DOI:10.16515/j.cnki.32-1745/c.2021.04.008

一种面向Web的英汉平行语料库的构建方法

徐润华[1],王东波[2]

(1.金陵科技学院人文学院,江苏 南京 210038;2.南京农业大学信息管理学院,江苏 南京 210095)

摘 要:随着自然语言处理领域各项研究的发展,平行语料库作为支撑自然语言处理技术的基础资源,发挥着越来越重要的作用。利用Web中的海量信息资源,采取信息抽取的方法,自动获取英汉双语平行语料资源。在获取过程中,首先确定抓取网站和制定词表,然后利用网络资源抓取工具GUN Wget自动获取网页中的英汉双语句子对资源,在对获得的平行句子对资源进行清洗和去重的基础上,利用条件随机场模型对汉语句子进行自动分词并导入数据库,最终完成大规模英汉双语平行语料库的构建。

关键词:平行语料库;GUN Wget软件;条件随机场;英汉双语;Web

中图分类号:H146;TP391 **文献标识码**:A **文章编号**:1673-131X(2021)04-0051-06

随着信息化时代的到来,基于Web的信息获取技术得到迅速发展,从拥有海量信息资源的互联网中获取有效的数据和知识以便更好地服务基础和应用研究成为一种趋势。英汉双语句子级平行语料库的构建有助于跨语言检索自动衍生英汉双语词典和潜语义自动标注,可以为辅助机器翻译和机器翻译系统的开发提供基本语法、语义和语用素材,也有助于英汉双语词典编纂者选取例证和确定词目。利用Web资源可以获取大量非结构化和结构化的英汉双语句子信息,从而达到构建更高质量英汉平行语料库的目的。

一、研究综述

传统的语料库加工主要依靠人工,费时费力且不便于数据更新。在搜索引擎和数据挖掘技术的推动下,利用互联网Web的双语平行资源自动构建语料库的方法逐渐引起学者们的注意。黄苏豪[1]提出一种利用互联网自动构建英汉平行语料库的系统设计及实现方法。该方法优化了双语网页对文本的句子抽取和对齐的效果,并在此基础上开发完成了语料检索平台,实现了检索服务。但是该平行语料库未对收录资源进行深度加工,对后续研究利用缺乏有力支撑。韩名利[2]重点探讨了《庄子》英译的翻译策略与方法,并提出了语料收集、对齐、标注、检索等操作的具体方案,拓宽了古籍文献的英译研究思路,但也存在语料规模较小的局限性。罗奋[3]介绍了一种构建大规模汉英平行语料库的方法,开发了一个双语语料挖掘自动获取系统,该系统采用B/S结构,使用最新的动态网页开发技术获取平行句子对资源,并为其他双语平行资源的开发提供基础,但利用此方法获取的双语词汇质量并不高。尹存燕[4]对中日双语平行语料库的自动构建进行了研究,提出了两种获取双语句子对信息的自动挖掘方案。实验证明,通过对网页文本进行分析可以获取更高质量的中日双语平行语料资源。程岚岚[5]针对Web上存在的大规模术语网页,提出了一种基于正则表达式的术语抽取方法。这种基于正则表达式的方法虽然在获取某一特定领域的术语资源时效率比较高,但缺乏可移植性,且没有对较为复杂的短语和句子资源进行抽取研究。

本研究在前人双语平行语料库研究的基础上,

收稿日期:2021-09-06

基金项目:江苏高校哲学社会科学研究基金项目"基于CSSCI的组块级汉英平行语料库构建及知识挖掘研究"(2018SJA0473);金陵科技学院高层次人才科研启动基金项目"大数据环境下面对论文相似性检测的学术资源预处理研究"(jit-b-2021-37)

作者简介:徐润华(1982-),男,江苏南京人,讲师,博士,主要从事自然语言处理、信息计量研究。

利用Web自动获取英汉平行句子对资源,并在自动获取的过程中加入语言学知识进行辅助,弥补了过去对统计方法过于依赖所造成的语料质量不高的缺陷[6]。获取英汉平行句子对资源之后,对其进行分词等加工处理,并进一步构建一个大规模的英汉双语平行语料库。经过深度加工处理的语料库资源可以为知识挖掘、机器翻译等其他领域研究提供更深层次的语言知识[7]。

