文章编号：1003-0077(2021)11-0051-09

面向司法领域的高质量开源藏汉平行语料库构建

沙 九[1,3]，冯 冲[1,3]，周鹭琴[2]，李洪政[1,3]，张天夫[1,3]，慧 慧[4]

（1. 北京理工大学 计算机学院，北京 100081；
2. 北京理工大学 信息与电子学院，北京 10081；
3. 信息智能处理与内容安全工业与信息化部重点实验室，北京 100081；
4. 甘肃省迭部县初级中学，甘肃 迭部 747400）

摘 要：面向司法领域的藏汉机器翻译面临严重的数据稀疏问题。该文从两个方面展开研究：第一，相较通用领域，司法领域的藏语需要有更严谨的逻辑表达和更多的专业术语。然而，目前藏语资源在司法领域内缺乏对应的语料、稀缺专业术语词以及句法结构。第二，藏语的特殊词汇表达方式和特定句法结构使得通用语料构建方法难以构建藏汉平行语料库。因此，该文提出一种针对司法领域藏汉平行语料的轻量级构建方法。首先，采取人工标注的方法获取一个中等规模的司法领域藏汉专业术语表作为先验知识库，以避免领域越界而产生的语料逻辑表达问题和领域术语缺失问题；其次，从全国的地方法庭官网采集实例语料数据，例如，裁判文书。优先寻找藏文实例数据，其次是汉语，以避免后续构造藏语句子而丢失特殊的词汇表达和句式结构。基于以上原则采集藏汉语料构建高质量的藏汉平行语料库，具体方法包括：爬虫获取语料，规则断章对齐检测，语句边界识别，语料库自动清洗。最终，该文构建了 16 万级规模的藏汉司法领域语料库，并通过多种翻译模型和交叉实验验证了构建的语料库具有高质量和鲁棒性等特点。另外，此语料库会开源以便相关研究人员用于科研工作。

关键词：司法领域；藏汉平行语料；数据稀疏

中图分类号：TP391 **文献标识码**：A

Constraction of High-quality and Open Source Tibetan-Chinese Parallel Corpus Judicial Domain

SHA Jiu[1,3], FENG Chong[1,3], ZHOU Luqin[2], LI Hongzheng[1,3], ZHANG Tianfu[1,3], HUI Hui[4]

(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;
2. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
3. Intelligent Information Processing and Contents Computing, Key Laboratory of MIIT, Beijing 100081, China;
4. Diebu County Junior High School, Diebu, Gansu 747400, China)

Abstract: The current Tibetan-Chinese (Ti-Zh) Machine Translation in the judicial domain suffers from a severe data-sparse issue. The high-quality Ti-Zh corpus in the judicial domain is obstructed by two issues: 1) rigorous logical expression and professional terminology vocabulary in judicial domain, and 2) unique lexical expression and specific syntactic structure of Tibetan. In this paper, we propose a lightweight Ti-Zh parallel corpus construction method for the judicial domain. First, we construct a medium-scale Tibetan-Chinese terminology glossary of the judicial domain to as the prior knowledge to avoid the missing of logical expression and domain terminology. Secondly, we collect the case data, such as judgment documents, from the official websites of Chinese courts in various places, with a priority of Tibetan case data. Finally, we build a high-quality Tibetan-Chinese parallel corpus 160,000-sentence Ti-Zh parallel corpus of the judicial domain, and we evaluate its quality and robustness via a variety of translation models and cross-validation experiments.This corpus will be provided as an open-source to for related research.

Keywords: judicial domain; Tibetan-Chinese parallel corpus; data-sparse

收稿日期：2021-02-21 **定稿日期**：2021-03-29

基金项目：国家重点研发计划(2018YFC0832104)；国家自然科学基金(61732005)

0 引言

最近神经机器翻译(Neural Machine Translation, NMT)的进步已经证明,其在某些特定领域内的翻译质量基本上可以达到专业人工翻译水准,这些翻译模型通常受益于大量的平行语料库,它们对NMT模型效果的提升最为明显。然而,平行语料库的获取和构建需要大量的时间和精力,且并非所有域或语言对都可以使用相同的数据集。为此,不少研究者通过数据增强来提升翻译质量。数据增强是一种具有显著价值的技术,其既可用于缓解数据量不足的问题,同时还可用于提升模型的稳健性。在图像分类和文本分类等应用中,几乎所有表现最好的机器学习以及深度学习模型都会用到数据增强技术。例如,启发式的数据增强方案、词汇替换、反向翻译、随机噪声注入和语法树操作等。启发式的数据增强方案往往需要依靠具有丰富领域知识的人类专家进行人工调整,但这可能导致其所得到的增强方案并非最优方案。词汇替换,这种方法试图在不改变句子主旨的前提下替换文本中的单词,包括基于词典的替换[1]、基于词向量的替换[2]、基于TF-IDF的词替换。反向翻译是一种利用单语数据生成伪平行语料的方法,具体做法是使用已有的目标语言到源语言的NMT系统,将较大规模的目标语言翻译为源语言,从而获取到新生成的伪平行语料,与已有平行语料合并后训练可提升模型效果。随机噪声注入,其思想为在文本中加入噪声,使所训练的模型对扰动具有鲁棒性。语法树操作可以解析原始句子的依存关系,使用规则对其进行转换,并生成改写后的句子。这些增强方法对性能的影响如何,目前仅针对某些特定用例进行了研究,如果能系统地比较这些方法,并分析它们对更多任务性能的影响,将是一个有意义的研究方向。

