计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900012

面向小语种机器翻译的平行语料库构建方法

刘 妍 熊德意
天津大学智能与计算学部 天津 300350
(yan_liu@tju.edu.cn)

摘 要 神经机器翻译模型的训练效果在很大程度上取决于平行语料库的规模和质量。除了一些常见语言外,汉语与小语种间高质量平行语料库的建设一直处于滞后状态。现有小语种平行语料库多采用自动句子对齐技术利用网络资源构建而成,在文本质量和领域等方面有诸多局限性。采用人工翻译的方式可以构建高质量平行语料库,但是缺乏相关经验和方法。文中从机器翻译实践者和研究者角度出发,介绍了经济高效的人工构建小语种平行语料库的工作,包括其总体目标、实施过程、流程细节和最后结果。在构建过程中尝试并积累了各种经验,形成了小语种到汉语平行语料库构建方法、建议的总结。最终,成功构建了波斯语到汉语、印地语到汉语、印度尼西亚语到汉语各50万条高质量平行语料。实验结果表明,所构建的平行语料库有较好的质量,提高了小语种神经机器翻译模型的训练效果。

关键词:平行语料库;小语种;神经机器翻译模型
中图法分类号 TP391

Construction Method of Parallel Corpus for Minority Language Machine Translation

LIU Yan and XIONG De-yi
College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

Abstract The training performance of neural machine translation depends heavily on the scale and quality of parallel corpus. Unlike some common languages, the construction of high-quality parallel corpora between Chinese and minority languages has been lagging. The existing minority language parallel corpora are mostly constructed by using automatic sentence alignment technology and network resources, which has many limitations such as domain and quality confined. Although high-quality parallel corpora could be constructed by manual, it lacks relevant experience and method. From the perspective of machine translation practitioners and researchers, this article introduces a cost-effective method to manually construct parallel corpus between minority languages and Chinese, including its overall goals, implementation process, engineering details, and the final result. This article tries and accumulats various experiences in the construction process, and finally forms a summary of the methods and suggestions for constructing parallel corpora from minority languages to Chinese. In the end, this paper successfully constructs 0.5 million high-quality parallel corpora from Persian to Chinese, Hindi to Chinese, and Indonesian to Chinese. The experimental results prove the quality of our constructed corpora, and it improves the performance of the minority language neural machine translation models.

Keywords Parallel corpus, Minority language, Neural machine translation

1 背景

深度学习技术的快速发展,使得神经机器翻译成为了机器翻译的主流方法,并且广泛应用于双语翻译中,如英汉、英法、英德等。神经机器翻译模型含有大量参数,其训练过程严重依赖于大规模、高质量的平行语料库。相比常见语种,小语种与汉语间神经机器翻译的研究起步较晚,其主要原因是平行语料库匮乏。现有的小语种语料库基本来自网络资源中的平行数据,依靠自动对齐技术来获取语言对。但该方法的相关技术尚未成熟,数据覆盖领域和规模都严重依赖于网络资源,并且难以评估句子的对齐质量。而人工构建平行语料库可以扩展数据领域范围,保障平行句对的高质量。但是,在小语种平行语料库的构建中仍面临着许多问题。例如,如何获得足够规模的原始语料;如何对原始语料进行清洗,筛选出合适的语句进行翻译;如何做好相关项目规划管理;如何以经济高效的方式构建语料库等。

针对上述问题,并考虑到神经机器翻译模型的训练,本文以波斯语、印地语、印尼语3个语种为例,探索小语种与汉语间高质量平行语料库构建的高效方法和路线,为小语种到汉

到稿日期:2021-09-01 返修日期:2021-10-18
基金项目:国家重点研发计划(2019QY1802)
This work was supported by the National Key Research and Development Program(2019QY1802).
通信作者:熊德意(dyxiong@tju.edu.cn)