二、英汉平行语料库的网络获取标准及来源

(一)英汉平行语料的网络获取标准

英汉平行语料在网络上的资源非常丰富,分布也较为广泛。从语料资源的具体分布情况入手,综合考虑通用型语料和专门型语料的获取效果,本文选取语料资源的覆盖度、语料资源的准确度和网站的开放度作为英汉平行语料的网络获取标准。

网站的信息、数据的覆盖度是获取英汉平行语料首先要考虑的一个标准。本文使用随机抽样法,在包含英汉平行语料资源的网站上获取一定数量的文本,利用语言学知识对其进行分析和判断,根据语料库覆盖情况给出"高覆盖度""中覆盖度""低覆盖度"三个不同层次的评价级别,分别用"2""1""0"进行量化表示。

英汉平行语料的准确度包括精确度和可行度两个方面的要求,其直接关系到基于英汉平行语料的后续深入研究的实验效果。本文同样使用随机抽样法对样本数据的准确度进行评定,给出"非常准确""基本准确""不太准确"三个不同层次的评价级别,分别用"2""1""0"进行量化表示。

网站的开放度主要是指从网站获取相关资源的方便程度,直接关系到资源获取的难易程度。根据资源获取难易度的不同,本文对网站给出"非常开放""较为开放""不太开放"三个不同层次的评价级别,分别用"2""1""0"进行量化表示。

(二)英汉平行语料的网络获取来源

依据英汉平行语料网络获取标准,本文制定如下获取流程:首先,利用人工或者网页自动抓取软件在包含英汉平行语料的不同网站上获取一定数量的网络数据,并从中提取全部英汉双语平行句子对资源;然后,使用随机抽样软件,对获取的全部英汉双语平行句子对进行随机选取;最后,在综合考虑英汉平行语料的网络获取标准基础之上,辅以专家判别,对随机获取的英汉双语平行句子对进行分析,并以此确定来源网站的评价等级。

按照上文的获取流程并综合考虑英汉平行语料的网络获取标准,排除资源覆盖度低、语料准确度差或者开放度达不到要求的网络获取途径,本文从56个网站中筛选出12个网站作为英汉平行语料的获取来源,具体情况见表1。从语料资源的呈现方式和功能用途两方面进行划分,可将这12个网站分为在线辞典、辅助翻译、英语论坛、搜索引擎、英语门户五个大类。从表1的数据可以看出,在线辞典是英汉平行语料最常见、最主要的获取来源,也是最稳定、最可靠的获取来源。观察表1中"资源覆盖度""资源准确度""网站开放度"三项数据可以发现:从在线辞典来源获取的英汉平行语料属于权威语料,拥有最佳的综合评价等级,整体质量比较高;由于论坛本身具有开放性,因此从英语论坛途径获取的英汉平行语料虽开放度较高,但在准确度方面不尽如人意[8];英语门户类网站的英汉平行语料(如阅读类语料、听力类语料)一般具有较强的针对性和领域性,但资源覆盖度往往不高;辅助翻译类网站主要是为翻译服务的,有较高的覆盖度,但准确度和开放度均无法保证;搜索引擎类网站与辅助翻译网站类似,搜索引擎的功能需求使得其覆盖度较广,但准确度和开放度都不是很理想。

表1 英汉平行语料的网络获取来源情况

网络站点	资源覆盖度	资源准确度	网站开放度	语料形式
金山在线词典	2	2	2	在线辞典
句酷在线翻译	2	2	1	在线辞典
译典通	1	2	2	在线辞典
谷词	2	2	0	在线辞典
海词	1	1	2	在线辞典
知网翻译助手	2	1	1	辅助翻译
词都	2	0	1	辅助翻译
沪江论坛	1	0	2	英语论坛
普特听力论坛	1	0	2	英语论坛
百度词典	2	1	0	搜索引擎
可可听力网	0	1	2	英语门户
酷悠网	0	1	2	英语门户

