面向司法领域的高质量开源藏汉平行语料库构建

沙九[1,4] 周鹭琴[2,4] 冯冲[1,4*] 李洪政[1,4] 张天夫[1,4] 慧慧[3,5]
1.北京理工大学计算机学院,北京市海量语言信息处理与云计算应用工程技术研究中心
2.北京理工大学信息与电子学院
3.甘肃省迭部县初级中学
4.北京市海淀区中关村南大街5号
5.甘肃省迭部县兴迭东街38 号
{shajiu,zhouluqin,fengchong,lihongzheng,zhangtainfu}@bit.edu.cn,huihui@163.com

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。本文将从两个方面展开研究: 第一，相比于通用领域，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。为此，本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，我们采取人工标注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，我们从全国的地方法庭官网采集实例语料数据，例如裁判文书。我们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。我们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，我们构建了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的高质量特点和鲁棒性。另外，此语料库会开源以便于相关研究人员用于科研工作。

关键词: 司法邻域 ; 藏汉平行语料 ; 稀缺资源

A High-quality Open Source Tibetan-Chinese Parallel Corpus Construction of Judicial Domain

Jiu Sha[1,4] Luqin Zhou[2,4] Chong Feng[1,4*] Hongzheng Li [1,4] Tianfu Zhang[1,4] Hui Hui[3,5]
1.Beijing Engineering Research Center of High Volume Lange Information Processing and Cloud Compution Applications,School of Computer Science & Technology,Beijing Institute of Technology
2.School of Information and Electronics Beijing Institute of Technology
3.Diebu County Junior High School, Gansu Province
4.No. 5, Zhongguancun South Street, Haidian District, Beijing
5.No. 38, Xingdie East Street, Diebu County, Gansu Province
{shajiu,zhouluqin,fengchong,lihongzheng,zhangtainfu}@bit.edu.cn,huihui@163.com

Abstract

To date, the Tibetan-Chinese (Ti-Zh) Machine Translation in the judicial domain confronts a data-sparse severe problem. In this work, we tackle the problem from two aspects: 1) judicial Tibetan needs more rigorous logical expression and professional terminology vocabulary than the public domain. However, there hardly exists the high-quality Ti-Zh corpus in the judicial domain, which contains professional terminology and syntactic structure. 2) It is challenging to construct a Ti-Zh parallel corpus due to the unique lexical expression and specific syntactic structure. To this end, 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 be our prior knowledge, which can avoid the logical expression

第十九届中国计算语言学大会论文集，第499页-第508页，海南，中国，2020年10月30日至11月1日。
(c) 2020 中国中文信息学会计算语言学专业委员会

and domain terminology missing problems caused by the out-of-domain phenomenon. Secondly, we collect the instance data, such as judgment documents, from the official websites of Chinese courts in various places. To avoid losing the Tibetan lexical expressions and syntactic structures, we firstly search for Tibetan case data, followed by Chinese. Based on the above principles, we build a high-quality Tibetan-Chinese parallel corpus，which consists of the following methods: crawling corpus, document segmentation alignment detection, sentence boundary recognition, automatic corpus cleaning. Lastly, we construct a 160,000 Ti-Zh parallel corpus of the judicial domain, and we evaluate the quality and robustness of the constructed corpus by performing a variety of translation models and cross-validation experiments. Besides, this corpus will be an open-source to provide to other researchers for related research.

