTI（The Parallel Text Identification System）先根据 URL 命名相关性先识别部分平行网页，再利用互译率查询剩余网页集中平行网页。Sakre 等[40]借助平行文本非语言信息的对应关系过滤伪平行网页。Patry 等[41]使用信息检索技术降低源语言和目标语言文本匹配的复杂性，将数字实体和 Hapax 词（blank separated strings of more than 4 characters）确认平行网页。Zhu 等[42]提出以文本长度、文本句数、数字序列为主要信息，以 HTML 序列为补充信息，使用 SVM 分类判定网页对的平行性。为解决内容信息过度依赖互译知识的缺陷，已有双语语料资源被研究者所采用，王洪俊等[43]提出将统计翻译模型替代双语知识计算两文档的相似度。Antonova 等[44]使用统计概率词典实现从可比文档中抽取句对齐语料。

平行网页识别从结构分析到内容分析，提高挖掘方法应对网页布局变化的性能。上述研究以平行网页词汇互译为基础，重点解决词汇互译查询效率偏低和双语知识不足的制约。如何发现更多平行文档相互关联的信息以提高平行网页识别效果有待进一步探究。

3.3.2 源自单一双语网页

单一双语网页是指平行文本存在于一个网页内，较双语平行网页具有双语对照更整齐和翻译质量更好的优点。因单一双语网页的出现滞后，且可挖掘平行语料数量和语言种类相对

较少，目前开展的相关研究偏少。

Jiang 等[29]提出自适应模式学习的方法抽取平行语料，利用翻译和音译模型找到网页中的翻译词对作为种子，再使用种子学习泛化模板，最后根据学习的模板获取所有的平行语料。林政等[45]采用尝试下载策略发现单一双语网页，利用互译信息确认网页，再根据长度、词典、标点符号和数字、缩略词等信息抽取平行句对。Mohler 等[46]为获取低密度语言（Low-density Languages）的平行文本，利用小语种词表查询 Google 引擎收集原始网页，再采用句对齐方法获取平行语料。

3.3.3 源自可比语料库

双语平行网页和单一双语网页是挖掘互联网双语语料的直接来源。然而，此类双语资源数量有限、多集中于少数大语种。可比语料拓展了平行语料挖掘的空间，使平行语料的来源不再局限于篇章对齐文本。

Munteanu 等[19]使用最大熵分类器（Maximum Entropy Classifier）实现中文、阿拉伯文和英文可比语料库句对划分为平行或非平行。Wu 等[47]开发了反向转换文法（Inversion Transduction Grammar）实现从大量非平行文档中检索平行句对。Utiyama 等[48]抽取日英可比专利中平行句对时，发现专利的实施方案详细描述和发明背景有较多的词语翻译。LU 等[49]联合句对齐工具 Champollion[50]和 MS aligner[51]，使用句长度和长度比信息过滤非平行句对。Rauf 等[52]使用统计机器翻译系统翻译源语言句子成为目标语言句子，然后使用 WER、TER 和 TERp 等机器翻译评测算法检测翻译句子和目标句子相似性。Ion 等[53]将特定翻译对作为对齐标识利用最大期望（Expectation Maximization）发现平行语料。Ştefănescu[54]评价句对使用 5 个词汇重叠和结构匹配特征，利用 Logistic 回归分类实现实词翻译、功能词翻译、对齐词位置、句首尾实词翻译、句标点一致等参数的最优化。Quirk 等[55]认为目标句能通过原句有条件的产生，提出通过生成模型（Generative Models）检测源语言和目标语言句子的相似性，模型参数采用源词和目标词以及源词的位置。

4 双语语料评价方法

平行语料的评价是假定源语言和目标语言文本为互译或非互译，根据准确率和召回率检测算法性能。针对平行语料的对齐程度探讨较少，Cartoni 等[56]度量不同子语料的相似度进而衡量平行语料的同质性（Homogeneity），通过计算词汇的 χ^2 值并融合 Since、Because 等原因连接词的分布情况综合考察平行语料的对齐性。研究者对可比语料相似性的定义并非完全一致[57]，对可比语料质量评价进行了大量的探索。