三、英汉平行语料词表的制定

制定适用于英语语料的词表是实现面向 Web 获取英汉平行语料库的前提和基础。之所以基于英语语料而不是汉语语料制定词表，主要是因为英语语料不需要额外进行分词，而汉语需要进行额外的自动分词工作，加之汉语缺少形态变化、一词多义现象普遍等特点会影响分词效果，进而影响语料库的获取精度。

（一）基于 BNC 语料库的词频统计步骤

词表中有两项数据需要统计获取：一是词语本身；二是词语的出现次数，即词频。考虑到基于 Web 获取语料的规模和效果，本研究利用大规模英语语料库——BNC 语料库进行英语词频的统计工作。BNC 语料库的语料规模达到了亿词次级别，并且该语料库的平衡性非常好，各个领域、各种题材的语言资源都有涉及和收录，既包括书面语语料，也涵盖谈话、聊天、座谈等口语资源。基于 BNC 语料库的英语词频统计步骤如图 1 所示。

图 1　基于 BNC 语料库的英语词频统计步骤

首先，要对 BNC 语料库中的句子进行数据清洗，因为 BNC 语料库中的每一个句子都有大量的标注信息[9]，包括词法、句法甚至是语义层面的标记，如“＜caption id＝BBAF091＞、＜v n＝"529"＞、＜/p＞”等。这些标记对于词频统计来说是冗余信息，需要将其清除。其次，在去除冗余的标记信息后，还需要将句子中词语的形态进行转换。英语是一门形态变化丰富的语言，但一个词语在词表中只能收录一种形式，需要根据英语形态变化的规律制定相应的形态转换规则，例如：将“selected”转换成“select”。这种需要进行转换的情况最常见于动词、名词和形容词。最后，基于排序容器动态插入“＜词语、词频＞”对，词频统计结果示例见表 2。

（二）规则和统计相结合的词表制定方法

在已获取的英语词频信息的基础上，本文参考英语词典中收录的词汇，并结合英语语言学知识，制定英汉双语平行语料库词表，具体步骤如下：首先结合基于 BNC 语料库统计得到的英语词表和英汉词典（英汉综合大词典）所收录的词表，整合形成一个覆盖度高、规模大的英语词表，对于部分形态变化不规则的动词、名词、形容词也将其添加到该词表中；其次，对初步得到的词表进行人工校对，发现其中的错误，并使用停用词词表对该词表进行过滤，剔除其中无意义的词语，得到规模约为 11 万个词汇的英语词表；最后，对初步得到的词表进行“瘦身”，利用词表多次进行基于 Web 的语料资源获取实验，每次实验都减少词表中的词汇数量，当词表中的词语数量减少到一定程度时，获取到的双语平行语料资源会无法覆盖 Web 站点的全部网页，并且覆盖度会随着词语数量进一步减少而继续下降，这个临界点就是制定词表的最佳规模。经过以上三个步骤，本研究最终制定了一个规模为 63 924 个词汇的英语词表，词表构成示例见表 3。从表 3 数据可以发现：有些词汇没有被 BNC 语料库统计，但在词典中被收录，如 alkali；有些词汇没有被词典收录，但被 BNC 语料库统计到了，如 depicting。运用语料库来表示语言现象是基于统计的思路，遵循词典中的专家知识是基于规则的思路，以上两种研究思路互为补充。表 3 中 BNC 语料库统计词语和词典收录词语的示例也恰好印证了这一点。

表 2　词频统计结果示例

序号	词语	词频	序号	词语	词频
1	like	73 987	6	busy	3 369
2	find	67 442	7	being	1 540
3	take	17 651	8	cause	1 279
4	must	10 908	9	action	975
5	person	4 981	10	amount	565

表 3　词表构成示例

序号	制定词表的词汇	形态变换不规则词汇	BNC 语料库统计词汇	英汉词典收录词汇
1	baboon	Y	Y	Y
2	amulet	Y	Y	Y
3	baffle	Y	Y	Y
4	dewy	Y	Y	Y
5	denote	Y	Y	Y
6	also-ran	N	N	Y
7	depicting	Y	Y	N
8	depict	Y	Y	Y
9	alkali	Y	N	Y
10	barmaid	Y	N	Y

