情报分析与研究

英汉双语句子级平行语料库自动构建*

王东波 苏新宁
(南京大学信息管理系 南京 210093)

【摘要】探讨如何基于网络自动构建大规模英汉双语句子级平行语料库的问题,即确定抓取网站和制定相应的抓取底表;利用网络抓取工具 Wget 自动获取含有英汉双语句子对的网页;对从网页中提取出来的英汉双语句子对进行后续加工以及基于条件随机场对汉语句子进行自动分词。最后从 675 308 个网页中共获取 1 017 963 对英汉双语句子对并把句子对导入到数据库中完成英汉双语句子级平行语料库的构建。

【关键词】英汉平行语料库 Wget 抓取底表 条件随机场
【分类号】TP391

Automatic Building of Sentence – Level English – Chinese Parallel Corpus

Wang Dongbo Su Xinning
(Department of Information Management, Nanjing University, Nanjing 210093, China)

【Abstract】This article gives an account of the steps of how to automatically build a large – scale sentence – level English – Chinese parallel corpus based on websites. Specifically speaking, the following questions are addressed: the criterions which are used to grab websites are set and words library is worked out; the websites are automatically grabbed by making use of the tool 'Wget'; the English – Chinese parallel sentences extracted from websites are subsequently processed and the Chinese sentences are segmented based on Conditional Random Field. Finally, the building of English – Chinese parallel corpus is completed which includes 1 017 963 English – Chinese parallel sentences stored in database which are automatically extracted from 675 308 websites.

【Keywords】English – Chinese parallel corpus Wget Words library Conditional random field

1 引 言

随着 Web 信息采集技术和自然语言处理技术的迅速发展,从拥有海量资源的网络上获取有效的数据和知识以便更好地服务于基础和应用研究日趋成为一种趋势。本文英汉双语句子级平行语料库的构建正是在这一趋势下的一种尝试。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[1]。

在搜索引擎和数据挖掘技术的推动下,基于网络的双语素材获取技术日益受到关注。程岚岚[2]针对 Web 上存在的大规模术语网页,基于 Web 挖掘技术,提出了一种基于正则表达式的术语对抽取方法。该方法是获取网页

收稿日期:2009 – 11 – 30
收修改稿日期:2009 – 12 – 08
*本文系国家社会科学基金青年资助项目"对双语语料库介入下学生译者翻译能力的计算机辅助实验研究"(项目编号:09CYY040)的研究成果之一。

源文件，并依据已定义的正则表达式从中抽取出正确的术语对。这种基于规则的方法虽然在获取某一特定领域的术语对效率比较高，但可移植性比较差，同时只获取了相对简单的术语对，而对于比较复杂的短语对和句子对则没有涉及。Zhang[3]、Huang[4]使用不同的方法构造各种查询词列表，然后通过搜索引擎来获取对应查询词的检索结果，进而从检索结果中使用统计的方法获取相应的翻译对。由于搜索引擎本身的局限性，使用这种方法获取大规模的平行语料对相对困难。张永臣[5]提出了一种从网络非平行语料中抽取特定领域双语词典的算法，给出了利用词间关系矩阵法从特定领域非平行语料中抽取双语词典的过程。由于种子词对双语词的抽取影响较大，从实验结果来看，获取的双语词汇对质量并不高。在上述研究的基础上，本文基于网络自动获取了大规模的英汉双语平行句子对，并在对汉语句子自动分词的基础上，进一步构建了英汉双语句子级平行语料库。平行语料对自动获取的过程中融入了大量的语言学知识，弥补了以往构建或获取平行语料对过于偏重技术和统计的不足之处。构建了一个大规模的平行语料库并对汉语进行了自动分词，与以往的平行语料库相比较，确保了从该平行语料对数据库中能获取到更丰富的语言信息和挖掘到更深层的语言知识。

2 研究的基本流程和主要方法

英汉双语句子级平行语料库构建的基本流程如下：确定抓取的网站和制定抓取底表；利用网络抓取工具自动获取可能含有英汉双语句子级平行语料对的网页；结合网页的标记以及平行语料对在网页中的分布特点，从网页中提取出句子级平行语料对并进行初步的整理和加工；对汉语句子进行自动分词并把英汉双语平行句子对存入数据库中。研究的基本流程如图1所示：

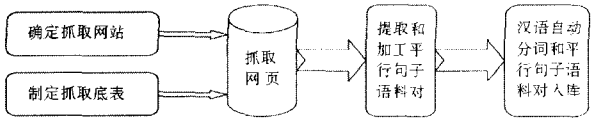

图1 英汉双语平行句子对自动获取流程图

研究中主要使用了语料库、人工内省、统计与比较的方法。使用BNC语料库统计词汇频率获取词汇表的过程中使用了语料库的方法；在统计词汇表的基础上，制定抓取底表的过程中用人工内省的方法把语言学知识融入到抓取底表中；在抓取网站的确定和抓取底表的制定过程中使用了大量的统计方法；由于自动获取的是英汉双语句子级平行语料，比较的方法会贯穿整个抓取过程，如比较英汉平行语料对在语言属性上的差异、汉语需要自动分词而英语不需要等。