Keywords: Judicial Domain , Tibetan-Chinese Parallel Corpus , Data-sparse

1 引言

最近神经机器翻译(Neural Machine Translation, NMT)的进步已经证明，在某些特定领域内，翻译质量基本上可以达到专业的人工翻译，这些翻译模型通常受益于大量的平行语料库，它们的效果在神经机器翻译模型中最为明显。然而，平行语料库的获取和构建需要大量的时间和精力，而且并非所有域或语言对都可以使用相同的数据集，为此，有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，它既可用于缓解数据量不足的问题，同时还用于提升模型的稳健性。在图像分类和文本分类等应用中，使用的几乎所有表现最好的机器学习以及深度学习等模型都会用到数据增强技术。例如：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，但这可能导致所得到的增强方案是次优的。词汇替换，这种方法试图在不改变句子主旨的情况下替换文本中的单词，包括基于词典的替换(Zhang X, Zhao J, LeCun Y., 2015)，基于词向量的替换(Jiao et al., 2019)，基于TF-IDF的词替换。反向翻译，利用NMT来解释文本，同时重新训练含义。文本表面转换, 使用正则表达式简单的模式匹配转换。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。语法树操作将解析和生成原始句子的依赖关系树，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仅针对某些特定用例进行了研究。系统地比较这些方法并且分析它们对许多任务性能的影响将是一项有趣的研究。然而与计算机视觉(Computer Vision, CV)中使用图像进行数据增强不同，在自然语言处理(Natural Language Processing, NLP)中文本数据增强是非常罕见, 其主要原因之一是图像的一些简单操作，如将图像旋转或将其转换为灰度，并不会改变其语义。语义不变变换的存在使其增强成为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进步空间。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。自动化数据增强领域具有挑战性的难题，从人工设计到自动搜索算法可以发现：1.不同于执行次优的人工搜索，我们要如何设计可学习的算法来寻找优于人类设计的启发式方法的增强策略？2.从实践到理论理解：虽然在实际应用中增强技术的设计研发进展飞速发展，可是由于缺乏分析工具，仍然难以挖掘这类技术的优点。该如何从理论上理解实践中使用的各种数据增强技术？3.从粗粒度到细粒度的模型质量保证：现有的大多数数据增强方法的关注重点都是提升模型的整体性能，通常还需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在相关指标上的表现差距？然而真正从根源上解决问题的并不多，为此本文探究最根本最实质性的构建高质量特定领域的平行语料库。本研究中，我们带着以上自动化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。我们研究针对稀缺资源司法领域的藏汉平行语料在神经机器翻译中的构建，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，我们还表明，使用CWMT2018的通用数据训练基线模型，并使用我们构建的数据集对模型

©2020 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

进行微调，将比基线模型显著提高翻译质量。我们的贡献如下：我们在稀缺资源司法领域公开了160K大小的高质量藏汉平行语料库。这是本文的主要贡献。我们比较了几种识别句子边界以及句对齐的方法，用于构建NMT的平行数据集。我们的实验表明，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。我们发现，对藏汉互译的140K或160K个句子对进行微调以及预训练，可以将大幅度的提升译文质量，在较大的数据集上翻译质量继续提升。

2 相关工作

数据增强是一种普遍存在的技术，通过利用保留类标签的特定于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(2017)提出了一种方法来自动化这个过程，通过学习生成序列模型在用户指定的转换函数使用生成对抗的方法。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足够真实的增强数据。2019年谷歌大脑提出了一种自动数据增强方法(AutoAugment) (Cubuk et al. , 2019)，该方法创建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它相似的数据集上。发表在ICLR 2019上的(Wei et al. , 2019)文章中介绍了几种NLP数据增强技术，具体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中数据增强问题(Gao et al., 2019)，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，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更加丰富的信息。在超越反向翻译这篇文章(Li et al., 2019) 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。他们以扩展有限的噪声数据，进一步提高NMT对噪声的鲁棒性，同时保持较小的模型。探索合并双语词典的方法(Nag et al., 2020)，以实现半监督神经机器翻译，通过一种简单的数据增强技术来解决在反向翻译中对低资源环境下合成句子产生不利影响的问题。结合了广泛使用的双语词典，解码时以逐次生成词从而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(Nishikawa et al., 2020)，利用无监督机器翻译模型生成的伪平行语料库，以促进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。在NEJM-enzh (Liu B and Huang L, 2020)中提出了一个在生物医学领域内构建英汉平行数据集。(Han L, Jones G J F, Smeaton A F., 2020)中介绍了多语语料库的构建方法，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，通过分析比较离散而实际应用的可获取数据，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，取得了突破性的进展。