与计算机视觉(Computer Vision, CV)中使用图像进行数据增强不同,CV中图像数据增强对模型性能的提升优于自然语言处理(Natural Language Processing, NLP)中文本数据增强对模型性能的提升,其主要原因之一是对图像的一些简单操作,如将图像旋转或转换为灰度,并不会改变其语义。语义不变这一性质使数据增强成为CV研究中的一个重要工具。常规数据增强方法的局限性表明这一领域还存在很大的研究空间。常规数据增强技术依赖相关领域的专家,耗时耗力成本高昂,因此研究者开始探索自动化数据增强技术。相对于人工设计的算法,自动化数据增强更具挑战性:如何设计可学习的算法来寻找优于人类的启发式方法来增强数据;如何从实践出发解决具体问题;如何把相关理论研究落实到实践中;现有的大多数数据增强方法的关注重点基本都在于提升模型的整体性能,通常还需要在更细的粒度上关注数据的关键子集。当模型在数据的重要子集上的预测结果不一致时,又该如何利用数据增强来缩减其在相关指标上的表现差距?然而真正从根源上解决问题的方法并不多,为此本文对其中最根本、最实质性的构建高质量特定领域的平行语料库技术进行探究。

本文针对以上数据增强技术面临的挑战问题,研究NMT中通过半自动方式构建高质量特定领域的数据增强技术。本研究面向稀缺资源司法领域的藏汉平行语料在NMT中的构建,通过半自动化数据增强技术获取了大量司法领域中的藏汉语料库,进一步构建了在司法领域中具有裁判文书以及法庭判决书等子领域的庞大藏汉平行语料。此外,本研究还表明,使用CWMT2018的通用数据训练基线模型,并使用本文构建的数据集对模型进行微调,能显著提高特定领域的翻译质量。

本文的主要贡献为:

(1) 在稀缺资源司法领域公开了规模为160K的高质量藏汉平行语料库;

(2) 比较了几种识别句子边界以及句对齐的方法,用于构建NMT的平行语料库。通过实验表明,利用不同的句子边界识别技术的消融策略可以获得更可靠的可用数据,对藏汉互译的140K或160K个句子对进行微调以及预训练,可以大幅度地改善译文质量,在较大的数据集上翻译质量将继续提升。

1 相关工作

数据增强是一种普遍应用的技术,通过利用保留类标签的特定任务的数据转换来增加带标记的训练集大小。为了解决这一难题,Ratner等人[3]提出了一种方法来实现这个过程自动化。具体是采用了对抗式方法训练变换函数序列生成器,以得到与训练数据相比更加真实的增强数据。2019年谷歌大脑提出了一种自动化数据增强方法(AutoAugment)[4],该方法创建了一个数据增强策略的搜索空间,利用搜索算法选取适合特定数据集的数据增强策略。此

外，从一个数据集中学到的策略能够很好地迁移到其他相似的数据集上。发表在 ICLR 2019 上的文章[5]中介绍了几种 NLP 数据增强技术，并提出了四种简单的操作来进行数据增强，以防止过拟合并提高模型的泛化能力。

