doi:10.6043/j.issn.0438-0479.202401010

领域专有平行语料库的敏捷构建方法

李 铭,张克亮*
(国防科技大学外国语学院,江苏 南京 210000)

摘要:[**目的**]针对领域平行语料库数量不足,构建困难的问题,研究能够敏捷构建满足垂直领域模型训练要求的领域平行语料库的方法.[**方法**]在构建的通用大规模跨领域平行语料库的基础上提出了领域专有平行语料库的自动构建方法.该方法结合语言学理论概念层次网络与词向量模型构建目标领域核心词汇,并依此抽取目标领域平行句对,从而实现领域专有平行语料库敏捷构建.[**结果**]以军事领域为例,进行领域专有平行语料库构建与领域机器翻译的测试结果表明:相较于使用关键词对比、预训练模型与语言大模型等方法,本文方法所构建的领域平行语料库具有更高的 F_1 值.使用该方法生成语料所训练的机器翻译模型在该领域的翻译结果相比于上述方法与商业翻译引擎具有更高的双语互译评估(BLEU)值.[**结论**]本文所提方法能够有效复用现有的高质量开源语料资源,并在数小时之内完成最贴合目标领域的平行语料库的构建,从而提升领域机器翻译的效果.

关键词:领域专有平行语料库;概念层次网络;句对抽取算法;语义距离计算
中图分类号:TP 391;G 35 **文献标志码**:A **文章编号**:0438-0479(2025)04-0586-11

Agile construction method of domain-specific parallel corpus

LI Ming, ZHANG Keliang*
(College of International Studies, National University of Defense Technology, Nanjing 210000, China)

Abstract: [**Objective**] In terms of the lack and the difficulty related to constructing domain-specific bilingual corpora, this study aims to generate an agile construction method of large-scale bilingual corpora so that needs of domain-specific model training can be met. [**Methods**] Based on large scales in discriminative multilingual corpus, we developed the method by jointly using readily available open-source monolingual and parallel corpora with self-owned corpora. Domain-specific Corpus Automatic generator (DsCAG), namely a method of automatic-construction of domain-specific parallel corpora, is explored in depth. First, domain-specific core words are extracted by jointly using the Concept-Net of Hierarchical Network of Concepts (HNC) and Word2Vec model. Then, parallel sentences, which are mostly approximate in meaning to the target domain, are extracted based on a multi-round computation of semantic distance between sentence and domain. [**Results**] Experimental results on Military field tests show that this method can construct the domain-specific bilingual corpus with the highest F_1 among comparative methods, such as key words match, Pre-trained language model (PLM) and Large language model (LLM). Translation tests on Military field show that the translation engine trained by this corpus secures the highest BLEU compared with PLM, LLM and commercial translation engines. [**Conclusions**] DsCAG can effectively reuse existing high-quality recourses of corpora to construct a domain-specific parallel corpus, which is mostly characterized with the target domain in hours. Optionally, DsCAG shows the capability of promotion on domain-specific translation.

Keywords: domain-specific parallel corpora; HNC; sentence extraction algorithm; semantic distance computation

平行语料库是机器翻译任务中不可或缺的资源, 随着近年来深度学习方法在机器翻译领域的大规模

收稿日期:2024-01-11 **录用日期**:2024-12-30
* **通信作者**:kliang99@sina.com
引文格式:李铭,张克亮. 领域专有平行语料库的敏捷构建方法[J]. 厦门大学学报(自然科学版),2025,64(4):586-596.
Citation: LI M, ZHANG K L. Agile construction method of domain-specific parallel corpus[J]. J Xiamen Univ Nat Sci, 2025, 64(4): 586-596. (in Chinese)

应用并取得了良好的效果,其重要性和必要性愈加凸显. 神经机器翻译模型的最终效果的优劣依赖于是否具有超大规模、高质量的平行语料库资源.

语料库语言学和机器翻译等领域经过数十年的发展,研究人员开发出了大量有标注的语料数据. 尽管有许多语料资源并没有公开分享,但依然存在大量经过高质量标注的通用开源语料资源. 面向特定领域的神经翻译模型训练需要大规模的领域专有平行语料,构建高质量平行语料库往往需要数月甚至数年的辛勤工作,而现有的领域专有平行语料在数量和规模上都无法满足需要. 因此,如何高效地利用已有的开源通用平行语料库并结合自有资源敏捷构建面向不同领域的专有平行语料库就成为了值得研究的问题. 基于此,本研究将对现有的开源平行语料库进行收集和整理,并结合自有语料资源进行整合,构建大规模通用平行语料库. 此外,本研究还尝试提出一种自动构建领域专有语料库的方法. 其核心思想是在已有的、经过不同层次加工过的语料的基础上最大限度地扩展可利用的语料数量,将高质量的平行语料库与领域相关的单语语料进行结合,从而实现大规模领域专有平行语料库的敏捷自动构建.

本文将所提方法命名为 DsCAG(domain-specific corpus automatic generator),并主要讨论其 4 个主要方面的具体步骤:

1) 构建可持续扩展的通用多语种大型平行语料库.

2) 利用语言学理论概念层次网络(hierarchical network of concepts,HNC)、词向量模型和基于词向量的语义计算方法得到领域核心词汇组.

3) 从构建的通用大型平行语料库中抽取与目标领域语料相关的平行句对.

4) 结合目标领域单语语料,使用反向翻译(back translation,BT)的方法生成更多的平行句对.

本研究将使用所提方法设计目标领域平行语料库自动生成系统,并在军事领域进行测试,验证所提方法的有效性.