3 语料库的构建

本节描述了我们构建句子对齐语料库的步骤。

3.1 构建平行语料库的基本流程

通过多语言网站获取语料库，从语料库中抽取语句构建平行语料包括以下几个步骤：(1)从多语言网站中抓取所需语言的文档；(2)从爬取的文档中提取纯文本并规范化；(3)两种语言的文档根据其内容进行匹配；(4)在每个文件中，段落被分解成单独的句子；(5)句子随后被排列成句子对；(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。其前两步基本是工程任务，而后四步正在不断的探索之中。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(Gomes and Lopes, 2016)，其中最佳词条依赖于匹配不同的双语短语对(Read et al., 2012)。对于步骤(4)而言，Read等(2012)系统地评估了9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。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，相比第(5)步简单容易实现。

3.2 法律领域内的语料库来源

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一种民族语言文书，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

(c) 2020 中国中文信息学会计算语言学专业委员会

期发布每季度的新词术语，例如：“带头攻坚克难、敢于担当”。在中国知网上可以获取法律领域相关的论文，进一步获取具有藏汉双语摘要部分。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，推送的相关法律立案以及法庭判决书等数据。以上四种种数据的历史可追溯至2015年。

3.3 获取语料库并断章对齐

我们使用Selenium (2014)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文章按段落标识符分段对齐，研究文章由相关的统计人员校对。

3.4 检测并识别语句边界

我们比较了以下三个方法并取交集: 首先，在汉文中，我们通过统计并发现，汉文的引号出现在断句之前，这使得很容易发现句子的边界。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。所以我们利用“！ 、？ 、。”来检测识别汉文句子的边界。在藏文中，我们同样做了统计实验，另外人工分析并归纳，我们使用跟汉文相似的方法，通过识别“.<ja>、,、.<jb>、.<jc>、.<jd>”(转写为拉丁)来判断藏文句子的边界。其次，我们针对藏文和汉文统一使用如下两个工具进行识别句子边界，分别为Read (2012)等人开源的无监督句子标记器Punkt和Ziemski (2016)等人开源的Eserix系统。最终，我们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。

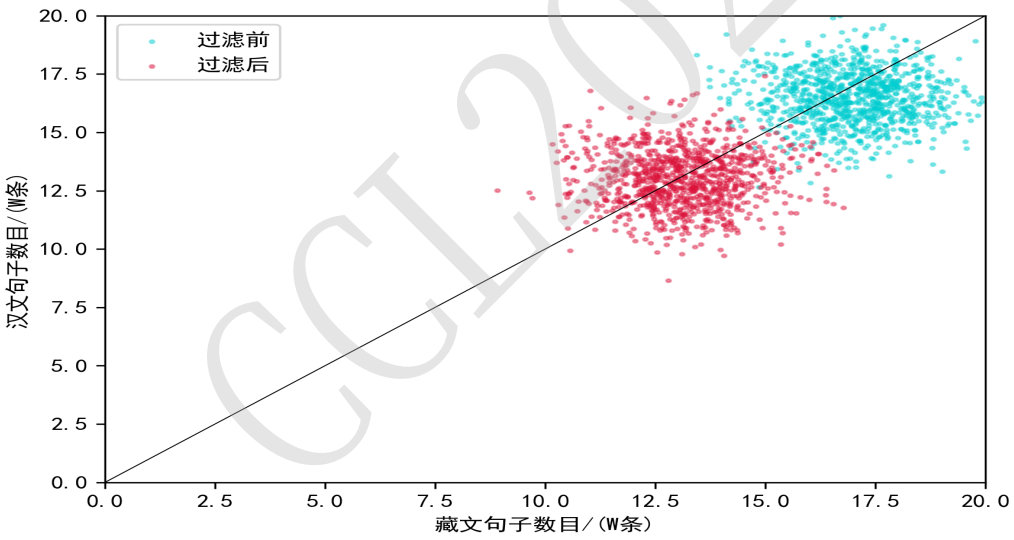

Figure 1: 藏文和汉文在句子数目过滤前后的对比，通过过滤之后藏文的句子数目越来越接近汉文的句子数目。

3.5 清洗和过滤语料库

对于这两种语言，我们过滤掉如下内容：(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远超过了汉文句子数，用蓝色点表示。是因为在藏语中短语往往构成短句，例如“on kyang |”(转写为拉丁)，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。

3.6 构建双语句对齐

虽然已经提出了一些句子对齐的方法(Simard M and Plamondon P, 1998; Repar et al.,, 2019)，但这些方法在稀缺资源司法领域的语言上表现缺乏共识。我们比较了以下三种方法: 基