在 NMT 中，通过上下文软连接的方式来处理 NMT 中的数据增强问题[6]，这篇文章跟以往在句子中随机删除或替换单词的增强方法有所不同，它将一个单词的一种表示替换为一个分布（由语言模型提供），将该词的嵌入替换为多个语义相似的词的加权组合。由于这些词的权重依赖于被替换词的上下文信息，因此新生成的句子比以前的增强方法捕捉到了更加丰富的信息。文献[7]中，作者通过增强数据来提高和扩展神经机器翻译的鲁棒性。他们通过扩展有限的噪声数据，进一步提高 NMT 对噪声的鲁棒性及较小的模型体量。合并双语词典的方法[8]用于实现半监督神经机器翻译，该方法通过一种简单的数据增强技术来解决反向翻译中低资源环境下合成句子产生不利影响的问题。并结合了广泛使用的双语词典，解码时先逐次生成词，再合成句子达到翻译的效果。在无监督机器翻译中，通过学习双语单词嵌入来提升数据增强效果[9]，利用无监督机器翻译模型生成伪平行语料库，以提高两个嵌入空间的结构相似性，提高映射方法中双语词嵌入的质量。NEJM-enzh[10]中提出了一种在生物医学领域内构建英汉平行数据集的方法。有学者[11]尝试使用库尔德语平行语料库，具体是从多语言网站中检索可能对齐的新闻文章，并根据文字的词汇相似性和音译在不同方言和语言之间手动对齐，从而构建对应的平行语料。另外还有类似的通过多语语料库的构建方法[12]，包括德语-英语和汉语-英语的平行语料库的提取方式。

本文将半自动化数据增强技术应用到 NMT 中，通过对现有藏汉数据的筛选和预处理，并利用一些技术和方法，针对司法领域资源稀缺的藏汉互译任务，构建了庞大的平行语料库，取得了突破性的进展。

2　语料库的构建

2.1　标题构建平行语料库的基本流程

通过多语言网站获取语料数据，从中抽取语句构建平行语料的操作包括以下几个步骤：

（1）从多语言网站中抓取所需语料数据文档；

（2）从获取的文档中提取纯文本并将其规范化；

（3）根据其内容进行两种语言文档的匹配；

（4）在每个文件中将段落分解成单独的句子；

（5）随后将句子排列成句子对；

（6）对对齐的句子对进行过滤，去掉重复的和低质量的句子对。

在这 6 个步骤中，前两步基本属于工程化任务，后四步的研究正处于不断探索之中。

对于第 3 步，在 WMT16 中使用了一个用于双语文档对齐的共享任务[13]，其中最佳词条依赖于匹配不同的双语短语对[14]。对于步骤 4 而言，Read 等人[14]系统地评估了九种现有的句子边界检测工具。第 5 步的句子对齐可能是目前最具挑战性的部分。与文档对齐相比，句子对齐使用更少的文本，但有更多的排列。第 6 步也属于工程化任务。

2.2　法律领域的语料库来源

本文的语料库主要来源于中国裁判文书网站、中国民族语文翻译局、中国知网以及一些官方微信公众平台。中国裁判文书网站提供了民族语言文书，其中，含有藏文和中文的刑事案件、民事案件、行政案件、赔偿案件、执行案件以及其他案件。中国民族语文翻译局每年会定期发布每季度的新词术语，例如，“带头攻坚克难、敢于担当”等。在中国知网上可以获取法律领域相关的论文，进一步获取其藏汉双语摘要部分。还有一些官方微信公众号，如“藏汉双语法律平台”“刚察藏汉双语普法平台”以及“TBL 酥油灯青年法客”等，它们会推送相关法律立案以及法庭判决书等数据。以上四种数据的历史可追溯至 2015 年。

2.3　获取语料库并断章对齐

本文使用 Selenium[15]抓取所有可用的藏文和相应的中文文章。首先，在爬取期间，为了易于检索内容，获取的文章都按时间顺序排列。其次，对应的文档对通过超链接连接，消除了文档对齐的需要。最后，把藏文和汉文两个对应的 PDF 或者 Word 文档按段落标识符分段对齐，研究文档由相关的统计人员校对。

2.4　检测并识别语句边界

本文比较了以下三种句边界方法并取其交集：首先，本文通过统计发现，汉文的引号出现在断句之前，这使得句子的边界很容易被发现。与欧洲语言不同，“.”在汉文中不能兼用作小数点或其他语言标记。

所以本文利用{“!、?、。”}来检测识别中文句子的边界。在藏文中,同样做了统计实验,另外人工分析并归纳其规律,本文使用与中文相似的方法,通过识别“།、།།”来判断藏文句子的边界。其次,本文针对藏文和中文统一使用如下两个工具识别句子边界,分别为Read[14]等人开源的无监督句子标记器 Punkt 和 Ziemski[16]等人开源的 Eserix 系统。最终发现取以上三种方法的交集误差最小,因此取三种方法的交集进行句边界识别。

2.5 清洗和过滤语料库

本文过滤掉如下内容:①数字以及数字说明;②表格和图片及对应的说明;③短语以及短句。图 1 对比了过滤前后不同来源下藏文和汉文句子的数量。在过滤之前,大量文章中藏文句子数远超汉文句子数,这是因为在藏语中短语往往构成短句,例如,“བོད་ཀྱི་”,而在汉文中很少出现“但是。”,一般是“但是,”,经过滤之后,每篇文章中藏文和汉文句子的数量变得更接近。