语平行语料库的构建搭建标准化流程。

波斯语、印地语、印尼语虽然属于小语种，但它们分别是西亚、南亚、东盟地区的重要语言，也是一带一路国家的重要官方语言。表1列出了3个语种的一些相关情况。

表1 波斯语、印地语、印尼语的相关情况

Table 1 Related information of Persian, Hindi and Indonesian

语言	作为官方语言使用的国家	母语使用人数(Million)	所属语系	相关语言
波斯语	伊朗、阿富汗、塔吉克斯坦	70×10^6 [1]	印欧语系	达利语、塔吉克斯坦语
印地语	印度、斐济	528×10^6 [2]	印欧语系	乌尔都语
印尼语	印尼、东盟	302×10^6 [3]	南岛语系	马来语

随着历史的发展和人口的流动，波斯语在整个西亚地区被广泛使用，并且很多国家都存在波斯语社群。值得注意的是，作为波斯语分支的达利语和塔吉克斯坦语分别是阿富汗和塔吉克斯坦的官方语言。印地语以使用国家数量来算是排名第8的语言，主要分布在印度南部。该语言与乌尔都语(与印地语相似)的使用人数加起来仅次于汉语的使用人数。印尼语是印度尼西亚的官方语言，也是世界上使用最广泛的语言之一[4]。综合以上考虑，本文选择了这3种语种用于构建小语种平行语料库。

本文详细介绍了波斯语-汉语、印地语-汉语、印尼语-汉语各50万人工平行语料库的处理、获取与构建过程。实验结果表明，本文所构建的语料库拥有较高的质量。

2 平行语料库构建的规划及模型

2.1 3个语种-汉语平行语料库的现状

根据公开平行语料库网站OPUS(The Open Parallel Corpus)的数据显示，波斯语、印地语、印尼语与汉语平行语料库的情况如表2所列。

表2 波斯语、印地语、印尼语与汉语平行语料库的基本情况

Table 2 Existed parallel corpora between Persian-Chinese, Hindi-Chinese and Indonesian-Chinese

平行语料库	平行句对(Million)	小语种单词数量(Million)
波斯语-汉语	6.0×10^6	126.6×10^6
印地语-汉语	3.1×10^6	78.5×10^6
印尼语-汉语	7.7×10^6	151.5×10^6

2.2 现有平行语料库分析

以波斯语为例，表2所列平行语料库的来源分别是WikiMatrix v1[5]，XLEnt v1.1[6]，Wikimedia v20210402[7]，Tanzil v1[7]，infopankki v1[7]，TED2020 v1[8]，QED v2.0a[9]，Ubuntu v14.10[7]。WikiMatrix v1是Facebook Research在多种语言句子向量基础上以自动方法从Wikipedia网站上抽取的包含85种语言的平行句对语料库。XLEnt v1.1通过挖掘CCAligned，CCMatrix，WikiMatrix所得。CCAligned，CCMatrix，WikiMatrix本身是通过网页快照和自动抽取技术得到的平行语料库。XLEnt v1.1是以英语为中心的实体对齐语料库。构建该语料库的研究者首先通过命名实体识别技术来发现英语语料中的实体，再对平行的120种语言中的句子进行实体对齐。Tanzil v1是由42种语言译本的《古兰经》所形成的语料库。TED2020 v1包含了由全球志愿者翻译的超过100种语言版本的将近4000场TED和TED-X的演讲记录。QED v2.0a是属于教育领域的语料库，来源于公开的多种语言版本的教育视频或教育课程的字幕。Ubuntu v14.10是关于Ubuntu操作系统本地文件的平行语料，共由244个译本组成。印地语和印尼语语料库的情况类似。可以看出，上述小语种平行语料库大都是在特定领域通过自动抽取技术和对齐方法进行抓取并构建而成，缺乏高质量人工翻译平行语料。