1　相关工作

当下,Vaswani 认为[1],在机器翻译领域广泛应用的 Transformer 模型[2]只有在训练样本数量达到 3 亿时,模型的 F_1 值才能够显著高于基于其他神经网络架构的模型. 平行语料库的建设往往决定了后续相关机器翻译任务研究的成败,又因语料收集的困难性、需要的人力物力都应具有相当的规模,研究人员提出了多种方法帮助构建平行语料库. 如 Liu 等[3]提出的平行语料库建设通用的 6 个基本步骤:1) 网页、文档爬取;2) 文本提取并规范化;3) 内容筛选与匹配;4) 文档分解,提取句子;5) 句子对齐,形成句对;6) 句对过滤与清洗. 现在的平行语料库研究和构建工作通常具有数量不足、耗时过长以及领域划分不清的缺点,例如:

1) 使用人工翻译的方法构建平行语料库. 如刘研等[4]使用人工翻译的方法构建了波斯语、印地语、印尼语 3 种语言到汉语每种约 50 万的平行句对,但未对领域进行区分.

2) 使用一些自动对齐的方法构建平行语料库. 如 Utiyama 等[5]利用现有的专利数据库,通过自动打分与降噪的方式使句子自动对齐,从而获得了约两百万日英专利句对,但并未对专利所属领域进行分类.

3) 从互联网上自动获取大量的平行句对,从而构建平行语料库. 如 Morishita 等[6]于 2019 年构建 JParaCrawl 语料库,其包含了 870 万余英日平行句对,这些句子都是从互联网上获取并进行质量过滤而获得的,研究者将这些平行句对训练的翻译模型作为预训练模型,并对具体领域的翻译模型训练进行指导. 然而其实验时所选取的开源领域专有平行语料库的领域分类标准模糊,领域的确定没有理论依据. 于 2022 年发布的 JParaCrawl V3.0 版本语料规模增大至 2 200 万句对. 它将领域划分为:科学论文、电影字幕、维基百科、TED 演讲、对话、新闻以及社交网络评论等,混淆了领域、体裁的概念. 此外,研究者对语料库的扩展花费了长达 3 年的时间[7].

4) 利用现有的大规模平行语料库进行融合整理,从而构建更大规模的通用平行语料库. 如 Agić 等[8]构建的 JW300 语料库就试图融合多个高质量开源平行语料库,从而构建超大规模、覆盖度广但未进行领域划分的通用平行语料库.

5) 使用一些自动构建平行语料库的方法,如使用回译(BT)的方法通过单语语料构建伪平行语料[9-11],基于词向量模型及神经网络模型计算句子与句子之间的语义距离的方法从大规模多语言语料数据中抽取句对从而构建伪平行语料库[12-13],以及其他一些自动方法,如元翻译信息的设计[14]. 然而这些工作往往未对领域、领域句对以及领域核心词汇进行有效的筛选和分辨,也无法保证平行语料库的质量.

还有很多工作针对性地构建了单一领域平行语料库,但其规模较小. 如沙九等[15]构建的司法领域藏

汉高质量平行语料库,其规模仅包含 16 万句对.

总的来说,国内外平行语料库建设工作已经取得了很大的成果,但对于领域平行语料库的构建而言,还有一些问题尚未得到解决. 首先,已有的平行语料库没有进行领域分类工作,或者其领域的划分没有理论依据,不能满足垂直领域机器翻译的需要. 其次,现有的领域平行语料库规模尚不能满足基于深度学习的机器翻译模型训练. 另外,使用人工参与的方法进行平行语料库构建工作往往需要数月甚至数年的时间,而仅使用自动方法则无法保证最终的语料库质量. 最终,语料重用问题也没有得到很好的解决.

针对上述问题,本研究首先构建了跨领域的大规模多语种通用平行语料库,并在此基础上提出了 DsCAG:一种大规模领域专有平行语料库的敏捷构建方法. 方法首先结合语言学理论 HNC 和词向量模型筛选出领域核心词汇组,然后通过基于随机抽样的多轮循环对语义距离进行筛选,自动从跨领域的、通用平行语料库中抽取出与目标领域最为贴合的领域平行句对. 该方法能够在 10 h 之内完成领域平行语料库的敏捷构建,最终抽取的领域平行句对还可以和高质量的目标领域单语语料进行融合,使用 BT 的方法扩大领域专有平行语料库的语料规模.

2 跨领域大规模通用语料库资源收集与构建

本研究通过对开源语料库资源进行收集和整理,并结合自有语料资源,设计并构建了跨领域的大规模多语种通用平行语料库. 此外,为了实现领域双语平行语料库自动构建方法,本研究也收集了若干英、汉单语语料资源用于词向量的训练.

本节将对收集整理的平行语料资源、单语语料资源和数据清洗所用到的语言处理工具资源进行简介. 在语料库的构建中,本研究严格遵守相关开源协议,对于没有开源授权的语料资源,仅给出相关介绍,未进行整合. 读者如使用此语料库也需要遵守各资源的原始协议和引用条约.

2.1 开源语料收集

为构建跨领域的大规模多语种通用平行语料库,本研究收集了大量的开源语料资源并进行加工、过滤和整理. 表 1 整理了在构建时主要参考的开源语料来源.

表 1 开源语料收集

Tab. 1 Collection of open-source corpora

语料类型	语料名称	包含内容	语料规模	语种数量	获取链接
平行语料	联合国平行语料库[16]	平行句对	1 136 万句对	6	https://conferences.unite.un.org/uncorpus/Home/Index/zh
	word2word 多语言词对语料库[17]	平行词对	3 564 词对	62	https://pypi.org/project/word2word/#description
	基于维基百科制作的大规模平行文本语料库[18]	平行句对	1.34 亿句对	85	https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix
	OPUS[19]	大型语料库、平行句对	1 210 语料库、459 亿句对	744	https://opus.nlpl.eu/
	nlp_chinese_corpus[20]	中文语料、平行句对	100 万篇维基百科中文词条、250 万篇新闻、150 万个百科问答、450 万个社区问答、520 万中英句对	2	https://github.com/brightmart/nlp_chinese_corpus
中文单语语料	THUOCL:清华大学开放中文词库[21]	中文词表	15.7 万词条	1	http://thuocl.thunlp.org/
	WebQA[22]	中文问答	4.2 万个问答	1	http://idl.baidu.com/WebQA.html
	CMRC[23]	中文问答	35 万个问答	1	https://github.com/ymcui/cmrc2017
	DRCD[24]	中文段落	10 014 个中文段落	1	https://github.com/DRCKnowledgeTeam/DRCD