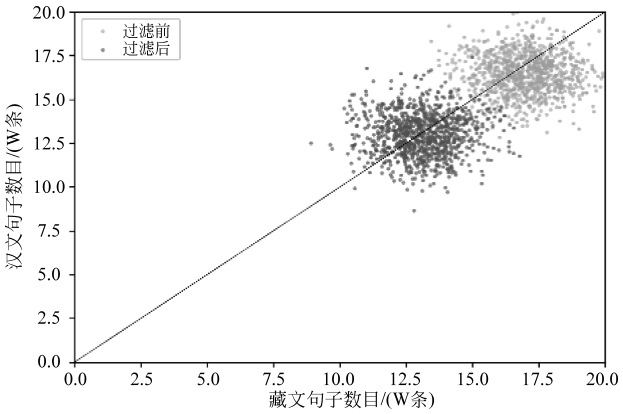

图 1 藏文和汉文在句子数目过滤前后的对比
过滤之后藏文的句子数目越来越接近汉文的句子数目

2.6 构建双语句对齐

虽然已经提出了一些句子对齐的方法[17],但这些方法在稀缺资源司法领域上的表现欠佳。本文比较了以下三种方法:基于长度对齐(Gale-Church)[18]的方法,通过假设源句和目标句的长度相似来寻找句子对;基于词典对齐(Microsoft Aligner)[18-19]的方法,把单词对与句子长度结合搜索句子对;基于翻译对齐(Bleualign)[20]的方法,将原始文本和翻译文本进行比较搜索锚定句,然后使用 Gale-Church 算法对其余的文本进行对齐。为了比较这些方法,本文人工构建了两种语言不同来源的 5 000 句测试集。表 1 显示了 Microsoft Aligner 在 5 000 句测试集上的对齐类型的分布,将近 82.64%的对齐是一对一的。

表 1 Microsoft Aligner 在 5 000 句测试集上的对齐测试结果

藏-汉	Count	Percent/%
1-0	202	4.04
0-1	204	4.08
1-1	4 132	82.64
1-2	164	3.28
2-1	298	5.96

因为大多数句子对都是一对一对齐的,对于一对多对齐的情况,所有算法的性能都会显著下降,因此在本研究中主要针对一对一对齐这一情况。精确度、召回率以及 F_1 值如图 2 所示。

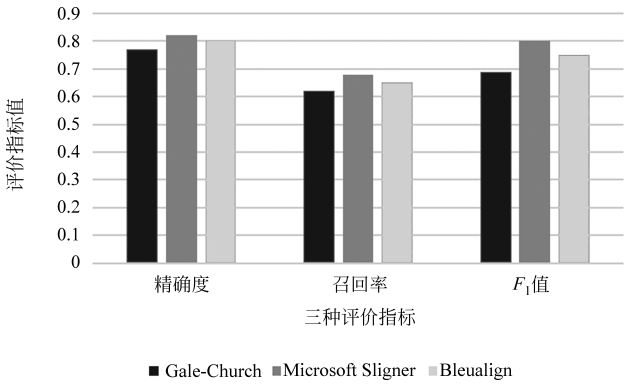

图 2 三种句子对准器在语料库上通过双向对齐(即藏对汉和汉对藏)的最终结果

其中,微软的 Aligner 获得最佳 F_1 得分,因此本文选用 Microsoft Aligner 构建句对对齐。随后所有的句子对由专业的翻译人员逐句校正,为了保证句子的流畅性,偶尔会进行句子的衔接和分割,即把一个藏文句子分成两个句子或者更多的中文句子,反之亦然。最终的数据由专门的编辑小组成员进行校对和统计,如表 2 所示,用于 NMT 的训练。

表 2 不同来源下获取的最终平行语料大小

藏-汉	数据大小/M	句子对数目/条	约占法律领域的数据/%
中国裁判文书网站	8.68	80 000	99.63
中国民族语文翻译局	5.43	50 000	89.56
中国知网	1.63	15 000	75.32
各官方微信公众号	2.17	20 000	93.23

3　实验设置

3.1　模型架构

本文使用基于 PyTorch 的 OpenNMT [21]框架，使用 Transformer-Base 模型训练，本次实验中所有的网络参数跟文献[22]中的参数设置保持一致。模型有 6 层编码器和解码器，每个输出大小为 512 个隐藏单元[16]，使用 8 个注意头和正弦位置嵌入。最后隐藏的前馈层大小为 2 048。模型总共训练了 100 000 步，训练耗时约为 1.5 天。使用 Adam 优化器[21]，其中，$Beta_1=0.9$，$Beta_2=0.98$，$Epsilon=10^{-9}$，$P_{drop}=0.1$。在藏译汉和汉译藏上使用相同的参数训练，使用 CWMT2018 官方评测工具衡量译文的质量，具体以 BLEU4 值为评测指标。本文在 8 台 Nvidia TitanX GPU 上训练模型。