此外，这些平行语料库的建设大都以英语为中心，很少存在以小语种和汉语为原始源语言、目标语言的平行语料库。以WikiMatrix v1语料库为例，其来源Wikipedia(维基百科)网站是包含多种语言版本的网络百科全书。然而，由于维基百科采取多人协作模式维护网站内容，因此各语言版本维基百科发展并不均衡[10]，相同含义词条的内容也并不完全对应。中文维基参与的元维基翻译计划，也是以翻译英文维基百科条目为主。综上所述，现有平行语料库质量不利于我国小语种到汉语的机器翻译研究工作的发展。

2.3 平行语料库构建方案

为了更好地开展面向小语种机器翻译的平行语料库构建的工作，为更多语种与汉语间平行语料库的建设提供思路，本文进行了各种尝试，提出了一个相对完整的、能够以较快的时间完成构建较高质量的小语种到汉语的平行语料库方法。

为保证翻译的质量，本文采用人工翻译的策略。所使用的翻译语料来自于涵盖多领域内容的新闻网站，以达到数据覆盖领域广的目的。根据预算以及训练一个较好模型所需数据量进行综合考虑，每个语种各构建50万句对平行语料。

构建流程如图1所示。该方法通过各个环节不同的辅助工具，尽可能地保证了平行语料库中语料的搜集、加工和组织工作。

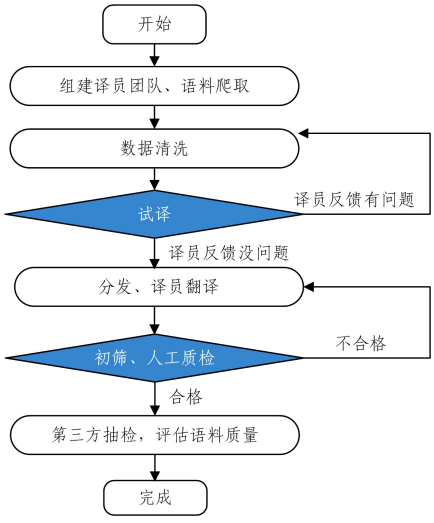

图1 平行语料构建总览图

Fig. 1 Overview of construction of parallel corpora

译员团队组建、语料爬取以及整理同期进行为本项目的第一阶段工作。第二阶段抽取所准备语料的部分内容构成试译文件，测试并统计译员翻译水平、速度和质量，以对项目工期进行把控。同时，收集译员对试译文件的反馈，了解所准备

语料存在的问题，以便及时调整。第三阶段正式进入项目完成阶段，主要做好项目进度记录、语料收发、检验等工作。

3　面向小语种机器翻译的平行语料库构建方法

本节内容主要介绍了平行语料库的人工构建过程，主要涵盖了原始语料的采集与处理、人工翻译项目管理流程、语料质量控制及语料统计分析几个部分。

3.1　原始语料来源

考虑到新闻领域的数据具有较好的时效性与覆盖性，本文采集的小语种原始语料均来源于互联网新闻。数据采集过程主要分为 3 个步骤：首先选取较著名的新闻网站作为网页爬虫工具 Common Crawl[11] 的初始网页。然后以这些网页为起点，大规模搜索波斯语、印地语、印尼语网页，以获得足够的原始文本数据。最后对爬取的文本数据做清洗，每个语种各得到 50 万条语料。

Common Crawl 是一个自动爬取互联网网页并将网页档案和网页数据集免费提供给公众的非盈利性组织。自 2008 年以来，Common Crawl 所爬取的网页档案已包含拍字节(petabyte)的数据，并且到目前为止，每月依然进行一次爬取与更新。本文使用 Common Crawl 提供的其中一种索引类型 Common Crawl Capture Index，CDX) 对网页档案进行筛选，找到与初始网页最相关的网页集合。