3 英汉双语平行句子对自动获取的前期准备工作

3.1 抓取网站

根据具体的研究需要和网络上英汉双语句子级平行语料资源的分布情况，确定所抓取的网站。

(1)抓取网站的标准

网站数据是否丰富是确定抓取网站的基本指标。通过随机抽样统计的方法，确定网站资源丰富与否的量化指标，即网站平行语料对达到多大量的时候才可以抓取。数据丰富性有“不丰富、丰富、很丰富”三个级别，分别用“+、+ +、+ + +”表示；网站数据是否优质是确定抓取网站的关键问题。以翻译学对语言翻译质量“信、达、雅”的衡量标准为前提，并结合语言研究者的内省，对随机从网站上获取的句子对进行评估，最终确定网站数据的优质级别。数据优质性分“一般、优质、很优质”三个级别，分别用“★、★★、★★★”表示。

(2)确定抓取网站

获取拟抓取网站的部分网页，从中提取出一定量的英汉平行句子对。使用随机抽样工具，选取英汉平行句子对。根据确定抓取网站的标准判断英汉平行句子对的数量和质量，进而确定是否抓取该网站上的数据。

最终确定了5大类含有英汉双语平行句子对的网站，主要是外语学习综合性论坛、搜索引擎辅助工具、外语学习听力门户、外语学习阅读门户和在线词典。5大类网站的具体属性如表1所示：

表1 5大类网站具体属性分布

网站类型	具体数量	数据丰富性	数据优质性	网站样例
外语学习综合性论坛	11	+ +	★★	沪江论坛
搜索引擎辅助工具	5	+ + +	★★	百度词典
外语学习听力门户	12	+	★	可可听力网
外语学习阅读门户	13	+	★★	酷悠网
在线词典	16	+ + +	★★★	译典通

3.2 抓取底表的制定

抓取底表即抓取网页过程中的种子数据,在网页抓取的过程中起着非常关键的作用,一定程度上将决定网页抓取的速度、规模和质量。

(1)基于语料库统计词表

一方面,从 BNC(British National Corpus)[6]语料库中用程序去掉每个词的词性标记和其他标记并转存到文本文件中,总体规模约 98 567 320 词次。另一方面,用动态数组实现英语词频统计,具体使用 C++完成程序设计,主要设计了"Found(CString Tag,Int &Id)"和"Insert(CString Tag)"这两个函数,用来完成查找和插入操作。

(2)基于人工内省和抓取实验制定抓取底表

在基于 BNC 语料库统计的词汇表基础上,结合人工内省的词汇表和抓取实验的具体表现,最终制定抓取底表。用程序比对统计方法获取的词汇表和人工内省确定的词汇表,进而合并两个词汇表;通过人工核对合并后的词汇表并增加其他词汇,尽可能地扩大词汇表的规模;在一定词汇量基础上通过逐步增加词汇的方法来进行抓取实验,观察当词汇达到何种数量的时候抓取实验是最理想的,即网页抓取速度快、数量多和质量高,从而最终制定抓取底表。本文制定了一个含有 47 218 个英语词汇的抓取底表。

4 英汉双语句子级平行语料库的构建

4.1 网页的自动获取

基于网页抓取稳定性和高效性的考虑,选取了 Wget 作为抓取工具。Wget 是从网络上自动下载文件没有交互式界面的工具[7],支持 HTTP,HTTPS 和 FTP 协议。Wget 主要有如下特点:可以跟踪 HTML、XHTML 和 CSS 页面上的链接进行依次下载,也就是实现"递归下载";在带宽很窄的情况下和不稳定的网络中表现出了良好的健壮性;能够快速获取不含图像、音频和视频的网页[8]。网页自动获取的流程具体如下:

(1)抓取文件的构建

抓取底表中的词汇与要抓取网站的网址绑定在一起形成抓取列表,构成抓取文件。5 大类下的 57 个网址与 47 218 个英语词汇按照查询的格式捆绑起来,构成 57 个抓取文件。"http://www.jukuu.com/show-urgent-0.html"是句酷网站抓取文件列表中句酷网址"http://www.jukuu.com/"与英语词汇"urgent"按查询的方式捆绑起来的一个样例。

(2)抓取工具 Wget 相关参数的设置

根据抓取网页的具体特性,设置抓取软件 Wget 的相关参数,从而满足特定的抓取需要。在抓取的过程中,本文主要对 Wget 的递归方式(-r)、断点续传(-c)和超时时间(-t)等几个参数进行了设置。