注：Y 表示符合，N 表示不符合。

四、基于 Web 的英汉平行语料库的构建

(一)包含英汉平行语料的网页自动获取

能够自动获取网页的软件数量众多,考虑到获取网页过程中所需要的稳定性、高效性和兼容性,本文使用 GUN Wget 软件进行基于 Web 的英汉平行语料的获取工作。GUN Wget 是一个功能强大的开放软件,能够从网络上获取各种数据、文件等资源,支持 TCP/IP 协议,支持 HTTP、HTTPS 以及 FTP 下载[10]。GUN Wget 的主要特点有:链接灵活,可以跟踪 HTML、XHTML 和 CSS 页面上的链接并进行依次下载,代理服务器也可以下载;链接稳定,在带宽很窄或者网络不稳定的情况下表现出较好的鲁棒性;链接快速,能够快速获取网页数据,通过数据缓存或者区域存储的方式实现抓取中止和接续。

网页的自动获取流程主要有制定词表及获取链接、设置 GUN Wget 参数、网页抓取三个步骤。为了应对格式各异的抓取底表,本研究共设置了两种网址与抓取底表中的词汇捆绑方式。网页获取词汇与网址链接生成的程序见图 2,获取词汇与网址生成的链接样例见表 4。基于词汇获取网页链接的特殊性,根据 GUN Wget 自身的文件处理参数、下载参数、目录参数和递归参数等各种参数,结合具体的词汇获取特性,对 GUN Wget 进行相应的参数配置,从而顺利完成各种词汇获取任务。

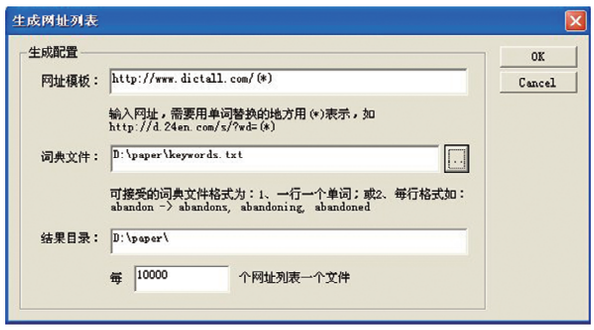

图 2 抓取网页与词汇链接生成程序

表 4 网页获取词汇与网址链接样例

网址链接	抓取底表词汇	生成的链接
http://www.cuyoo.com/	kinetic	http://www.jukuu.com/show-kinetic-1.html
http://www.jukuu.com/	capability	http://www.jukuu.com/show-capability-1.html
http://www.jukuu.com/	majesty	http://www.jukuu.com/show-majesty-1.html
http://www.jukuu.com/	scarcity	http://www.jukuu.com/show-scarcity-1.html
http://www.jukuu.com/	regiment	http://www.jukuu.com/show-regiment-1.html

(二)英汉双语平行语料库的构建

英汉双语平行句子对的抽取要考虑到其在网页中的分布情况以及网页的 XML 标记特点。在抽取平行句子对的基础之上,还需要对其进行数据清洗、去重等操作。

1.英汉双语平行句子对的抽取。在网页中,英汉平行双语句子对都是由 HTML 语言存储,其格式符合 XML 标准,例如:"</td><td> She is almost 20 points ahead of Mr Obama in the latest Gallup poll, and appeals much more strongly to blue-collar workers. </td></tr><tr class=c><td></td><td> 在最新的盖洛普民意调查中她几乎领先奥巴马先生 20 个点,而且对蓝领工人来说她的吸引力更强烈地存在。</td></tr>。"利用这种格式上的规律性可以通过程序设计方便地对英汉双语句子对进行抽取,抽取程序样例见图 3。