3.2　原始语料爬取

以波斯语语料爬取为例，选取表 3 中所列的 5 个网页作为 Common Crawl CDX 检索网站的初始网页。The Khaama Press News Agency，TOLOnews 等均为阿富汗伊斯兰共和国著名的新闻服务机构。本文主要选取年份较新的网页数据，以保证所收集语料的时效性。随后，以 Common Crawl 抓取到的网页为基础，分别为波斯语、印地语、印尼语爬取了上百万条语句作为原始语料，经过清洗后每个语种各得到超过 50 万条的数据用于人工翻译。

表 3　波斯语初始网站示例

Table 3　Initial website of Persian

新闻机构	网址
The Khaama Press News Agency	https://www.khaama.com/persian/
TOLOnews	https://tolonews.com/fa/
Radio Azadi	https://da.azadiradio.com/
Pajhok	https://pajhwok.com/fa/
1TVNews.af	https://1tvnews.af/fa/

本环节还实现了多进程网页文本爬取工具作为 Common Crawl 爬取网站方式的补充。该工具通过 3 个步骤获取网页文本：首先通过广度优先搜索算法爬取网页中的所有和初始网页在同一个域名下的链接；然后，从远程服务器拉取网页的前端代码，去除代码中的 html 标签，得到网页中的文本；最后使用分词工具对文本进行分句，按照每行一条语句的方式将爬取的文本按顺序保存在文件中。

3.3　数据清洗

挑选出合理的原始语料进行人工翻译是语料库构建工作中非常重要的一环，在一定程度上影响着所构建语料库的质量。本文对原始语料进行挑选进行了以下几项工作。

首先是去重，将所爬取语料与现存的小语种到汉语的平行语料库进行对比去重。其次，由于在神经机器翻译模型训练时过长的句子会造成计算资源消耗大、模型显存占用不稳定等问题。因此，本文去除了过长的句子，保留了每个语种长度在 20 词到 50 词之间的句子。然后，保留信噪比较高的句子。例如，一些语料文本中可能会含有部分其他语言的引用，或者存在无关符号比例过大的现象。这些包含了许多与目标语言无关信息的句子不应保留。本文将所爬取语料中小语种相应 Unicode Block 字符含量低于 75% 的句子进行去除。最后，经过译员试译后的反馈，本文在原始语料的挑选过程中尽可能地保留了所爬取语料的顺序，使语料具有上下文参考性，方便译员在翻译过程中保持连贯性和效率。

为了在有限的语料库建设上提高语料质量，充分发挥人工构建语料库的优势，尽可能地挑选出多样性的文本。本文尝试采取文献[12]提出的主动学习[13]方法对原始语料进行筛选。信息增益的概念可以衡量双语平行语料中句子的重要性，筛选出真正对模型训练有用的句子。该文献将所有未经过挑选的句子看作一个集合，通过计算每个句子所带来的信息增益对所有句子进行排序。句子的信息增益情况根据句子中没有在挑选语料集中出现过的 n-gram 个数除以句子长度计算得到。在每轮挑选中都选出得分最高的句子加入被挑选语料集。句子的信息增益的计算公式如下：

$$sentence\_weight=\frac{\sum_{n=1}^{2}\#(unseen\_n\_grams)}{|sentence|}$$

其中，首先统计未在被挑选语料集中出现过的 uni-gram 和 bi-gram 个数，然后除以以词为单位所统计的句子长度。

经过试译后，译员反馈：使用该方法筛选出的句子包含了大量的人名、地名和机构名，并且都是长难句。在翻译过程中，译员需要花费大量的时间去核对这些人名、地名、机构名的译文并商量保持统一，严重降低了译员的翻译效率。虽然本文在使用该方法的过程中考虑了 n-gram 出现频率的问题，并且尽量在保持文本多样性的情况下挑选了包含出现频率较高 n-gram 的句子，但筛选后得到的原始语料依然存在较为严重的相同问题。我们猜测是以下原因造成了该主动学习方法的不适用：所爬取语料属于新闻领域，多种多样的人名、地名和机构名通常是新闻报道中的必要元素，使用 n-gram 方法不可避免地会挑选出长难句。虽然目的都是挑选出对模型训练有益的句子，但该方法更侧重于在已有的双语语料的基础上进行挑选，对于单语数据筛选，则可能需要在已有主动学习方法的基础上补充更多限制。