早期研究多从定性角度讨论可比语料匹配质量的高低，Maia 等[3]将其分为两类指标阐述，在相似度方面，考虑的标准有形式和内容、结构和功能、语料模式等；在可比度方面，考虑语料的规模、构建原则、语料类型、语料领域等指标。Braschler 等[8]将可比较语料库的相似性分为如节 1 中的 5 个等级。Fung[58]定义了 3 等语料可比度，噪声平行语料（Noisy Parallel Corpus）：包含很多平行句对，但非全部对齐；可比语料（Comparable Corpus）：相同主题但非彼此翻译；准可比语料（Quasi-comparable Corpus）：文档部分探讨相同话题。Gliozzo 等[59]认为两个语料词表中共同词语所占的比例可作为语料可比度的一个标准。Munteanu[60]根据可比语料文档含有的对齐数据的多少，将语料分为完全对齐、噪声对齐和非对齐。

在定性评价双语语料的基础上，定量分析工作也被逐步展开。Kilgarriff 等[61]从对比语料库中抽取 Top-N 高频词汇进行卡方统计（Chi-square Statistic）确定相似性。Leturia 等[23]计算语料关键词的统计值衡量领域语料的可比度；文章先使用双语词典进行翻译映射，再通过关键词的余弦相似度或最频繁关键词的 χ^2 值计算语料的可比度。Vasiljevs[62]采用对称方法计算词语的 χ^2 值，先将语料 A 中词语的相对频率视为期望值，其翻译在语料 B 出现频率作为观测值；再互换 A、B 重新产生新的 χ^2 值，避免单一方向出现的偏差。Li 等[57]计算语料

中每个词翻译的期望值（Expectation）来衡量可比度。Su 等[63]采用与 Fung 类似的平行文档、强可比文档、弱可比文档 3 类指标；在机器翻译基础上，综合检测词汇特征、结构特征、关键词特征和命名实体特征实现对语料的评价。Denoual[64]计算语料与参考语料的相似度来评价可比性。不同于基于词汇级对比的分析，Sharoff[65]利用聚类和主题模型等分析语料，从更高层面研究语料间的差异性。Saralegi 等[66]融合领域、报道类型、主题分布和发表日期等估计新闻语料的总体可比度。

双语语料评价方法的出发点是检测源语言和目标语言文本的匹配程度。从定性分析到定量衡量、从词汇级到主题级，主题相似性进一步选择和过滤低可比度的双语语料，为双语语料挖掘提供新的思路。可比语料要求源语言和目标语言各自产生于独立的语言环境，目前的评价方法缺乏针对语言真实性的检验。

5 总结与展望

跨语言信息处理技术的发展迫切需求双语语料资源的支持，国内外研究者围绕 Web 挖掘双语语料开展了大量工作。在可比语料和平行语料挖掘中有很多挖掘方法可相互借鉴，如可比语料挖掘中新闻发表时间和跨语言信息检索可作为平行语料挖掘的预过滤措施；平行语料挖掘使用已有的翻译模型可作为可比语料匹配的双语知识。然而，原始资源存在形式的差异导致各自具有特定的挖掘方法，如维基百科中语言链接是可比语料特有的信息；平行网页 URL 命名相关性和 HTML 结构相似性是平行语料特有的信息。然而双语语料挖掘在执行效率、外部资源依赖和文档内容分析等仍有较多问题需要进一步解决：

1）在可比语料挖掘方面，基于维基百科获取可比语料是当前的研究热点，但维基百科知识的语种不足，不能满足获取所有语种可比语料的需求，特别是低密度语言可比语料；维基百科中高密度语言（High-density Languages）的信息量也有限，无法实现大规模、多领域的可比语料库构建。基于特征信息和跨语言信息检索匹配双语相似文档是从词汇及相关外部特征层面出发，少有从主题的角度来开展可比语料构建技术研究[8][59]。Preiss[67]首次将文档隐含主题分析用于构建可比语料，文中将源语言主题模型翻译成目标语言模型后计算主题的相似性，词汇翻译的歧义性制约了可比语料质量。利用文档主题信息挖掘可比语料的研究需深入开展。