续表

语料类型	语料名称	包含内容	语料规模	语种数量	获取链接
英文单语语料	C³[25]	中文对话	13 369 篇对话	1	https://dataset.org/c3/
	NLPCC-2019 Shared Task[26]	中文句子	3.8 万句	1	http://tcci.ccf.org.cn/conference/2019_0826/taskdata.php
	Brown	英文文档	100 万词文档	1	https://varieng.helsinki.fi/CoRD/corpora/BROWN/
	LOB	英文文档	500 个文档	1	https://varieng.helsinki.fi/CoRD/corpora/LOB/index.html
	BNC	英文文档	1 亿词文档	1	https://www.english-corpora.org/bnc/
	COCA	英文文档	10 亿词文档	1	https://www.english-corpora.org/coca/
	ANC	英文文档	15 亿词文档	1	https://anc.org/
多语种单语语料	*N*-gram annotation form digitized books[27]	多语种书籍	519 万书籍	7	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/

2.2　跨领域大规模通用平行语料库构建

上述语料资源中的文本格式、文件格式、标注规范往往都不统一,本研究对所有语料进行统一处理,进行了数据清洗工作,其中包括:文件格式统一化、非语言字符清理、相似句子去重、非法句子过滤、非标准句对修改,从而构建了跨领域的大规模通用平行语料库.

其中汉语部分借助 jieba(https://github.com/txsjy/jieba)和 hannlp(https://github.com/hankcs/pyhanlp)对语料进行分词,并对分词结果进行对比匹配,对少数不一致的情况进行人工校对. 分词所用的算法如下:

1) 基于前缀词典实现词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG).

2) 采用动态规划查找最大概率路径,以词频为基准选择概率最大的切分组合.

3) 对于未登录词,采用基于汉字成词能力的隐马尔可夫模型,使用 Viterbi 作为解码算法.

英文语料处理方面,本研究使用 Facebook 提供的 LAMA 工具包对文本进行标注和清洗[28].

最终,构建的跨领域的大规模通用平行语料库采用. translate 文件格式对语料数据进行保存,并分别对双语语料进行句对齐. 最终构建的跨领域大规模通用平行语料库共包含超过 85 种语言的平行句对,其中英汉平行句对超过 1 亿句对、10 亿字词,其他低资源语种每种语言的平行句对数量为 10 万到 100 万对.

3　领域专有平行语料库生成方法

本研究所提出的领域专有语料库自动构建方法的具体步骤分为:

1) 领域核心词汇的纵向收集:利用 HNC 将领域相关的概念节点及其下位节点纵向获得领域的核心词汇.

2) 领域核心词汇的横向扩展:利用训练好的词向量模型进行词汇语义距离计算,获得领域核心词汇的横向扩展语汇.

3) 句对抽取:利用扩展后领域核心词汇群从大规模通用语料库中抽取领域专有句对.

4) 与目标领域单语语料进行融合:使用 BT 的方法将抽取的领域专有句对与领域单语语料资源进行融合,最终生成领域专有平行语料库.

经过上述 4 个步骤,实现了领域平行语料库的自动构建,下文对 4 个步骤分别进行详细的介绍.

3.1　纵向领域核心词汇收集

领域指的是一种特定的范围或区域,是人类根据所感受到的事物和知识根据共同特征抽象出来的概念. HNC 对于具有基元性和系统性的抽象概念具有非常详细的描述,蕴涵了丰富的世界知识[29].

HNC 以概念为切入点,试图构建模拟出存在于人类大脑中的概念激活联想脉络,从而实现跨语言的概念建模. 具体地,HNC 参考跨语言百科全书构建第一类扩展基元概念中的 a 行概念,即专业活动. a 行概念网络将专业活动分为 8 个一级领域子网络(政治、经济、文化、军事、法律、科技、教育、卫生). 本研究将上述一级领域子网络作为领域分类理论依据,并对子网络下层概念基元进行详细分析,从而映射出相关领域纵向核心词汇. 与其他语言学理论如:WordNet、

HowNet 等更加关注词汇或是词汇的意元的语言学理论相比,HNC 更加着重于从概念的层次性着手,具有概念的纵向特征,更适合作为领域区分的参考和纵向核心词汇的分析与选取. 本文测试主要基于英汉双语进行,但 HNC 理论的使用提供了将本文所提方法应用于跨语言的使用环境中的可能性[30].

以军事领域为例,在 HNC 概念网络中,通过对 a4(军事)节点及其子节点的分析,可得出军事网络的主要子概念基元,它们分别为:a40 一般军事活动、a41 组织、a42 战争、a43 战争效应、a44 军事行动和 a45 军事技术. 其中 a45 中相关概念过于专业化,所以核心词汇应该来自于 a40、a41、a42、a43、a44 节点. 继续对这些节点的子概念基元进行分析,发现在英文和汉语中,很多军事领域概念所映射的词汇都由 military 和军事组合而来,显然 military 和军事分别是英汉语言中军事领域最核心的词汇.

最终经过分析得到的军事领域纵向核心词汇见 4.2 节系统测试中的表 1. 同时,为了证明 HNC 理论的有效性,在军事领域实验验证时,选取了语言学理论 WordNet 与 HowNet 对军事领域进行了核心词汇组的分析与抽取并与本文基于 HNC 理论的方法进行对比.