最终，本文筛选出了超过 50 万个句子作为原始语料提供给译员进行翻译。

3.4　译员组织

由于专业译员翻译费用非常高昂，小语种翻译尤甚，且国内小语种翻译专业译员数量有限，难以在一定时间内保证较大规模语料的构建。从实际角度考虑出发，本文尝试了启用相关语种专业的高年级学生进行翻译。本次语料翻译工作中

波斯语、印地语、印尼语分别招募30名左右学生译员，均来自国内外著名外语翻译专业高年级学生及毕业生，如北京外国语大学、北京第二外国语大学、西安外国语大学、暨南大学、伊朗德黑兰大学等。

3.5 项管流程

项目管理流程是人工构建语料中非常重要的环节，需要做到有序组织。在本次语料库构建工作中，将项管流程分为以下几个步骤。

3.5.1 材料组织

每个语种的50万条数据采用Excel文件进行组织。Excel文件中A列为句段编号、B列为原文、C列为译员需要填写的正式译文。译员在翻译过程中若认为原文句子有问题不能翻译，那么在原文表格前端输入“ERROR”进行标记。对于人名、地名、机构名等专有名词，使用共享文档由译员协商登记在术语汇总表中，尽量使这些专有名词的翻译保持统一。对于句子中存在的英文词，若是一些公开的缩写，比如IBM，则直接在译文中写成IBM，否则，进行正常翻译。

3.5.2 翻译策略

在翻译过程中，译员可以直接根据自己的知识进行翻译，也可以根据所提供的小语种机器翻译模型采用MTPE(机器翻译后编辑)的方式进行翻译，根据机译质量或者自己的偏好选择一种翻译方式。据统计，本次翻译工作大部分译员采用了MTPE的方式进行翻译。

3.5.3 语料分发

以每100条语料为单位建立Excel文件，将完整语料进行分割。按各批次时间节点将相应Excel文件分配给译员进行翻译。并另建立批次进度表统计文件名称、负责译员、译文返回时间、是否初筛、质检结果和完成状态等。

3.5.4 译文回收

本环节主要进行译文回收及初筛。随后对初筛合格数据进行抽样、提交给第三方翻译公司进行人工质检。

3.6 质量控制

3.6.1 译文初筛

译员按时返回译文后，对译文进行初筛，初筛的主要目的是检查译员对机器翻译的依赖程度，对译文进行译后编辑率检查，若译后编辑率低于10%，则需要返工重新翻译。

译后编辑率的检查方式为：首先调用谷歌翻译(经对比，谷歌翻译在此3个语种的翻译效果上优于其他翻译平台)对波斯语、印地语、印尼语原始语料进行翻译。然后使用编辑距离[14]、Needleman-Wunsch算法[15]、n-gram precision的算术平均数计算译员翻译译文与谷歌翻译译文的相似度，得到译员的译后编辑率。

3.6.2 人工质检

对于初筛合格的译文，每个Excel文件抽样10%的数据组成抽样语料提交给第三方翻译公司，由职业译员按照表4中的验收标准进行人工质检，若质检不合格，则判定该译员的整体翻译都不通过，译员需及时对负责的所有Excel文件进行返工并重新进行新一轮的质量控制检查，直至合格为止。

表4 人工质检验收标准

Table 4 Artificial quality inspection and acceptance criteria

分数	打分标准
5分	可以接受，翻译得很好(挺不错的，用词恰当，翻译通畅、贴切)
4分	可以接受，翻译得较好(还不错的，意思相符，用词恰当，但可以翻译得更好)
3分	可以接受，翻译得一般(句义大致正确，没明显错误，个别词语暧昧，部分非关键词未翻译到或多译都归到此类)
2分	不能接受，翻译得较差(句子不通顺，句义表达不清)
1分	不能接受，大段漏译(大部分内容漏译或错译)
0分	不能接受，翻译与原文完全无关