(3)抓取网页

调用各个抓取文件,运行抓取工具 Wget 自动获取可能含有英汉双语平行句子对的网页,并根据各个网站的共享程度,做具体的抓取调整。具体的网页抓取过程如图 2 所示:

图 2 具体的网页抓取页面

在 47 218 个英语词汇的抓取底表基础上,使用 Wget 抓取工具共获取了可能含有英汉双语平行句子对的 675 308 个网页。

4.2 英汉平行句子对的提取和加工

在获取的大规模网页基础上,结合网页的标记以及英汉平行句子对在网页中的分布特点,自动提取英汉双语平行句子对并对其进行相应加工,具体内容如下:

(1)句子对的自动提取

根据不同网页的标记特征和英汉双语平行句子对在网页中的分布特点,总结提取平行句子对的规则。在总结的平行句子对提取规则基础上,基于 C++中的字符串类 CString,通过设计程序把平行句子对字符串提取出来并临时存储到文本文件中。具体的提取软件如图 3 所示。

(2)句子对的后续加工

由于抓取网页中存在大量的重复页面并且有些网页的字符编码不一致,所以对于提取出来的平行语料

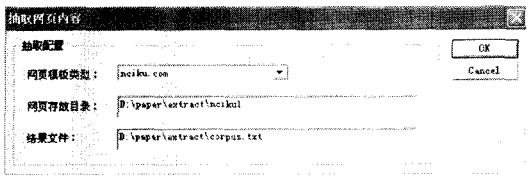

图3　英汉双语平行句子对提取软件

对字符串必须进行去重处理和编码转换。为了确保去重的精确性，本文通过比对平行句子对中的汉语和英语是否完全一致来达到去重的目的。为了解决编码不一致性的问题，去重后的平行句子对统一以 UTF－8 编码的方式存储。

从网页中共提取和加工出 1 017 963 对英汉双语平行句子对，汉语句子以“。?!”等标号结尾，英语句子以“?. !”等标号结尾。

4.3　汉语句子自动分词和英汉双语平行句子对入库

汉语是表意文字，词与词之间没有明显的分隔标记，而英语是表音文字，词与词之间用空格自然间隔开来。通过两种语言之间的对比，英文不需要再进行自动分词，而汉语由于自身的特殊性以及为了有效地进行后续的词性标注、组块识别和自动句法分析必须进行分词。为了达到英汉双语句子级平行语料真正的“平行”和获取英汉对应词，本文基于条件随机场对英汉双语平行句子对中的汉语句子进行了自动分词。

(1)基于条件随机场的汉语自动分词简介

条件随机场(CRF)是一个无向图的判别模型，对于一组长设为 n 的观察序列 $X = X_1X_2X_3\cdots X_n$(要标记的汉字序列)，输出为 $Y = Y_1Y_2Y_3\cdots Y_n$(相应的标注序列)[9]。基于条件随机场模型的这一特性，汉语自动分词的问题就相应地转化为标注问题了，而条件随机场的最大优点是能够加入任意的特征作为输入特征并能够充分利用训练集的统计信息和中文的构词特点，从而向模型提供更多的汉语信息，进而使分词软件的精确率和召回率大为提高[10]。

(2)基于条件随机场的汉语句子自动分词

本文所训练的条件随机场模型的字位数是 4，标记为：B、M、E、S。除了汉字本身的信息外，还在条件随机场的训练模型中增加了以下语言学知识：成语表、部首、汉人姓氏、汉人名字、人名译音、地名译音、词缀、单字名词、单字动词、单字形容词、单字其他词、名词首字、名词尾字、动词首字，动词尾字、形容词首字、形容词尾字、其他词首字、其他词尾字、声调、词首、词尾、译名和姓名。这些语言学知识与汉字本身构成了条件随机场模型训练中的特征。本文所使用的有关 CRF 的相关实验都是使用基于 C++ 语言开发的 CRF++ 工具进行的[11]，训练语料为人民日报语料。

从英汉双语平行句子对中单独提取出汉语句子，按照一定的格式对其进行预处理，然后通过条件随机场模型完成汉语句子的分词，最后把经过分词的汉语句子与英语句子再对应起来。这样，汉语句子与英语句子都是在词汇级别上的平行。

(3)英汉双语平行句子对入库

把英汉双语平行句子对自动导入到数据库中，由于语料来源的多样性和数据的复杂性，专门设计导入程序以解决语料对入库过程中的问题。考虑到英汉双语平行句子对的数据规模比较大以及后续加工和检索的需要，选择 MySQL 数据库来存储平行语料对。之所以要选择 MySQL 数据库，是因为其有下面几个特点：可移植性强，可以运行在不同的操作系统系上；安全性能好，有安全权限和加密口令；数据存储量大，支持上千万条记录的存储[12]。数据库中共导入了1 017 963 对英汉双语平行句子对，数据库中存储具体的英汉双语句子级平行语料对如图 4 所示：