3.2　训练语料

本文利用公开数据 CWMT2018 提供的新闻领域的藏汉数据集。以 2017Dev 作为开发集，2018Test 作为测试集；另外把本文构建的数据 JusticeCorpus 分割为训练集 JusticeTraining、开发集 JusticeDev 以及测试集 JusticeTest。具体实验所用的数据如表 3 所示。

表 3　两种不同训练数据的大小

藏-汉	训练集/句对	开发集/句对	测试集/句对
CWMT2018	147 434	650	1 000
JusticeCorpus	163 000	1 000	1 000

在本文中，汉文统一使用 Jieba[23]分词，随后处理为子词(Byte-Pair-Encoding，BPE)[24]。藏文先使用西北民族大学开源的 TIP-LAS[25]分词，随后按照文献[26]中音词融合的方式处理，使用 80K 的源端和目标端词汇表。最终实验所用的所有数据的汉文以 BPE 为粒度单位，藏文以音词融合为粒度单位。主要设置了五种方案进行实验：

(1) 单独用 CWMT2018 和 JusticeCorpus 数据分别训练模型作为基线系统 Topic1 和 Topic2；

(2) 先用 CWMT2018 数据进行训练，其次用 JusticeCorpus 数据进行微调 Topic3；

(3) 先用 JusticeCorpus 数据进行训练，其次用 CWMT2018 数据进行微调 Topic4；

(4) 把 CWMT2018 和 JusticeCorpus 数据合成再训练 Topic5；

在方案(2)和方案(3)中具体微调方式参照之前相关工作[27]。另外还做了一些预训练的实验用以验证本文所构建的数据是否可靠。预训练使用 BERT[28]和 XLM[29]，具体方式参考了 Weng R 等人的文章[30]。

4　实验结果

4.1　主要结果

本文的基线系统为 CWMT2018 和 JusticeCorpus 数据集上训练的模型，实验结果如表 4 所示。不难发现，单独在数据集 CWMT2018 和 JusticeCorpus 上训练时，不管在藏译汉还是汉译藏上，2017Dev 和 2018Test 在 JusticeCorpus 数据上的 BLEU 值低于 CWMT2018 数据上的值。相反，JusticeDev 和 JusticeTest 在JusticeCorpus 数据上的 BLEU 高于 CWMT2018 上的值，至少提升了 3.21 个 BLEU 值。首先，我们发现同一领域内的数据具有较强的相似之处，所以针对特定领域的测试集使用跟它相同领域的训练集训练是至关重要的。为此，构建特定的法律领域数据是很有必要的；其次，本文所构建的数据集在新闻领域的测试集上虽然 BLEU 值偏低，但是跟用新闻领域的数据集训练的结果相比，相差最多也不到 1.31 个BLEU 值。相反，在法律领域的测试集上，BLEU 值远远超越了用新闻领域训练的结果。因此，可以认为本文所构建的平行数据集具有较高的质量。

表 4　本文主要的实验结果

数据		Topic1	Topic2	Topic3	Topic4	Topic5
汉藏	2017dev	47.29	45.98	48.22	50.22	52.04
	2018Test	35.75	34.48	36.72	38.32	40.10
	JusticeDev	19.33	22.54	23.81	24.81	27.12
	JusticeTest	20.24	23.49	24.76	25.98	28.82

续表

数据		Topic1	Topic2	Topic3	Topic4	Topic5
藏汉	2017Dev	51.74	50.49	52.73	54.67	56.24
	2018Test	38.07	36.78	39.02	40.64	42.08
	JusticeDev	23.56	26.87	28.14	28.68	30.86
	JusticeTest	24.67	27.98	29.25	30.13	32.52

以上实验分别采用CWMT2018数据进行训练,在JusticeCorpus数据进行微调,以及采用JusticeCorpus数据进行训练,在CWMT2018数据进行微调。当用CWMT2018数据进行训练,以及用JusticeCorpus数据进行微调时,在2017Dev和2018Test测试集上的效果优于JusticeDev和JusticeTest测试集上的值;当用JusticeCorpus数据进行训练,以及用CWMT2018数据进行微调时,在JusticeDev和JusticeTest测试集上的效果优于2017Dev和2018Test测试集上的值。因此,虽然通过微调能提升一定的翻译效果,但是不如用该领域内的数据直接训练的效果好,故本文所构建的数据集很有价值。