3.2 横向领域核心词汇扩展

首先需要利用收集的领域的单语语料训练词向量模型. 常见的词向量训练模型有能够有效捕捉全局统计信息的基于共现矩阵构造的方法和能够有效获取局部句法信息的基于窗口词的预测方法. 本研究选用了 GloVe 模型[31]训练静态向量,该方法利用滑动窗口遍历语料库,能够同时捕捉局部的语法信息和全局的统计信息,其具体算法如下:

统计 word-word 共现矩阵 $\boldsymbol{X}$,其中 X_{ij} 指的是 word i 的上下文中出现 word j 的次数,X_i 指的是 word i 在语料中出现的总次数,其统计方式:

$$X_i = \Sigma_\kappa X_{ik}. \tag{1}$$

P_{ij} 为 word j 出现在 word i 上下文的概率:

$$P_{ij} = P(j \mid i) = \frac{X_{ij}}{X_i}, \tag{2}$$

而比值 F 的计算函数为

$$F(w_i, w_j, w_k) = \frac{P_{ik}}{P_{jk}}. \tag{3}$$

本研究利用收集的语料资源训练英、汉静态词向量,其中向量维度为 300. 下面以经济领域和军事领域为例,图 1 给出了核心词组词向量在二维空间的投射.

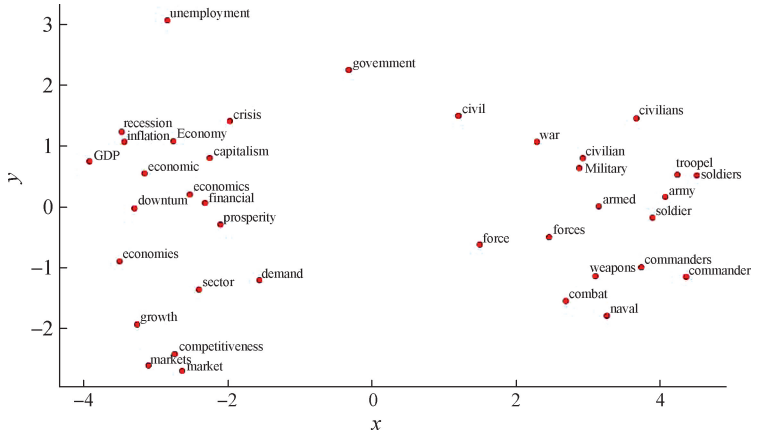

图 1 经济与军事领域核心词汇

Fig. 1 Core words of economy and military

可以看出,某一个领域在本质上指的就是该领域的若干核心词组成的向量空间,而不同领域的文本通常使用不同的词,通过计算句子中的核心词汇到该向量空间的距离就可以确定该句子到该领域的距离.

因此有了词向量之后,就可以通过计算词向量的余弦距离,在向量空间中寻找总是和领域核心词汇共现的相关词汇,实现对领域核心词汇的横向扩展.

3.3 句对抽取算法

从通用大规模平行语料库中选取高质量的句对是本文提出的方法模型中最重要的步骤. 为了从已有通用句对中找出与目标领域最相关的句子集合,不能简单地使用关键字、词查询的方法. 本文提出了一种基于词嵌入的领域计算模型,该模型可以计算出句子和目标领域的语义距离,通过语义距离的比较筛选,提取出与目标领域最为相关的句对集合,从而为接下来和目标领域语料进行融合并构建目标领域专有语料库做好准备. 需要注意的是,该模型包含对语料进行质量评价并筛选的模块,当句子被取出并进行处理时,模型首先判断该平行句对是否符合一定的质量要求,对长度过短、双语长度差别过大的句子进行过滤,

并自动处理非关键标点符号.

句对抽取算法的简单描述如下.

算法输入:目标领域 d、目标领域核心词汇 D_s、需要的语料数量 n、待处理语料库 C,8 个一级领域分类、各 50 句领域代表句集合 S_t.

算法输出:领域平行语料库 C_d.

根据句对距目标领域距离 d_s、d_T 按比例 d_1、d_2、d、d_4、d_5 分别保存 5 个等级的句对集合,根据所需的平行语料规模依次进入最终领域平行语料库 C_d.

其中,计算语义距离值 d、d_s、d_T 的方法如下:

1) 对于任意包含源语言 S 和目标语言 T 的平行语料,通过 3.1 和 3.2 节给出语言 S 的目标领域代表词集合 D_s,其中包含 i 个词,i 为模型的超参数. 之后给出语言 T 目标领域代表词集合 D_T. D_T 可以根据 D_s 进行翻译得到,也可以选择使用构建 D_s 同样的方法.

2) 对于某一个句对 D_P,计算语种 S 的句子 S_{source} 到目标领域词集合 D_s 的语义距离 d_s,与语种 T 的句子 S_{target} 到目标领域词集合 D_T 的语义距离 d_T. 以 S 为例,其具体实现如下:对 S_{source} 进行过滤,删除功能词、停用词等特殊词汇,得到句子词集合 W_s.

这里使用向量的余弦距离计算某两个词 $\boldsymbol{w}_1$ 与 $\boldsymbol{w}_2$ 之间的语义距离:

$$\text{distance}(\boldsymbol{w}_1,\boldsymbol{w}_2)=\cos\theta=\frac{\boldsymbol{w}_1\boldsymbol{w}_2}{\|\boldsymbol{w}_1\|*\|\boldsymbol{w}_2\|}. \tag{4}$$

要计算 W_s 和 D_s 的语义距离 d_s,可设 W_s 有 j 个单词,设 D_s 有 i 个单词,其计算式如下:

$$d_s=\frac{1}{i}\frac{1}{j}\sum_1^j\sum_1^i\text{distance}(W_{s_j}D_{s_i}). \tag{5}$$

从式(5)可以看出,句子到领域的距离就是句子的核心词汇向量到领域的词袋词汇向量平均的距离.

同样的方法得到 d_t,则返回语义距离 d:

$$d=\text{avg}(d_s,d_t). \tag{6}$$

在算法中通过同时计算平行句对双语句子到领域的向量距离能够减少源文符合领域特征而译文不符合或译文符合而源文不符合的情况.