(c) 2020 中国中文信息学会计算语言学专业委员会

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	80000	99.63%
中国民族语文翻译局	5.43	50000	89.56%
中国知网	1.63	15000	75.32%
各官方微信公众号	2.17	20000	93.23%

Table 1: 不同来源下获取的最终平行语料大小。

于长度对齐(Gale-Church)(Gale and Church, 1993)，它是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)(Moore, 2002)，是把单词对应与句子长度结合搜索句子对；基于翻译对齐(Bleualign)(Sennrich and Volk, 2010)，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，

藏-汉	Count	Percent
1-0	202	4.04%
0-1	204	4.08%
1-1	4132	82.64%
1-2	164	3.28%
2-1	298	5.96%

Table 2: Microsoft Aligner在5000句测试集上的对齐测试结果，其中大多数是1-1对齐。

我们人工构建了两种语言不同来源的5000句测试集。如表2显示了Microsoft Aligner在5000句测试集上对齐类型的分布。将近82.64%的对齐是一对一的。 因为大多数句子对都是一对一对齐

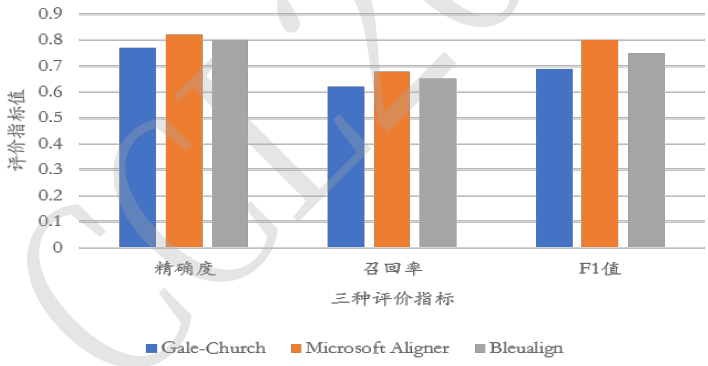

Figure 2: 三种句子对准器在语料库上通过双向对齐即藏对汉和汉对藏的最终结果。

的，对于一对多对齐，所有算法的性能都会显著下降，因此在本研究中我们主要针对一对一对齐。精确度，召回率以及F1值的分数如图2所示。其中微软的Aligner获得最佳F1分，所以我们选择Microsoft Aligner用于构建句对对齐。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汉文句子，反之亦然。最终的数据由专门的编辑小组成员校和统计如表1所示，便用于NMT的训练。

4 实验设置

4.1 模型架构

我们使用基于PyTorch的OpenNMT (Klein et al.,, 2017)框架，使用Transformer-base模型训练，本次实验中所有的网络参数跟论文(Vaswani et al.,, 2017)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(Ziemski et al.,, 2016)，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2048。模型总共训练了100,000步，训练耗时约

(c) 2020 中国中文信息学会计算语言学专业委员会

为1.5天。使用Adam优化器(Klein et al.,, 2017)，其中β_1=0.9，β_2=0.98，ϵ=10^{-9}，P_{drop}=0.1。我们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。我们在8台Nvidia TitanX GPU上训练模型。

4.2 训练语料

我们利用公开数据*CWMT*2018提供的藏汉数据集，该数据都属于新闻领域。以2017*Dev*作为开发集，2018*Test*作为测试集；另外我们把本文构建的数据*OurCorpus*分割为训练集*OurTraining*、开发集*OurDev*以及测试集*OurTest*。具体实验所用的数据如表3所示。在本文中，汉文统一使用Jieba(Sun J, 2012)分词，随后处理为子词(Byte-Pair-