从整体实验结果可以发现,所有测试集的值在藏译汉上的评分值都高于汉译藏上的评分值。因此可以认为,目标端的分词质量以及切分粒度相当重要。这是因为,藏译汉时目标端为汉文,而汉文具有很多开源的分词工具且比较成熟,但是藏语几乎没有统一成熟的开源工具,导致每个机构或者高校在藏语相关的分词任务上使用的工具各不相同,并且在很大程度上具有不同的分词粒度,直接影响了下游的工作。因此判断目标端的分词甚至比源端的分词更重要。从表4的最后一行(Topic5)可以看出,本次实验在CWMT2018和JusticeCorpus数据合并后的训练模型上译文质量最佳,比单独实验和其他的微调的结果都要好,证明只有具有高质量且大规模的平行语料训练模型,NMT才能获得最佳结果。

4.2 消融实验

本节通过预训练方法进行训练[30],因为GPT是单向语言模型,而BERT屏蔽语言模型可以获得更多的上下文信息。GPT可以对顺序信息进行建模。因此,本文用BERT初始化编码器,用GPT初始化解码器。将表5中的Topic6行中JusticeCorpus作为初始化参数语料,CWMT2018作为后期的NMT训练语料。将表5中的Topic7行中CWMT2018作为初始化参数语料,JusticeCorpus作为后期的NMT的训练语料。当编码器由BERT初始化并且解码器由GPT初始化时,BLEU值在四个测试集上都有所提升,并且本次实验结果都优于表4中的微调方法,说明这样的预训练方法比微调方法能够更有效地从预训练模型中获取更多知识。我们比较了两种不同语料作为初始化参数的方案,在Topic7中四个测试集上的实验结果始终比Topic6中的实验结果强,并且在新闻领域的测试集上明显大幅度提升了BLEU值。我们认为,通过利用领域内的数据进行预训练并初始化,其次用不同领域的数据训练,这样不仅保留了原领域内的信息特征,而且从更多的层次融合了外部领域的知识,因此让模型获得了更好的性能。由此说明,预训练方法不仅能提升译文质量,而且本文所构建的数据质量是值得信赖的。只有高质量的平行语料训练NMT,才能从输入的句子中获取语义,获得更多的上下文信息从而提升增益。当把CWMT2018作为初始化参数语料,JusticeCorpus作为后期NMT训练语料时,在藏译汉的JusticeDev测试集上,相比将JusticeCorpus作为初始化参数语料,CWMT2018作为后期的NMT训练语料,提升了1.04个BLEU值,在汉译藏上提升了2.12个BLEU值。无论在CWMT2018数据集上进行微调还是用CWMT2018数据初始化,均取得不错的效果。

表5 消融实验结果

System	数据	Topic6	Topic7
汉藏	2017dev	49.32	52.04
	2018Test	37.32	38.72
	JusticeDev	24.21	26.33
	JusticeTest	25.36	27.56

续表

System	数据	Topic6	Topic7
藏汉	2017Dev	53.37	54.95
	2018Test	40.20	41.94
	JusticeDev	29.44	30.48
	JusticeTest	30.45	31.53

4.3　译文分析

为了更好地展示本文构建的平行语料库的效果，本文手动检查了 Topic1 至 Topic7 中的输出，并在图 3 中展示了部分示例。Topic7 在四种测试集上都能较好地翻译出源文所含有的所有词。通过图 3 可知，CWMT2018 数据和 JusticeCorpus 数据合并后的训练模型最好，如图 3 中的 Topic5 行能够正常翻译“毁损、灭失、承担、损害、赔偿以及责任”。其次为预训练方法，用 CWMT2018 数据初始化参数并用 JusticeCorpus 数据进行训练，如图 3 中的 Topic7 行能够准确翻译“毁损、灭失、损害、赔偿”。翻译质量随着数据集的增长不断提高。此外，使用领域外数据进行预训练也有助于提高翻译质量，甚至在全集级别上均是如此。