句对抽取算法伪代码如下:

```
Input:d,Ds,n,C,ST.
Output:Corpus Cd
def extact(d,c,Ds):
    ES1,ES2,ES3,ES4,ES5=[]
    d1,d2,d4,d5=d*[0.5,0.75,1.25,1.5]
    for s in c:
        ds=distance(s.source(),Ds)
        dT=distance(s.target(),Ds)
        if abs(s.source().length()-s.target().length())>10:
            continue();
        if ds<d1 and dT<d1:
            ES1.append([s,avg(ds,dT)])
        elif ds<d2 and dT<d2:
            ES2.append([s,avg(ds,dT)])
        elif ds<d and dT<d:
            ES3.append([s,avg(ds,dT)])
        elif ds<d4 and dT<d4:
            ES4.append([s,avg(ds,dT)])
        elif ds<d5 and dT<d5:
            ES5.append([s,avg(ds,dT)])
    Return([ES1,ES2,ES3,ES4,ES5])
Dt=distance(ST,Ds)
Dt=DT.sort(reverse=True)
d=DT[int(n/C.length()*1 000)]
ES=extract(d,c,Ds)
Cd=[]
For S in ES:
    if (S.length()+Cd.length())<n:
        Cd.append(sorted(S,key=lambda x:x[1],reverse=True))
    else:
        Cd.append(sorted(S,key=lambda x:x[1],reverse=True)[0:(n-S.length()-Cd.length())])
```

3.4　与目标领域单语语料融合

BT 的方法可以结合使用真实的目标语言语料和已有的双语平行语料从而伪造大量的平行语料[32]. 对于伪造后的重复语句可以使用左世亮等[33]提出的相似句段去重算法进行去重,对于其中的低质量句对使用刘婉月等[34]提出的基于熵的统计语言模型去噪方法保证伪造平行句对的质量. 如果用户拥有高质量的目标领域平行语料,可以直接和本方法抽取的句对进行融合. 针对单语言领域高质量语料,可以使用 BT 的方法与抽取的平行句对进行融合. 因为上述方法抽取的句对保证了其中的句子和翻译方式是最贴近目标领域的,从而能够保证在进行融合之后的平行句对能够最大限度地保证领域相关性.

本方法集成了上述相关方法,其步骤如下:

1) 基于 3.1～3.3 节方法最终抽取的句对,使用 Transformer 模型训练语言 S 到 T 的翻译模型 $\text{MT}_{S\text{-}T}$.

2) 使用 $\text{MT}_{S\text{-}T}$ 翻译准备好的语言 T 的单语语料资源,获得伪平行句对.

3) 对伪平行句对进行过滤,过滤掉熵值过低和句

对字符数量差距过大的句对.

4 系统设计与系统测试

4.1 系统架构

根据本文讨论过的相关方法,设计并实现了一个目标领域平行语料库自动生成系统. 该系统的架构图如图 2 所示.

下面对其中的核心模块进行简要阐述:

1) 通用大规模平行语料库模块. 系统收集大量的开源语料库,并将其中的平行语料库进行处理和整合,形成具有很大规模的通用平行语料库,其来源多样化,翻译方式多样化,但都已进行过对齐和清洗处理.

2) 用户接口模块. 该模块接受用户的输入,其中包含目标领域名称、目标领域单语语料、期望构建的平行语料库句对数量 n 等.

3) 词袋生成模块. 该模块结合 HNC、词向量模型和目标领域的词频统计给出能够代表输入领域的核心词汇.

4) 目标语料库构建模块. 核心功能,接受用户输入和处理过的目标领域语料后,根据已有数据,自动构建目标领域平行语料库,其中包含语句抽取模块和 BT 模块.

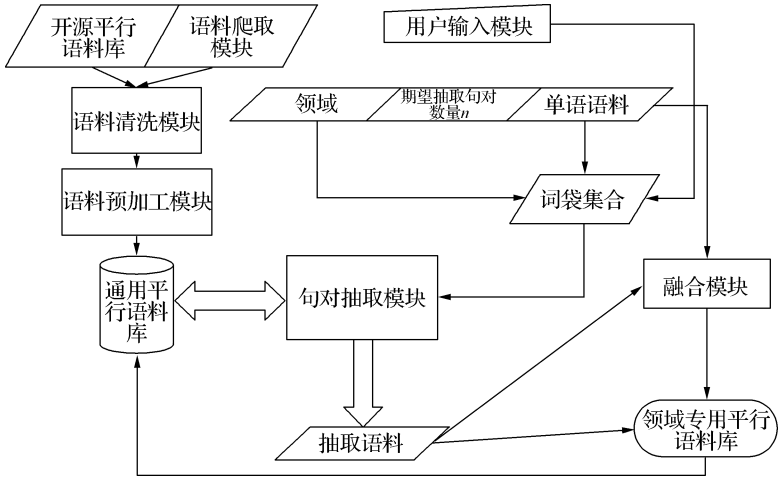

图 2 系统架构

Fig. 2 System architecture

4.2 军事领域核心词汇抽取

系统构建完成后,我们使用所提方法并选取军事作为测试领域进行了该领域平行语料库的自动构建测试. 测试首先通过对 HNCa4 军事节点进行分析,得到军事领域纵向核心词汇组,其具体词汇及 HNC 概念代码如表 2 所示.

对上述词汇进行分析,deploy、security、withdraw 等单词也常出现于概念网络其他领域节点中,所以需要将其去掉.

有了 HNC 给出的纵向核心词汇,就可以使用词向量模型对它们进行扩展. 使用更多的词汇可以提升抽取的语句与目标领域的相关性. 使用更多的核心词汇能够在向量空间上划取更广阔的领域语义空间,但同时,过多的核心词汇也会使得语义空间缺少鲜明的领域特征,如果把处于领域语义空间边缘的词汇也囊括进来反而会影响句对抽取效果,同时加大计算量. 在测试中,最终选取最合适领域核心词的数量 i 为 20 个,并设置 8 个核心词汇以及 40 个核心词汇作为对比.