2）在平行语料挖掘方面，平行网页识别从分析结构到分析内容，避免了对平行网页 URL 命名相似性和 HTML 结构对称性的假设，反映了平行网页的本质要求。然而，伴随内容分析出现了两个问题：a）过度依赖双语资源，解决双语知识严重匮乏现有的方法是假定某时间段内不同语言描述的事件基本一致，词汇分布相同表征词汇存在某种相关性[15]。此方法是基于特定的数据源，且算法的时间复杂度过高。b）互译词汇搜索耗费大量时间，Enright 和 Patry 等选择特定词汇进行查询篇章的互译性。如何选择最能体现篇章内容的关键词语，如何依据上下文信息有效降低互译词汇的歧义，保证准确率同时降低搜索复杂度，这方面的工作仍需研究者开展。

上述问题的存在，使挖掘的双语语料与实际的应用相距较远。互联网双语资源多分布于独立更新的不同网站，源语言与目标语言文本常报道相同事件，语言的真实性强，具有一定的相似性但非绝对相互翻译。Web 双语资源更适宜于挖掘可比语料，然而，平行语料对互译性的严格要求有利于挖掘更细粒度的双语知识。因此，构建可比语料库和平行语料库都十分迫切，且需公开的双语语料用于开展算法性能评测，推动大规模语料挖掘技术的发展。

参考文献

[1] Baker M. Corpora in Translation Studies: An Over view and Some Suggestions for Future Research [J]. Target, 1995, 7(2): 223-243.

[2] MCENERY T. Multilingual corpora-current practice and future trends[C]. Proceedings of the 19th ASLIB Machine Translation Conference, London, England, 1997:71-83.

[3] Belinda Maia. What are Comparable Corpora? [EB/OL] [2010-12-28]. http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

[4] skadina l, Aker A, Giouli V, et al. A Collection of Comparable Corpora for Under-resourced Languages [C].