藏文-汉文	句子
源文	དོན་ཚན་ སུམ་བཅུ་ དང་ གསུམ་པར་ ལྡན་ལས་ཁེད་ རིང་འགྲུལ་པ་ ས་ ཁྱེར་བ་ ལེ་ ཅ་དངོས་ ཚགས་རྒྱོག་ རྒྱུ་ དང་ ཚ་བརྡག་ ཏུ་ སོང་བ །སྐབས་ ཁྱབ་འཚོལ་ ཆོར་བ་ ཡིན་ན་ ཐོང་ཁུན་ ཀྱིས་གཡར་ ཀྱི་ འགན་འཁུར་ དགོས །
参考文	第 三百零三 条 运输 过程 中 旅客 自带 物品 毁损 、 灭失 ， 承运人 有 过错 的 ， 应当 承担 损害 赔偿 责任 。
TOPIC1	条 三百 和 第三 送 春风 中旅客 我 带 的 加快 损坏 发生 和 灭失 中 去 人夫 我 过错 丧失 的 损失 垮 补 的 负责 要
TOPIC2	第 三百零三 条 搬运 路途 中 旅客 自拿 东西 毁坏 、 灭失 ， 他人 有 错过 的 ， 应当 承担 损害 赔偿 责任 。
TOPIC3	第 三百零三 条 搬运 路途 中 旅客 自带 东西 毁坏 、 灭失 ， 运者 有 犯错 的 ， 相应 承担 损害 赔偿 责任 。
TOPIC4	第 三百零三 条 搬运 路途 中 旅客 自拿 东西 毁坏 、 灭失 ， 承运者 有 错误 的 ， 需要 承担 损害 赔偿 责任 。
TOPIC5	第 三百零三 条 运输 过程 中 旅客 自带 物品 毁损 、 灭失 ， 承运人 要 过错 的 ， 相应 承担 损害 赔偿 责任 。
TOPIC6	第 三百零三 条 运输 过程 中 旅客 自带 物品 毁损 、 灭失 ， 承运人 要 过错 的 ， 需要 承担 损失 补偿 义务 。
TOPIC7	第 三百零三 条 运输 过程 中 旅客 自带 物品 摧毁 、 灭失 ， 承运人 要 过错 的 ， 相应 承担 损害 补偿 责任 。

图 3　藏译汉上同一条测试句在 7 个不同训练方法中的实验结果

5　总结及未来工作

本文中已经证明用 CWMT2018 藏汉平行语料库训练的基准模型在稀缺资源司法领域上可泛化性是有限的。针对此问题，本文构建了一个高质量的藏汉平行数据集。利用本文所构建的数据集训练司法领域的 NMT，极大地提高了翻译质量，同时发现随着数据集的增长，翻译质量也不断得到提高。

我们的数据集大小为 160K 个句子对，这也弥补了到目前为止仅有公开的藏汉新闻领域数据的局限性。我们的数据集将会被公开供研究者们使用，以促进少数民族语言信息处理的发展。同时具有公开透明的可比性。在未来，我们计划针对藏汉语料的翻译模式进行一些语言知识调查，把稀缺资源司法领域语料库扩展到其他领域，包括政治和教育等领域，即通过某种领域的平行数据来构造另外一种领域的平行语料库。

参考文献

[1] Zhang X, Zhao J, Lecun Y. Character-level convolutional networks for text classification[C]//Proceedings of the Neural Information Processing Systems, 2015.

[2] Jiao X, Yin Y, Shang L, et al. TinyBERT: Distilling BERT for natural language understanding [J]. arXiv preprint arXiv: 190910351, 2019.

[3] Ratner A J, Ehrenberg H R, Hussain Z, et al. Learning to compose domain-specific transformations for data augmentation [J]. Advances in Neural Information Processing Systems, 2017, 30: 3239-3249.

[4] Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation strategies from data[C]//Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, 2019: 113-123.

[5] Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks [J]. arXiv preprint arXiv: 190111196, 2019.

[6] Zhu J, Gao F, Wu L, et al. Soft contextual data augmentation for neural machine translation [J]. arXiv preprint arXiv: 190510523, 2019.

[7] Li Z, Specia L. Improving neural machine translation robustness via data augmentation: Beyond back translation [J]. arXiv preprint arXiv: 191003009, 2019.

[8] Nag S, Kale M, Lakshminarasimhan V, et al. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation [J]. arXiv preprint arXiv: 200402071, 2020.

[9] Nishikawa S, Ri R, Tsuruoka Y. Data augmentation for learning bilingual word embeddings with unsupervised machine translation [J]. arXiv preprint arXiv: 200600262, 2020.

[10] Liu B, Huang L. NEJM-enzh: A parallel corpus for English-Chinese translation in the biomedical domain [J]. arXiv preprint arXiv: 200509133, 2020.

[11] Ahmadi S, Hassani H, Jaff D Q. Leveraging multilingual news websites for building a Kurdish parallel corpus [J]. arXiv preprint arXiv: 201001554, 2020.

[12] Han L, Jones G J, Smeaton A F. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora [J]. arXiv preprint arXiv: 200510583, 2020.

[13] Gomes L, Lopes G. First steps towards coverage-based document alignment[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 697-702.