表 2 HNC 军事领域纵向核心词汇组

Tab. 2 Vertical core words of military by HNC

军事领域子节点	核心词汇(概念符号)	
军事活动	military(a4)	security(a40)
军事组织	army(a41b)	troops(a41a * 1)
战争	warfare(a42)	combat(a42a)
战争效应	casualties(a438)	attack(aa43e11)
军事行动	deploed(a44e21)	withdraw(a44e22)
军事技术	weapons(a451)	armed(a452)

表 3 展示了通过计算语义距离,选出距军事领域其他核心词汇的平均语义距离最近的 20 个单词,最低相似度为 0.474 7.

表 3　横向核心词汇及其与语义距离

Tab. 3　Horizontal core words and ics average semantic distance to else core words

单词	平均语义距离	单词	平均语义距离
military	0.680 9	troops	0.639 9
army	0.636 3	soldiers	0.610 0
forces	0.587 7	war	0.585 1
combat	0.583 8	armed	0.560 5
civilian	0.558 4	Commanders	0.529 5
infantry	0.527 6	tactical	0.524 9
weapons	0.521 6	battlefield	0.520 7
naval	0.509 1	enemy	0.509 1
warfare	0.508 4	marines	0.485 5
battalion	0.483 6	Iraq	0.474 7

使用同样方法可以得到 20 个中文核心词汇组,如表 4 所示.

表 4　中文核心词汇组

Tab. 4　Chinese core words

军事	部队	军队	士兵
武装	战争	战斗	战场
指挥官	平民	步兵	武器
装备	敌军	海军	空军
南海	海军陆战队	营	作战

同时,实验使用 WordNet 与 HowNet 等语言学工具对军事领域进行了纵向核心词汇组分析,分析结果如表 5 所示.

表 5　WordNet 与 HowNet 得到的核心词汇

Tab. 5　Core words from WordNet and HowNet

语言学理论	输出词汇
WordNet	military;armed_forces, armed_services, military_machine, military_machine
HowNet	SMAC, work in three shift, matter of no great urgency, harlotry, prostitution, whoredom, cause, course of study, employment, job, occupation, profession, practice, professional work, vocational work, middle shift, main business, core business, major business, main subject

对比表 3 与表 5 可知,WordNet 从同义词入手对 military 进行扩展,其同义词汇并不能有效涵盖并构建军事领域的语义空间. 而 HowNet 则通过 military 词汇的军事义元入手,所获得的词汇过多地与其他领域进行了交叉,无法有效区分军事领域与其他领域的语义空间. 相较而言,使用 HNC 理论分析获得的词汇,能够有效涵盖军事邻域语义空间并与其他领域具有明显的区分度.

4.3　军事领域平行语料库构建测试

在得到核心词汇后,进行军事领域平行语料库构建的实际测试,测试环境如下.

CPU:AMD R7 7950X;

OS:Windows 11;

Program language:Python 3.7;

Memery:32 GB.

使用关键词对比的方法发现,每个句子中出现表 3 中关键词 2 次的句子约有 30 万句、出现 1 次的句子约有 300 万句. 测试结果中分别使用上述两个结果作为基线系统 2 和 1.

测试使用本文所述方法分别选取了 1 000 万、500 万、300 万以及 100 万的句对抽取数量,选用 8,20,40 个核心词汇,并加入关键词的方法进行对比. 除此之外,本文选取文献 7 所提的预训练模型以及大规模语言模型(LLM),让其对句对进行军事领域判断从而构建军事领域平行语料库.

为了测试方法的召回率 r,提前将人工精选的长度适中的 1 万个军事领域句对加入通用语料库,将句对抽取结果中包含精选军事领域测试句对(test sentence pair,TSP)的数量 D_{TSP} 和其总数的比率作为系统召回率:

$$r=\frac{\mathrm{count}(D_{\mathrm{TSP}})}{100\ 000}. \tag{7}$$

正确率的计算方法为使用人工进行评测. 人工对抽取的句子中进行十轮随机抽样,每次抽取 100 个句子,正确率的计算如下式:

$$c=\frac{\mathrm{count}(\text{领域句对})}{100}. \tag{8}$$

最终求取 10 次准确率的平均值作为评测标准. 这里需要注意的是抽取的语料数量 n 的大小会显著影响抽取结果的准确率,因当前通用平行语料库的总句对数为 1.1×10^8,当 $n>10^7$ 时,不管使用什么方法,准确率都会低于 60%.

最终,F_1 值的计算方式为:

$$F_1=c*r*\frac{2}{c+r}. \tag{9}$$

最终准确率的测试时,在输出语料 n 为 300 万时,又加入了 BT 的方法额外构建 100 万平行句对,对结果进行数据增强后测试其正确率 c,最终测试结果 c,r,F_1 以及花费时间如表 6 所示. 当 $i=20$,$n=3\times10^6$ 时,DSCAG 的 F_1 值最高,分别比基线系统 1 和 2、预训练模型[7]、LLM[7]的 F_1 值高 32.0、33.1、20.1 和 3.9 个百分点. 当 i 一定时,随着 n 的增加,c 呈明显的下降趋势,r 呈明显的上升趋势,且 i 越大,这种趋势越明显. 这是因为语料库越大,领域不相关的语料会越多,c 必然会越小,r 越大;且核心词汇越多,越容易将领域不相关的语料划分进语料库,c、r 受 n 变化的影响越明显. 而当 n 一定时,核心词太少容易漏掉领域相关的语料,而核心词太多又容易将领域不相关的语料划入语料库,因此,c、r 均随 i 的增大呈先增后减的趋势.