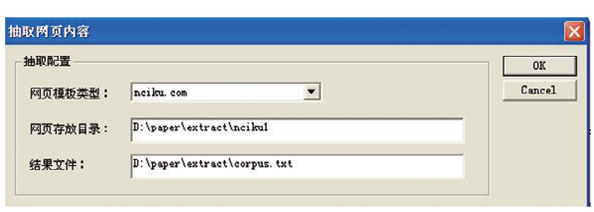

图 3 英汉双语平行句子对抽取程序

2.英汉双语平行句子对的数据清洗。从网页中抽取得到的英汉双语平行句子对存在乱码和数据缺失等问题。乱码现象是由字符编码不一致造成的,为了符合国际惯例,本研究将文本的编码统一设定为 UTF-8 格式;数据缺失主要表现为与英文对应的中文句子缺失或者相反,这种情况的处理方式是直接删除该句子对。

3. 英汉双语平行句子对的去重。由于获取的网页数量较多,英汉双语平行语料库的文本规模又非常大,因此抽取得到的英汉双语平行句子对会不可避免地存在重复现象,去重就成为语料库建设过

程中不可或缺的一道环节。去重主要涉及两个问题：一是对“重复”的定义，二是去重的方式。由于句子对本身文本形式的特殊性，在获取和保存的过程中会出现相似甚至相同句子对的情况，因此需要利用相似度比较算法界定何种相似是因为句子对中存在个别字词错漏而造成的，何种情况只是相似而非相同。解决第二个问题的关键是降低去重工作的时间复杂度。

4. 英汉双语平行语料库的加工和建设。英汉平行语料库有别于一般语料库的特殊之处在于它包含了两种语言文字。英语表意的基本单位是词语，所以英语文本中词语和词语之间有着天然的分隔边界；汉语表意的基本单位是汉字，汉语文本中词语之间没有任何分隔标记。虽然这并不影响人们阅读，但在双语平行语料库中，两种语言的句子是一一对应的，也就是所谓的“平行”。“平行”不仅仅指对应的两个句子在句意上完全一致，还包括句读、短语结构、词语的对应，这些都是基于平行语料库进行语义标注、组块分析、句法分析等研究的基础。为了达到在英汉双语平行语料库中汉语文本和英语文本真正的平行对应效果，需要对汉语文本进行分词处理。

“由字构词”是一种较为常见的中文分词理论。本研究采用六位词分词原理，把汉语词语中的汉字分为六类：单字词(S)、词语首字(B)、词语第二字(F)、词语第三字(G)、词语中部(M)、词语尾字(E)。基于六位词分词原理，本研究使用了目前在中文分词领域表现较为出色的条件随机场(CRF)模型来进行机器学习。条件随机场是一个在给定输入节点条件下计算输出节点的条件概率的无向图模型，擅长处理序列标记问题。对于输入序列 x 和输出序列 y，可以定义一个线性的 CRF 模型

$$P(y \mid x) = \frac{1}{Z(x)} \exp\Big(\sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x) \Big)$$

基于条件随机场模型的这一特性，可以把对汉语进行自动分词的任务转化为序列标注问题。条件随机场的一个重要特点就是支持在机器学习过程中加入任意多个特征进行训练以提高标注的效果。据此，除了汉字六位词的词位特征外，本研究还在条件随机场的训练语料中增加了多个对汉语分词有帮助的语言学特征，如部首、姓氏、外族人名地名音译字、词缀、声调等，从而为自动分词模型提供更多的汉语信息，有效提高模型对汉语进行自动分词的精度[11]。

在对汉语进行自动分词前，应先从英汉平行句子对中把所有的汉语句子提取出来；然后将所有汉语句子形成的文本按照条件随机场模型的要求调整格式，并添加特征信息；完成分词之后，将汉语句子文本的格式恢复，并重新与英文句子一一对应起来。经过获取、抽取、预处理、分词等环节的工作后，本研究最终得到一个词汇层面平行对齐的英汉双语平行语料库，该语料库共包含 1 017 963 个英汉双语平行句子对，语料库样例如表 5 所示。