Proceedings of HLT2010. Riga. Latvia. 2010:161-168.
[5] McEnery A, Xiao R. Parallel and comparable corpora: What are they up to? [C]. Proceedings of Incorporating Corpora: Translation and the Linguist Translating Europe Multilingual Matters. Clevedon, UK. 2007.
[6] Ji H. Mining name translations from comparable corpora by creating bilingual information networks[C]. Proceedings of BUCC 2009. Suntec, Singapore, 2009: 34-37.
[7] Lynne Bowker, Jennifer Pearson. Working with Specialized Language: A Practical Guide to Using Corpora [M]. London/New York: Routeledge, 2002.
[8] Braschler M, Schauble P. Multilingual Information Retrieval based on document alignment techniques[C]. Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries. Heraklion, Greece.1998: 183-197.
[9] Mohr G, Stack M, Ranitovic, I., et al.. An Introduction to Heritrix[C]. Proceedings of 4th International Web Archiving Workshop. 2004.
[10] Baroni M, Bernardini, S. BootCaT: Bootstrapping corpora and terms from the web[C]. Proceedings of LREC 2004:1313-1316.
[11] Ardo A. 2005. Combine web crawler, Software package for general and focused Web-crawling, http://combine.it.lth.se/.
[12] Roche, X. 2007. http://www.httrack.com.
[13] Sheridan P, Ballerini J P. Experiments in multilingual information retrieval using the SPIDER system [C]. Proceedings of the 19th ACMSIGIR conference. Zurich, Switzerland. 1996: 58-65.
[14] Aker A, Kanoulasy E, Gaizauskas R. A light way to collect comparable corpora from the Web, 2012.
[15] Tao T, Zhai C X. Mining comparable bilingual text corpora for cross-language information integration[C]. Proceedings of ACM SIGKDD, Chicago, Illinois, USA. 2005:691-696.
[16] Thuy Vu, Ai Ti Aw, Zhang M. Feature-based method for document alignment in comparable news corpora[C]. Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece. 2009: 843-851.
[17] OARD D W, DIEKEMA A R. Cross-Language Information Retrieval [J]. Annual Review of Information Science and Technology, 1998, 33: 223-256
[18] Munteanu D S, Fraser A, Marcu D. Improved machine translation performance via parallel sentence extraction from comparable corpora. HLT-NAACL, 2004: 265–272.
[19] Munteanu D S. Marcu D. Improving machine translation performance by exploiting non-parallel corpora[J]. Computational Linguistics, 2005, 31(4):477–504.
[20] Huang D G, Zhao L, Li L S, et al. Mining Large-scale Comparable Corpora from Chinese-English News Collections[C]. Proceedings of Coling 2010, Beijing China. 2010:472-480.
[21] Fiser D, Ljubesic N. Building and using comparable corpora for domain-specific bilingual lexicon extraction[C]. Proceedings of BUCC 2011. Portland, Oregon. 2011: 19-26.
[22] Talvensaari T, Pirkola A, Jarvelin K, et al. Focused web crawling in the acquisition of comparable corpora [J]. Information Retrieval. 2008. 11(5): 427-445.
[23] Leturia l, Vicente I S, Saralegi X. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet [C]. Proceedings of the WAC5. Basque Country, Spain. 2009: 53-61.
[24] Smith J R, Quirk C, Toutanova K. Extracting Parallel sentences from comparable corpora using document level alignment[C]. Proceeding of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. Los Angeles, California, US. 2010: 403-411.
[25] Yu K, Tsujii J. Bilingual dictionary extraction from Wikipedia[C]. Proceeding of MT Summit Xll. Ottawa, Canada, 2009.
[26] Otero P G, L'opez I G. Wikipedia as Multilingual Source of Comparable Corpora[C]. Proceedings of the 3rd Workshop on BUCC, LREC2010. Malta. 2010: 21-25.
[27] Ion R, Tufis D, Boros T, et al. On-Line Compilation of Comparable Corpora and their Evaluation[C]. Proceedings of the FASSBL 2010. Dubrovnik, Croatia. 2010:29-34.
[28] LI B, Gaussier E, Aizawa A. Clustering Comparable Corpora for Bilingual Lexicon Extraction [C]. Proceedings of the 49th ACL. Portland, Oregon. 2011: 473-478.
[29] Jiang L, Yang S, Zhou M, et al. Mining Bilingual Data from the Web with Adaptively Learnt Patterns[C]. Proceedings of 47th ACL, 2009: 870-878.
[30] 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, pages 72-82, 1998.
[31] Philip Resnik, Noah A. Smith. The Web as a parallel corpus[J]. Computational Linguistics, 2003, 29: 349-380.
[32] Chen J, Nie J Y. 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.
[33] Lei Shi, Cheng Niu, Ming Zhou, et al. A DOM Tree Alignment Model for Mining Parallel Data from the Web[C]. ACL, 2006.
[34] Lei Shi, Ming Zhou. ImProved Sentenee Alignment on Parallel Web Pages Using a Stochastie Tree Alignment Model [C]. EMNLP, 2008: 505-513
[35] Zhang Y, Wu K, Gao J F, et al. Automatic acquisition of chinese-english parallel corpus from the web. Proceedings of ECIR-06, ACL, 2006.
[36] 叶莎妮, 吕雅娟, 黄赟, 刘群. 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.
YE Shani, LV Yajuan, HUANG Yun, LIU Qun. Automatic Parallel Sentences Extraction from Web[J]. Journal