表 6 测试结果

Tab. 6 Test results

方法	i	$n/10^6$	$c/\%$	$r/\%$	$F_1/\%$
基线系统 1	20	3	75.6	48	58.8
基线系统 2	20	0.3	92.3	42	57.7
预训练模型[7]	\	1.5	65.8	76.3	70.7
LLM[7]	\	2	81.9	87.5	86.9
DsCAG	20	10	51.3	**99.8**	67.8
DsCAG	8	5	61.9	97.3	75.7
DsCAG	20	5	70.3	99.6	82.4
DsCAG	8	3	76.7	86.7	81.3
DsCAG	20	3	88.7	93.2	**90.8**
DsCAG	40	3	74.6	78.1	76.3
DsCAG	8	1	94.5	74.6	83.4
DsCAG	20	1	**98.9**	79.9	88.3
DsCAG+BT	20	5	81.2	93.2	86.8

4.4 军事领域机器翻译测试

进一步使用上文所提 DsCAG 以及 DsCAG+BT 所构建的军事领域专有平行语料库分别训练基于 Transformer 架构的端到端英汉机器翻译模型,模型训练参数选取注意力头个数为 6、隐藏层维度 1 024,批大小为 256,轮次为 40,学习率为 0.003 并使用 Adam 学习率优化器.

本文额外构建了 1 000 句经过人工精校的军事领域测试句对,选取常见商业翻译软件进行英汉翻译译文 BLEU 值的对比,BLEU 的计算使用 ScareBLEU API. 对比结果如表 7. 其中 DsCAG+BT 方法在测试集上取得的 BLEU 值最高,为 43.77%.

表 7 测试 BLEU 值

Tab. 7 BLEU for tests

翻译引擎	BLEU/%	翻译引擎	BLEU/%
百度	38.67	谷歌	21.87
预训练模型[7]	23.59	DsCAG	42.59
LLM[7]	33.9	DsCAG+BT	**43.77**

4.5 测试结果分析

本节首先于表 8 列举几个军事领域句对判断实例.

通过上述实例可以看出本文所提方法相比于对比方法能够结合原文与译文更加准确的识别出正确的军事领域.

因为测试所用的大量句对都不含有相关关键词,所以使用关键词匹配的方法无法有效地抽取领域专有句对,而使用本文方法则可以有效地快速自动构建领域专有平行语料库,最终测试的 F_1 值在使用本方法选取 20 个核心词汇、抽取句对数量为 300 万的时候最高(表 6). 过高的抽取句对的数量会降低准确率,过低则会降低召回率. 同时过低的核心词汇数量会降低召回率,而过多的核心词汇数量会模糊领域之间的界限从而显著降低测试结果.

另外,在使用 BT 方法后,测试结果的 F_1 值有所减弱,但能够对语料库规模进行扩充,效果仍然大幅优于基于关键词的方法. 使用本文所提方法构建的领域平行语料库进行机器翻译模型训练,能有效提升最终模型的领域机器翻译效果.

在测试中,军事领域平行语料库的自动构建方法皆在 10 h 以内运行完成,相比于传统的语料构建方法,大幅地缩短了构建时间,增加了语料复用的效果.

总结来看,本文所提方法具有以下特点:

1) 与其他语言学理论相比,使用 HNC 作为理论依据筛选出的核心词汇有效性明显,可以实现用词汇代表领域的效果,并可对领域进行明确、细致的划分.

2) 将 HNC 与词向量模型进行结合能扩展出更加灵活、准确的领域核心词汇.

3) 使用关键词进行语料抽取的方法具有一定的效果,但是非常不灵活,比如无法识别关键词未出现的领域句对、无法对识别出的句对进行语义距离排序并优中选优,也无法避免某句话出现关键词但其译文不属于该领域的情况.

表 8　实例的领域判断

Tab. 8　Domain judgment of examples

中文	英文	DsCAG	预训练模型[7]	LLM[7]
到目前为止,尚未对该项目进行系统性分析,且研究尚未评估这些部署对平民或非战斗用途人员在部署后心理健康护理利用方面的影响.	To date, no systematic analysis of this program has been conducted, and studies have not assessed the impact of these deployments on mental health care utilization after deployment for civilians or non-combat use.	军事领域	医疗领域	公共卫生政策、心理健康研究或政策研究领域
那个连因为在人员和医疗保障上花费了过多的经费而导致后勤严重不足.	The company's logistics were severely strained due to excessive spending on personnel and medical support.	军事领域	经济领域	商业或企业管理领域
这次实战演习在公司搞得如火如荼,是近年来最为重要的一次战斗.	The practical combat exercise is in full swing within the company, representing the most significant battle undertaken by the company in recent years.	非军事领域	军事领域	军事领域

4) 使用 BT 的方法可以在对正确率影响不大的情况下进一步对领域专有平行语料库进行扩充.

5) 使用领域专有语料库对模型进行训练,能够有效提升领域专有机器翻译模型的效果.

总的来说,本文所提方法能有效地利用开源语料资源与自有资源构建大规模通用平行语料库,结合语言学理论提出的 DsCAG 实现了领域专有平行语料库的快速构建,与其他人工构建领域专有平行语料库的方法相比,该方法构建速度更快,具有更为有效的领域划分.

5　结论与未来工作展望

当下自然语言处理领域的研究离不开大规模语料库的应用,对于机器翻译来说,常见的端到端模型更是需要大规模的平行语料库. 平行语料是由不同语言之间进行翻译对齐的句对构成,在海量的数据需求面前、有限的人力无法实现翻译、标注、对齐的工作. 尽管如此,经过数十年的发展,众多研究者已将他们标注好的语料数据开源共享,互联网上也有很多高质量的平行句对和带一定标注的单语语料.

本文的主要贡献分为两个方面:第一、充分利用开源和互联网上的公开数据,构建了跨领域大型多语言通用平行语料库,解决了语料复用问题;第二、提出了领域平行语料库自动构建方法,该方法结合了 HNC 和词向量理论,将先验知识和统计知识进行结合,实现了敏捷并且效果较优的语料抽取.

该方法在实现的过程中依然需要人为的定义领域、对领域词汇进行筛选,在下一步的研究中,会进一步地使用基于自监督训练的模型对该大型语料库进行训练,从而实现语料库的自动分类、领域自动提取. 另外,还需要对该大型语料库中的部分语料进行高精度的标注,从而实现脱离人工干预的自动评测.

参考文献:

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 5998-6008.