表 5　英汉双语平行语料库样例

序号	汉语语料	英语语料
1	我现在辞去主席职位，交由我的副手接替。	I am resigning as chairman and handing over to my deputy.
2	他踢足球时，膝盖上的一根骨头错了位。	He displaced a bone in his knee while playing football.
3	科学家们已分离出引起这种流行病的病毒。	Scientists have isolated the virus causing the epidemic.
4	他特意地来这儿看望老朋友。	He has come here express to visit his old friends.
5	各就各位 —— 预备 —— 跑！	On your marks — get set — go!
6	他有着极出色的记忆单词的能力。	He has a rarely seen ability to memorize words.
7	她没有办法离开她的孩子们。	She would never be able to leave her kids.
8	超声波扫描能够检测出胎儿是否畸形。	An ultrasound scan can detect abnormality in a fetus.
9	科举制是在 1905 年废止的。	The imperialist examinations system was abolished in 1905.
10	定期复习可以帮你记住新的知识。	Regular reviews can help you retain new knowledge.

五、结语

高质量、大规模的英汉双语平行语料库有着巨大的研究价值。随着互联网的发展，不同语言间的交流变得日益频繁，双语平行语料库已经成为机器翻译、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，在比较语言学研究等领域发挥着重要作用。但语料库的建设是一个漫长而烦琐的过程，作为一项重要的语言资源，双语平行语料库在规模和质量上都远不及起步更早的单语语料库[12]。而利用互联网 Web 的双语平行资源自动获取方法则是构建双语平行语料库的一种方便、快

捷、高效的途径。

考虑到网络获取语料的来源多样性和数据复杂性,下一步研究的方向是将英汉双语平行语料库存储到专业数据库软件中进行管理和维护。相较于一般的文本编辑工具,数据库软件的存储量更大,对操作环境的兼容性更强,具备可移植性,安全性能也更出色,能够更好地满足英汉双语平行语料库的后续加工、检索等研究任务的需要。

参考文献:

[1] 黄苏豪.英汉平行语料库自动构建系统的设计与实现[D].南京:东南大学,2018

[2] 韩名利.《庄子》汉英平行语料库的创建与应用展望[J].广西民族师范学院学报,2021(1):96-99

[3] 罗奋.基于WEB的汉英平行语料库构建系统开发[D].成都:电子科技大学,2014

[4] 尹存燕.中日双语平行语料库的自动构建技术研究[D].南京:南京大学,2012

[5] 程岚岚.基于正则表达式的大规模网页术语对抽取研究[J].情报杂志,2008(11):62-64,68

[6] 王东波,苏新宁.英汉双语句子级平行语料库自动构建[J].现代图书情报技术,2009(12):47-51

[7] 白拴虎,夏莹,黄昌宁.汉语语料库词性标注方法研究[M].北京:电子工业出版社,1991:89-90

[8] 常宝宝,詹卫东,张华瑞.面向汉英机器翻译的双语语料库的建设及其管理[J].术语标准化与信息技术,2003(1):28-31

[9] 顾益军,樊孝忠,王建华,等.中文停用词表的自动选取[J].北京理工大学学报,2005(4):337-340

[10] 黄昌宁,李涓子.语料库语言学[M].北京:商务印书馆,2002:56-57

[11] 吴琳,魏星,霍翠婷.基于Web的专利双语语料自动获取研究及实现[J].现代图书情报技术,2009(9):57-63

[12] 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(5):67-73

(责任编辑:刘 鑫)

A Web-based Construction Method of English-Chinese Parallel Corpus

XU Run-hua[1], WANG Dong-bo[2]

(1. Jinling Institute of Technology, Nanjing 211169, China;
2. Nanjing Agricultural University, Nanjing 210095, China)

Abstract: With the development of various researches in the field of natural language processing, parallel corpus is playing an increasingly important role as a basic resource supporting natural language processing technology. This study automatically obtains English-Chinese bilingual parallel corpus resources by using the method of information extraction based on massive information resources in the Web. In the process of obtaining, first determine the crawling websites and formulate a vocabulary, then use the crawling tool GUN Wget of network resources to automatically obtain the English-Chinese bilingual sentence pair resourcesin Web pages. On the basis of cleaning and removing duplication for the obtained parallel sentence pairs, the conditional random field model is used to automatically segment Chinese sentences and import them into the database. Finally, the construction of large-scale English-Chinese bilingual parallel corpus is completed.

Key words: parallel corpus; GUN Wget; conditional random field; English-Chinese bilingual; Web