4 语料库统计分析

4.1 语料库完成情况

本次小语种语料库各语种完成情况为：波斯语-汉语共完成501 000句对，印地语-汉语共完成511 932句对，印尼语-汉语共完成501 700句对。

4.2 人工质检结果的统计及分析

我们共抽检波斯语50031条数据、印地语51727条数据、印尼语50344条数据。由表5可以看出：波斯语得分为3分的句子数量占绝大部分比重，翻译质量为中等。印地语得分为3分、4分、5分的句子分布较平衡，翻译质量较优。印尼语得分为3分、4分的句子较多，翻译质量中等偏上。

表5 人工质检结果统计

Table 5 Statistics of manual quality inspection results

(单位：%)

语言	小于3分	3分	4分	5分
波斯语	1.5	94.6	3.7	0.1
印地语	0.5	32.1	45.8	21.5
印尼语	0.5	30.9	49.1	19.5

4.3 第三方质检结果统计

为保证所构建平行语料库的最终质量，在第三方翻译公司人工质检结果基础上，本项目还请了第三方质检进行第二次人工质检检查。这次质检分别抽取第一次人工质检打分数据中不合格数据、3分数据、4分数据、5分数据和未打分数据若干混合后交给第三方质检进行打分。最终结果如表6所列。

表6 第三方质检结果统计

Table 6 Statistics of third-party quality inspection results

(单位：%)

语种	小于3分	3分	4分	5分
波斯语	1.2	7.8	30.4	60.6
印地语	4.1	13.4	12.0	70.5
印尼语	6.9	23	27.2	42.9

由表6可以看出，第三方质检的评分对3个语种的翻译结果整体评价较高，相比第三方翻译公司打分集中在3分、4分，第三方质检评分打分则较多地给到了5分。此外，各语种评分上升条数较多，可证明第三方翻译公司在判定时打分较为严格，实际语料的质量可能优于第三方翻译公司估计的质量。

5 实验

本文在波斯语到汉语、印地语到汉语、印尼语到汉语翻译

任务上评估了所构建语料库的有效性。

5.1　实验设置

实验所用数据集分为3类。第1类采用2.1节介绍的各个语种已有语料库经过清洗合并后的数据集训练模型，各语种训练集规模如表7所列。第2类采用本文所构建语料库训练模型。第3类将第1类数据集和本文所构建语料库合并在一起作为模型训练的语料。

表7　各语种3类数据集训练集规模

Table 7　Training set size of three types of datasets in each language

语种	第一类数据集	第二类数据集	第三类数据集
波斯语-汉语	2.36×10^6	0.5×10^6	2.86×10^6
印地语-汉语	0.42×10^6	0.52×10^6	0.94×10^6
印尼语-汉语	2.26×10^6	0.42×10^6	2.69×10^6

3类数据集的验证集和测试集均采用所构建语料库中划分出的语料，其中波斯语922条，印地语1 000条，印尼语841条。

实验模型采用目前主流的Transformer[16]，其编码器解码器设定为6层，模型嵌入尺寸为512，前馈层维度为2 048，注意力头数为8。批尺寸为5 000词，学习率为0.000 5，dropout率为0.25。实验的翻译性能采用BLEU值[17]进行评估。

5.2　实验结果及分析

实验结果如表8所列，可以看出，波斯语、印地语翻译模型在第2类和第3类数据集上的训练效果较好，证明了所构建的波斯语、印地语小语种平行语料库的有效性。

表8　BLEU值的实验结果

Table 8　Experimental results of BLEU

语种	第一类数据集	第二类数据集	第三类数据集
波斯语-汉语	4.19	42.71	44.12
印地语-汉语	4.17	28.30	28.85
印尼语-汉语	25.15	15.15	16.42