藏-汉	训练集(句对)	开发集(句对)	测试集(句对)
CWMT2018	147434	650	1000
OurCorpus	163000	1000	1000

Table 3: 两种不同训练数据的大小

Encoding，BPE)(Sennrich et al.,, 2015)。藏文先使用西北民族大学开源的TIP-LAS(李亚超, 江静等, 2015)分词，随后按照(沙九，冯冲等, 2020)中音词融合的方式处理，使用80K的源端和目标端词汇表。最终我们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。我们主要设置了五种方案进行实验：(1)单独用*CWMT*2018和*OurCorpus*数据分别训练模型作为基线系统(*topic*1&*topic*2)；(2)先用*CWMT*2018数据进行训练，其次用*OurCorpus*数据进行微调(*topic*3)；(3)先用*OurCorpus*数据进行训练，其次用*CWMT*2018数据进行微调(*topic*4)；(4)把*CWMT*2018和*OurCorpus*数据合成再训练(*topic*5)。在(2)和(3)中具体微调方式我们参照了(Guo et al.,, 2019)。另外我们还做了一些预训练的实验加以验证我们所构建的数据是可靠的。预训练我们使用BERT (Devlin et al., , 2018)和XLM (Lample and Conneau, 2019)，具体方式我们参考了(Weng et al.,, 2019)文章。

5 结果

5.1 主要结果

实验结果如表4所示，我们的基线系统为*CWMT*2018和*OurCorpus*数据集上训练的模型，我们不难发现，单独在数据集*CWMT*2018和*OurCorpus*上训练时，不管在藏译汉还是汉译藏上，2017*Dev*和2018*Test*在*OurCorpus*数据上的BLEU值偏低于*CWMT*2018数据上的值，相反*OurDev*和*OurTest*在*OurCorpus*数据上的BLEU值胜于*CWMT*2018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用跟它相同领域的训练集训练是至关重要的，为此，我们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远远超越了用新闻领域训练的结果。从此，我们发现本文所构建的平行数据集具有较高的质量。

用*CWMT*2018数据进行训练，其次用*OurCorpus*数据进行微调和用*OurCorpus*数据进行训练，其次用*CWMT*2018数据进行微调在本文中也提升了翻译的质量。当*CWMT*2018数据进行训练，其次用*OurCorpus*数据进行微调时，在2017*Dev*和2018*Test*测试集上的效果优于*OurDev*和*OurTest*测试集上的值；当*OurCorpus*数据进行训练，其次用*CWMT*2018数据进行微调时，在*OurDev*和*OurTest*测试集上的效果优于2017*Dev*和2018*Test*测试集上的值。为此，我们发现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。为此我们肯定本文所构建的数据集是很有价值的。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。为此我们认为，目标端的分词质量以及切分粒度相当重要。因为，当藏译汉时，目标端为汉文，而汉文具有很多开源的分词工具并比较成熟，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各不相同，并且很大程度上具有不同的分词粒度。这直接影响了下游的工作。所以我们判断目标端的分词甚至比源端的分词更重要。在表4的最后一行*topic*5可以看出，本次实验通过*CWMT*2018和*OurCorpus*数据合并后的训练模型上译文质量最佳，相比之前单独实验和其

(c) 2020 中国中文信息学会计算语言学专业委员会

System	汉⇒ 藏		汉⇒藏		藏⇒ 汉		藏⇒汉	
	2017*Dev*	2018*Test*	*OurDev*	*OurTest*	2017*Dev*	2018*Test*	*OurDev*	*OurTest*
*topic*1	47.29	35.75	19.33	20.24	51.74	38.07	23.56	24.67
*topic*2	45.98	34.48	22.54	23.49	50.49	36.78	26.87	27.98
*topic*3	48.22	36.72	23.81	24.76	52.73	39.02	28.14	29.25
*topic*4	50.22	38.32	24.81	25.98	54.67	40.64	28.68	30.13
*topic*5	52.04	40.10	27.12	28.82	56.24	42.08	30.86	32.52

Table 4: 本文主要的实验结果，“*topic*1”为单独用*CWMT*2018数据训练的模型；“*topic*2”为单独用*OurCorpus*数据训练的模型；“ *topic*3”先用*CWMT*2018数据进行训练，其次用*OurCorpus*数据进行微调；“*topic*4”先用*OurCorpus*数据进行训练，其次用*CWMT*2018数据进行微调；“*topic*5”把*CWMT*2018和*OurCorpus*数据合成再训练的实验结果。