[2] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations. Washington DC: CLR, 2020: 11929.

[3] LIU B X, HUANG L. ParaMed: a parallel corpus for English-Chinese translation in the biomedical domain[J]. BMC Medical Informatics and Decision Making, 2021, 21(1): 258.

[4] 刘妍,熊德意. 面向小语种机器翻译的平行语料库构建方法[J]. 计算机科学, 2022, 49(1): 41-46.

[5] UTIYAMA M, ISAHARA H. A Japanese-English patent parallel corpus[EB/OL]. [2024-01-01]. https://api.semanticscholar.org/CorpusID:17728686.

[6] MORISHITA M, SUZUKI J, NAGATA M. JParaCrawl: a large scale web-based English-Japanese parallel corpus[EB/OL]. [2024-01-01]. https://arxiv.org/abs/1911.10668v2.

[7] MORISHITA M, KATSUKI C, SUZUKI J, et al. JParaCrawl v3.0: a large-scale English-Japanese parallel corpus[C]// Language Resources and Evaluation

Conference. Marseille:ELRA,2022:6704-6710.
[8] AGIĆ Ž,VULIĆ I. JW300:a wide-coverage parallel corpus for low-resource languages[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg:ACL,2019:3204-3210.
[9] 曹宜超,高翊,李淼,等. 基于单语语料和词向量对齐的蒙汉神经机器翻译研究[J]. 中文信息学报,2020,34(2):27-32,37.
[10] 贾承勋,赖华,余正涛,等. 基于枢轴语言的汉越神经机器翻译伪平行语料生成[J]. 计算机工程与科学,2021,43(3):542-550.
[11] 李响,刘洋,陈伟,等. 利用单语数据改进神经机器翻译 压缩模型的翻译质量[J]. 中文信息学报,2019,33(7):46-55.
[12] 王克非. 中国英汉平行语料库的设计与研制[J]. 中国外语,2012,9(6):23-27.
[13] MARIE B,FUJITA A. Efficient extraction of pseudo-parallel sentences from RawMonolingual data using word embeddings [C] // Annual Meeting of the Association for Computational Linguistics. Stroudsburg:ACL,2017:392-398.
[14] KAJIWARA T,KOMACHI M. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings [C] // International Conference on Computational Linguistics. Osaka:COLING,2016:1147-1158.
[15] 沙九,冯冲,周鹭琴,等. 面向司法领域的高质量开源藏汉平行语料库构建[J]. 中文信息学报,2021,35(11):51-59.
[16] ZIEMSKI M, JUNCZYS-DOWMUNT M, POULIQUEN B. The United Nations parallel corpus v1.0[C]// Language Resources and Evaluation Conference. Portorož:ELRA,2016:3530-3534.
[17] CHOE Y J,PARK K,KIM D. Word2word:a collection of bilingual lexicons for 3 564 language pairs[C]// Language Resources and Evaluation Conference. Marseille:ELRA,2020:3036-3045.
[18] SCHWENK H, CHAUDHARY V, SUN S, et al. WikiMatrix:mining 135M parallel sentences in 1620 language pairs from wikipedia[C]// Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg:ACL,2021:1351-1361.
[19] WRÓBLEWSKA A,PRZEPIÓRKOWSKI A. Towards a weighted induction method of dependency annotation[C]// Advances in Natural Language Processing. Cham:Springer,2014:164-176.
[20] XU B. NLP Chinese corpus:large scale Chinese corpus for NLP[EB/OL]. [2024-01-01]. https://doi.org/10.5281/zenodo.3402023.
[21] 韩世依,张钰晖,马云山,等. THUOCL:清华大学开放中文词库[EB/OL]. [2024-01-01]. http://thuocl.thunlp.org/.
[22] LI P,LI W,HE Z Y,et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering[EB/OL]. [2024-01-01]. https://arxiv.org/abs/1607.06275v2.
[23] CUI Y, LIU T, CHEN Z, et al. Dataset for the first evaluation on Chinese machine reading comprehension [C]// Language Resources and Evaluation Conference. Miyazaki:ELRA,2018:L18-1431.
[24] SHAO C C, LIU T, LAI Y T, et al. DRCD:a Chinese machine reading comprehension dataset [EB/OL]. [2024-01-01]. https://arxiv.org/abs/1806.00920v3.
[25] SUN K, YU D, YU D, et al. Investigating prior knowledge for challenging Chinese machine reading comprehension[J]. Transactions of the Association for Computational Linguistics,2020,8:141-155.
[26] PENG X, LI Z H, ZHANG M, et al. Overview of the NLPCC 2019 shared task: cross-domain dependency parsing[C]// Natural Language Processing and Chinese Computing. Cham:Springer,2019:760-771.
[27] MICHEL J B,SHEN Y K,AIDEN A P,et al. Quantitative analysis of culture using millions of digitized books[J]. Science,2011,331(6014):176-182.
[28] PETRONI F,LEWIS P,PIKTUS A,et al. How context affects language models' factual predictions[EB/OL]. [2024-01-01]. https://arxiv.org/abs/2005.04611.
[29] 苗传江. HNC(概念层次网络)理论导论[M]. 北京:清华大学出版社,2005.
[30] 黄曾阳. 论语言概念空间的总体结构:图灵脑理论基础之五[M]. 北京:科学出版社,2015.
[31] PENNINGTON J,SOCHER R,MANNING C. Glove:global vectors for word representation[C]// Conference on Empirical Methods in Natural Language Processing. Stroudsburg:ACL,2014:1532-1543.
[32] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data [C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg:ACL,2016:86-96.
[33] 左世亮,刘稳良. 融合多源信息的平行语料库相似句段去重算法[J]. 计算机仿真,2021,38(8):344-347,416.
[34] 刘婉月,艾山·吾买尔,敖乃翔,等. 基于熵的机器翻译伪并行语料库选择方法[J]. 现代计算机,2021,27(19):9-14,18.

(责任编辑:汪 军)