然而，波斯语和印地语模型在第1类数据集上训练效果较差，经分析，可能存在测试集与第1类数据集领域分布差别较大的原因。印尼语模型在第3类数据集上训练效果与第1类数据集差距较大。可能原因是组成第3类数据集的两个语料库数据类型分布存在差异。

6　相关工作

目前机器翻译领域平行语料库的构建工作主要集中在自动构建方法的改进上。文献[18]通过对已有的86个Common Crawl网页快照库进行挖掘，识别互为译文的网络篇章文本，从而得到多个语言对的平行语料库。该文使用的主要方法为网页对齐技术，将嵌入在网页网址(URL)中的标志用来标记网页篇章文本并进行对齐，在各个语言对上达到了平均94.5%的准确率。最终，该工作成功发表了包含8 144个语言对的平行语料库，其中，137个是包含英语的语言对。同等量级的工作还有文献[19]，其从327亿条句子中挖掘出了45亿平行句对，包含38种语言。其中，20个语言对都超过了3 000万平行句对，112个语言对超过了1 000万平行句对。文献[20]为解决所爬取网络文本中噪声数据过多的问题，提出了一种通过大规模预训练语言模型过滤噪声数据的方法。该文还采用了BERT[21]模型用于衡量平行句对的相似度，以及利用具有生成能力的预训练语言模型GPT[22]平衡各领域的语料数量。文献[23]设计筛选方法来提升伪平行语料的质量。一般的伪平行语料仅通过一次回译[24]得到。该文献对使用回译方法得到的伪源语言语料再次进行回译，得到目标语言的伪语料，最后计算目标语言原语料与伪语料之间的句子相似度，只保留相似度较高的伪平行句对。

其他小语种平行语料库的构建情况依然主要以英语为中心。文献[25]为低资源语言对英语-泰语构建平行语料库。为了达到一定规模，该文献充分利用了各个领域的文本，包括网络爬取文本、政府文件文档、预训练语言模型生成的文本以及公开的适用于自然语言处理任务的英文文本。在收集好原始语料后，对双语语料进行对齐工作。对较难的单语文本进行人工翻译，而对较简单的单语文本采用众包的方式进行翻译。在低资源语言对英-缅甸语平行语料的构建上，文献[26]提出了对回译方法所构建的伪平行语料中噪声平行句对进行过滤的系统。该系统的构建基于余弦相似度和孪生BERT网络训练得到跨语言句子向量。此外，针对缅甸语的语言特点，该文献还提出了有助于提升英文与缅甸语翻译模型性能的适用于缅甸语子词分割的无监督分割模型。文献[27]研究了以推特内容为源构建适用于机器翻译的平行语料库的方法。该方法具有普适性，可以用于补充低资源语言对的语料来源。最终，该文献以英语-阿拉伯语为例成功构建了适用于机器翻译的平行语料库。

7　建议

平行语料库构建是一项需要注意细节的工作，在已有构建流程的基础上，建议对相应的语种特点进行了解，做好语料的预处理工作。另外，在正式开始语料库构建工作前，建议先进行试译工作。试译过程可充分反映译员素质、语料库情况，这对原始语料的及时调整、项目进度把控有着重要影响。

结束语　本文成功构建了波斯语到汉语、印地语到汉语、印尼语到汉语3个小语种与汉语间的平行语料库。本文从神经机器翻译模型所需训练数据出发，提出了一个相对完整的人工平行语料库构建方法。在未来的工作中，我们将继续探究更加经济高效的语料库构建方法，为小语种语料库的规模进行扩充。

参 考 文 献

[1]　GERNOT W. The Iranian languages[M]. Routledge, 2009.

[2]　LIAO B. The Language Situation in India-An Analysis Based on the Language Survey Data of the Indian Census in 2011[J]. Journal of PLA University of Foreign Languages, 2020, 43(6): 7.