of Chinese Information Processing, 2008, 22(5): 67-73.
[37] Xiaoyi Ma, Mark Y. Liberman. Bits: A method for bilingual text search over the Web[C]. Proceedings of the Machine Translation Summit VII, 1999.
[38] Jessica Enright, Grzegorz Kondrak. A Fast Method for Parallel Document Identification[C]. Proceedings of NAACL HLT 2007, Rochester, NY, April 2007:29–32.
[39] Jisong Chen, Rowena Chau, Chung-Hsing Yeh. Discovering parallel text from the World Wide WEB [C]. Proceedings of CRPIT'32. Australia, 2004:157-161.
[40] Mohammed M. Sakre, Mohammed M. Kouta, AliM. N. Allam. AUTOMATED CONSTRUCTION OF ARABIC-ENGLISH PARALLEL CORPUS[J]. Computer Science and Network Security, 2009.
[41] Alexandre Patry, Philippe Langlais. Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia[C]. Proceedings of BUCC, ACL, Portland, Oregon, 2011:87-95.
[42] Zede Zhu, Miao Li, Lei Chen, et
al. Automatic Construction of Chinese-Mongolian Parallel Corpora from the Web Based on the New Heuristic Information[C]. IALP 2011.
[43] 王洪俊, 施水才, 俞士汶, 肖诗斌. 跨语言相似文档检索[J]. 中文信息学报, 2007, 21(1): 30-37.
WANG Hongjun, SHI Shuicai, YU Shiwen, XIAO Shibin. Cross-language Similar Document Retrieval [J]. Journal of Chinese Information Processing, 2007, 21(1): 30-37.
[44] Alexandra Antonova, Alexey Misyurev. Building a Web-based parallel corpus and filtering out machine translated text[C]. Proceedings of BUCC, ACL, Portland, Oregon, 2011:87-95.
[45] 林政, 吕雅娟, 刘群, 马希荣. Web 平行语料挖掘及其在机器翻译中的应用[J]. 中文信息学报,2010,24(5):85-91.
LIN Zheng, LV Yajuan, LIU Qun, MA Xirong. Mining Parallel Corpora from Web and Its Application in Machine Translation[J]. Journal of Chinese Information Processing, 2010, 24(5):85-91.
[46] Michael Mohler, Rada Mihalcea. BABYLON Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages[C]. Proceedings of the BUCC, ACL, 2010.
[47] Wu D, Fung P. Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora. Proceedings of IJCNLP 2005.
[48] Utiyama M., Isahara H. A Japanese-English patent parallel corpus[C]. Proceeding of MT Summit XI. 2007:475–482.
[49] Bin LU, Tao JIANG, Kapo CHOW, et al. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT[C]. Proceedings of BUCC, LREC 2010, Malta, 2010:42–49.
[50] Ma, X. Champollion: A robust parallel text sentence aligner[C]. Proceedings of LREC 2006. Genova, Italy.
[51] Moore R C. Fast and accurate sentence alignment of bilingual corpora[C]. Proceedings of AMTA. 2002:135-144.
[52] Rauf S A. Schwenk H. Parallel sentence generation from comparable corpora for improved SMT. Machine Translation, 25, 2011: 341-375.
[53] Ion R. Ceauşu A. Irimia E. An Expectation Maximization Algorithm for Textual Unit Alignment[C]. Proceedings of BUCC, ACL. Portland, Oregon, USA, 2011:128-135.
[54] Dan Ştefănescu, Radu Ion, Sabine Hunsicker. Hybrid parallel sentence mining from comparable corpora[C]. Proceedings of EAMT 2012, Trento, Italy.
[55] Quirk C, Udupa R, Menezes, A. Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction[C]. Proceedings of the MT Summit XI, Copenhagen, Demark, September, 2007: 321-327.
[56] Cartoni B, Zufferey S, Meyer T, Popescu-Belis A. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives[C]. Proceedings of BUCC 2011. Portland, Oregon, 2011: 78-86.
[57] LI B, Gaussier E. Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora[C]. Proceedings of Coiling 2010. Beijing, China. 2010: 644-652
[58] Fung P, Cheung P. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus[C]. Proceedings of COLING 2004, Geneva, Switzerland.
[59] Gliozzo A, Strapparava C. Cross language text categorization by acquiring multilingual domain models from comparable corpora[C]. Proceedings of the ACL. Morristown, NJ, USA. 2005: 9-16.
[60] Munteanu D S. Exploiting Comparable Corpora [D]. Information Sciences Institute, University of Southern California, USA, 2006.
[61] Kilgarriff A, Rose T. Measures for corpus similarity and homogeneity[C]. Proceedings of EMNLP 1998, Granada, Spain.
[62] [63] Vasiljevs A. ACCURAT: Metrics for the evaluation of comparability of multilingual corpora[C]. Proceedings of the workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods, LREC2010. Malta, 2010.
[63] Denoual E. A method to quantify corpus similarity and its application to quantifying the degree of literality in a document [J]. International Journal of Technology and Human. 2006.
[64] Su F Z, Babych B. Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents[C]. Proceedings of the 13th Conference of the European Chapter of the ACL, Avignon, France, 2012.
[65] Sharoff S. Analyzing Similarities and Differences between Corpora[C]. Proceedings of Language Technologies. Ljubljana, Slovenia. 2010.

[66] Saralegi X, Vicente I S, Gurrutxaga A. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain[C]. Proceedings of BUCC, LREC. Marrakech, Morocco. 2008: 27-32.
[67] Judita Preiss. Identifying Comparable Corpora Using LDA[C]. 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montre´al, Canada, 2012:558-562.