[14] Read J, Dridan R, Oepen S, et al. Sentence boundary detection: A long solved problem? [C]//Proceedings of COLING 2012: Posters, 2012: 985-994.

[15] Salunke S S. Selenium webdriver in Python: Learn with examples [M]. CreateSpace Independent Publishing Platform, 2014.

[16] Ziemski M, Junczys-dowmunt M, Pouliquen B. The united nations parallel corpus v1. 0[C]//Proceedings of the 10th International Conference on Language Resources and Evaluation, 2016: 3530-3534.

[17] Simard M, Plamondon P. Bilingual sentence alignment: Balancing robustness and accuracy [J]. Machine Translation, 1998, 13(1): 59-80.

[18] Gale W A, Church K. A program for aligning sentences in bilingual corpora [J]. Computational Linguistics, 1993, 19(1): 75-102.

[19] Repar A, Podpečan V, Vavpetič A, et al. TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment [J]. Terminology International Journal of Theoretical and Applied Issues in Specialized Communication, 2019, 25(1): 93-120.

[20] Sennrich R, Volk M. MT-based sentence alignment for OCR-generated parallel texts[C]//Proceedings of the 9th Corference of the Association for Machine Translation in the Americas, 2010: 1-10.

[21] Klein G, Kim Y, Deng Y, et al. OpenNMT: Open-source toolkit for neural machine translation [J]. arXiv preprint arXiv: 170102810, 2017,

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systerms, 2017: 6000-6010.

[23] SUN J. Jieba chinese word segmentation tool[CP/OL]. https://github. com//txsjy/jieba. [2020-08-09].

[24] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [J]. arXiv preprint arXiv: 150807909, 2015.

[25] 李亚超, 江静, 加羊吉, 等. TIP-LAS: 一个开源的藏文分词词性标注系统 [J]. 中文信息学报, 2015, 29(6): 203-7.

[26] 沙九, 冯冲, 张天夫, 等. 多策略切分粒度的藏汉双向神经机器翻译研究 [J]. 厦门大学学报 (自然科学版), 2020, 59(2): 213-219.

[27] Guo J, Tan X, Xu L, et al. Fine-tuning by curriculum learning for non-autoregressive neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 7839-7846.

[28] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 181004805, 2018.

[29] Lample G, Conneau A. Cross-lingual language model pretraining [J]. arXiv preprint arXiv: 190107291, 2019.

[30] Weng R, Yu H, Huang S, et al. Acquiring knowledge from pre-trained model to neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 9266-9273.

沙九(1994—)，硕士研究生，主要研究领域为自然语言处理、机器翻译。
E-mail：shajiu@bit.edu.cn

冯冲 (1977—)，通信作者，博士，教授，主要研究领域为机器翻译、信息抽取和知识图谱。
E-mail：fengchong@bit.edu.cn

周鹭琴(1995—)，硕士研究生，主要研究领域为数字信号处理、图像识别、目标检测。
E-mail：zhouluqin1217@163.com

热烈祝贺中国中文信息学会多位理事荣获 2020 年度国家科学技术奖！

2021 年 11 月 4 日，2020 年度国家科学技术奖励名单正式公布，奖项共评选出 264 个项目、10 名科技专家和 1 个国际组织。其中，中国航空工业集团有限公司顾诵芬院士和清华大学王大中院士分获国家最高科学技术奖。**中国中文信息学会副理事长王海峰博士，常务理事李涓子教授，理事唐杰教授、赵世奇博士，专委会主任李凤华研究员，作为主要完成人的项目，荣获多个奖项**。主要获奖情况如下：

【国家技术发明奖二等奖】

项目名称：知识增强的跨模态语义理解关键技术及应用
主要完成人：
王海峰(北京百度网讯科技有限公司)
吴　华(北京百度网讯科技有限公司)
赵世奇(百度在线网络技术(北京)有限公司)
贾　磊(北京百度网讯科技有限公司)
丁二锐(北京百度网讯科技有限公司)
孙　宇(北京百度网讯科技有限公司)

项目名称：物联网系统数据安全关键技术及应用
主要完成人：
马建峰(西安电子科技大学)
李凤华(中国科学院信息工程研究所)
郑志彬(华为技术有限公司)
吴　昊(北京交通大学)
沈玉龙(西安电子科技大学)
翁　健(暨南大学)

【国家科学技术进步奖二等奖】

项目名称：智能型科技情报挖掘和知识服务关键技术及其规模化应用
主要完成人：唐杰，李涓子，杨红霞，许静芳，许斌，邢春晓，张阔，刘德兵，高博，张帆进
主要完成单位：清华大学，北京搜狗科技发展有限公司，阿里巴巴(中国)有限公司