[3]　JIANG S Y, LI S S, FU S H, et al. An Overview of Natural Language Processing for Indonesian and Malay[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(6): 12.

[4]　JAMES N S. The Indonesia languages: Its history and role in Modern Society[M]. UNSW Press, 2004.

[5]　SCHWENK H, CHAUDHARY V, SUN S, et al. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia[J]. arXiv: 1907.05791, 2019.

[6] El-KISHKY A, RENDUCHINTALA A, CROSS J, et al. XLEnt: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment[J]. arXiv: 2104.08597, 2021.

[7] TIEDEMANN J. Parallel Data, Tools and Interfaces in OPUS [C]// Lrec. 2012: 2214-2218.

[8] REIMERS N, GUREVYCH I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation [J]. arXiv: 2004.09813, 2020.

[9] GUZMAN F, SAJJAD H, VOGEL S, et al. The AMARA Corpus: Building Resources for Translating the Web's Educational Content[C] // International Workshop on Spoken Language Translation(IWSLT). 2013.

[10] ZHAO F, ZHOU T, ZHANG L, et al. Research Progress on Wikipedia[J]. Journal of University of Electronic Science and Technology of China, 2010(3): 321-334.

[11] SMITH J R, SAINTAMAND H, PLAMADA M, et al. Dirt cheap web-scale parallel text from the Common Crawl[C]// Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013). 2013.

[12] ECK M, VOGEL S, WAIBEL A. Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF [C]// Proceedings of International Workshop on Spoken Language Translation. 2005.

[13] SETTLES B. Active Learning Literature Survey [J]. Science, 1995, 10(3): 237-304.

[14] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions and reversals[C]// Soviet Physics Doklady. 1996: 707-710.

[15] NEEDLEMAN S B. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of Molecular Biology, 1970, 48(3): 443-453.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[17] PAPINENI S. Blue: A method for Automatic Evaluation of Machine Translation[C]// Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2002.

[18] EL-KISHKY A, CHAUDHARY V, GUZMAN F, et al. CCAligned: A massive collection of cross-lingual web-document pairs[J]. arXiv: 1911.06154, 2019.

[19] SCHWENK H, WENZEK G, EDUNOV S, et al. Ccmatrix: Mining billions of high-quality parallel sentences on the web[J]. arXiv: 1911.04944, 2019.

[20] ZHANG B, NAGESH A, KNIGHT K. Parallel Corpus Filtering via Pre-trained Language Models[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8545-8554.

[21] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// NAACL-HLT (1). 2019.

[22] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[23] IMANKULOVA A, SATO T, KOMACHI M. Improving low-resource neural machine translation with filtered pseudo-parallel corpus[C]// Proceedings of the 4th Workshop on Asian Translation (WAT2017). 2017: 70-78.

[24] GRAÇA M, KIM Y, SCHAMPER J, et al. Generalizing Back-Translation in Neural Machine Translation[C]// Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). 2019: 45-52.

[25] LOWPHANSIRIKUL L, POLPANUMAS C, RUTHERFORD A T, et al. A large English-Thai parallel corpus from the web and machine-generated text[J]. Language Resources and Evaluation, 2021, 55(1): 1-23.

[26] ZIN M M, RACHARAK T, LE N M. Construct-Extract: An Effective Model for Building Bilingual Corpus to Improve English-Myanmar Machine Translation[C]// ICAART (2). 2021: 333-342.

[27] MUBARAK H, HASSAN S, ABDELALI A. Constructing a bilingual corpus of parallel tweets[C]// Proceedings of the 13th Workshop on Building and Using Comparable Corpora. 2020: 14-21.

LIU Yan, born in 1992, postgraduate. Her main research interests include neural machine translation and discourse parsing.

XIONG De-yi, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include natural language processing, especially in machine translation, dialogue, and natural language generation.

(责任编辑:李亚辉)