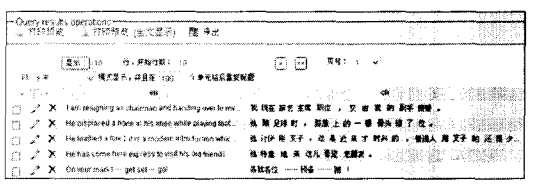

图4　数据库中英汉双语句子级平行语料对存储样例

5　结　语

本文基于 Wget 抓取工具，在一定规模的抓取底表基础上，针对含有英汉双语平行句子对的网站进行抓取，从抓取的过程看，无论抓取的速度还是规模基本上都是令人满意的。从 675 308 个网页中共提取和加工出了 1 017 963 对英汉双语平行句子对，并基于条件随机场对汉语句子进行了自动分词，最后把英汉双语平行句子对导入到 MySQL 数据库，完成了英汉双语句子级平行语料库的构建。下一步将加强对英汉平行句子对语料标注的力度，如词性标注和简单组块标注等，同时开发辅助校对工具对抓取的平行句子对语料进行适

当的人工校对,从而使英汉双语句子级平行语料库真正变成深加工、高质量的语料库。

参考文献:

[1] 王克非. 双语对应语料库研制与应用[M]. 北京:外语教学与研究出版社,2004:232 - 233.

[2] 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*,2008,27(11):62 - 63.

[3] Zhang Y, Vines P. Using the Web for Automated Translation Extraction in Cross - language Information Retrieval[C]. In: *Proceedings of SIGIR*. Sheffield: University of Sheffield, 2004: 162 - 167.

[4] Huang F, Zhang Y, Vogel S. Mining Key Phrase Translations from Web Corpora[C] In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada. Morristown, NJ, USA: Association for Computational Linguistics, 2005: 483 - 490.

[5] 张永臣,孙乐,李飞,等. 基于 Web 数据的特定领域双语词典抽取[J]. *中文信息学报*,2006,20(2):16 - 23.

[6] 王丽,王同顺. 中国英语学习者语用标记语习得研究——一项基于 SECCL 和 BNC 的实证研究[J]. *现代外语*,2008,31(3):294.

[7] Wget Manual[EB/OL]. [2009 - 12 - 06]. http://www.gnu.org/software/wget/manual/wget.html.

[8] Ma X, Liberman M. BITS: A Method for Bilingual Text Search over the Web[C]. In: *Proceedings of Machine Translation Summit VII*. Singapore: National University of Singapore, 1999.

[9] 章成敏,许鑫,章成志. 条件随机场标引模型的性能影响因素分析[J]. *现代图书情报技术*,2008(6):34 - 40.

[10] 李双龙,刘群. 基于条件随机场的汉语分词系统[J]. *软件天地*,2006(10):178 - 179.

[11] The Features of CRF++[EB/OL]. [2009 - 12 - 06]. http://crfpp.sourceforge.net/#features.

[12] Definition of MySQL[EB/OL]. [2009 - 12 - 06]. http://en.wikipedia.org/wiki/MySQL.

(作者 E - mail:jisuanyuyan@163.com)

Gale 公司推出一项针对图书馆的 iPhone 应用

圣智学习出版公司(Cengage Learning,原汤姆森学习出版集团)旗下的 Gale 公司近日宣布推出一种使用 iPhone 查阅图书馆资源的移动应用产品 AccessMyLibrary (AML),用户只需在手机上点击一次鼠标即可实现检索。AccessMyLibrary 实际上是一种图书馆网络门户,它能够帮助 Web 检索者访问当地图书馆,并从图书馆丰富的资料库中准确寻找到高品质的信息。

AccessMyLibrary 在 iPhone 用户和图书馆之间建立可信的网络连接,使读者可以通过手机就能查寻图书馆的海量数据库,并且迅速地获取各种权威解答。目前,AccessMyLibrary 能够帮助用户找到用户所处位置方圆 10 英里范围以内的图书馆。

通过 AccessMyLibrary,iPhone 变成一种非常有价值的检索工具,它使得手机用户在任何地方都能够使用各种电子资源集,并且定位到能够帮助他们获取更多信息的本地图书馆。目前,iPhone 的应用程序是免费的,内容资源也由被访问的图书馆付费,读者不需支付费用。

这项应用软件可以通过苹果公司的 iTunes 商店下载,也可以随时随地通过互联网访问 http://www.accessmylibrary.com 获取,要获取它的详细应用和服务信息可以访问 http://galesupport.com/iphone/aml.html。

(编译自:http://newsbreaks.infotoday.com/Digest/Gale - Announces - New - iPhone - Application - for - Library - Research - 60244.asp)

(本刊讯)