System	汉⇒ 藏		汉⇒藏		藏⇒ 汉		藏⇒汉	
	2017*Dev*	2018*Test*	*OurDev*	*OurTest*	2017*Dev*	2018*Test*	*OurDev*	*OurTest*
*topic*6	49.32	37.32	24.21	25.36	53.37	40.20	29.44	30.45
*topic*7	52.04	38.72	26.33	27.56	54.95	41.94	30.48	31.53

Table 5: 用BERT初始化编码器，用GPT初始化解码器，分别用*CWMT*2018作为初始化参数语料*OurCorpus*作为后期NMT训练语(*topic*6)；用*OurCorpus*作为初始化参数语料*CWMT*2018作为后期NMT训练语(*topic*7)。

他的微调都要好，为此，我们证明只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。

5.2 消融实验

本节我们按照(Weng et al.,, 2019)中的预训练方法训练，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。为此，本文用BERT初始化编码器，用GPT初始化解码器。在表5中的*topic*6行中*OurCorpus*作为初始化参数语料，*CWMT*2018作为后期的NMT训练语料。在表5中的*topic*7行中*CWMT*2018作为初始化参数语料，*OurCorpus*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分数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。通过这样的预训练方法比微调方法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中实验结果有所不同，在*topic*7中四个测试集上的实验结果总比*topic*6中的实验结果强，并且在新闻领域的测试集上明显提升了大幅度的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，同时更多的层次融合了外部领域内的知识，让模型获得了更好的性能。为此说明，预训练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。只有高质量的平行语料训练NMT，才能从输入句子中获取语义，获得更多的上下文信息从而提升增益。当*CWMT*2018作为初始化参数语料，*OurCorpus*作为后期NMT训练语料时，在藏译汉的*OurDev*测试集上相比*OurCorpus*作为初始化参数语料，*CWMT*2018作为后期的NMT训练语料提升了1.58个BLEU值，在汉译藏上提升了2.12个BLEU值。我们的数据不管在*CWMT*2018数据的上进行微调还是用*CWMT*2018数据初始化都取得不错的效果。

5.3 译文分析

为了更好的看到本文构建的平行语料库的效果，我们手动检查了*topic*1至*topic*7中的输出，并在表3中展示了一些示例。我们的*topic*7在四种测试集上都能较好的翻译出源文所含有的所有词。在表3中不难发现，当*CWMT*2018数据和*OurCorpus*数据合并后的训练模型最好，如表3中的第*topic*5行能够正常的把“毁损、灭失、承担、损害、赔偿以及责任”都能准确的翻译。

(c) 2020 中国中文信息学会计算语言学专业委员会

其次为预训练方法，用$CWMT$2018数据初始化参数并用$OurCorpus$数据进行训练，如表3中的第$topic$7行能够准确翻译“毁损、灭失、损害、赔偿”。随着数据集的增长，翻译质量不断提高。此外，使用领域外数据进行预培训有助于提高翻译质量，甚至在全集级别上也是如此。

藏文-汉文	句子
源文	དོན་ཚན་ སུམ་བརྒྱ་ དང་ གསུམ་པར་ སྐྱེལ་ འདྲེན་བྱེད་ རིང་འགྲུལ་པ་ ས་ ཁྱེར་བ་ འི་ ཅ་དངོས་ ཆག་སྐྱོན་ བྱུང་བ་ དང་ ཆ་བརླག་ ཏུ་ སོང་བ །སྐྱེལ་པ ས་ ཐུས་འཛོལ་ ཤོར་བ་ ཡིན་ན་ གྱོང་གུན་ རྒྱུད་གསབ་ 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 。

Figure 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。

6 总结及未来工作

在本文中，我们已经证明，用CWMT2017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。为此，我们针对稀缺资源司法领域的藏汉平行语料库，构建了一个高质量的藏汉平行数据集。我们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，同时我们发现随着数据集的增长，翻译质量也将不断提高。我们的数据集大小为160K个句子对，这也弥补了到目前为止公开的只有新闻领域CWMT数据的局限性。我们的数据集将会公开便于研究者使用，使得让研究少数民族语言信息处理快速发展。同时具有一个公开透明的可比性。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。我们将稀缺资源司法领域语料库扩展到其他领域上，包括政治和教育等领域。我们通过一种领域来构造另外一种领域内的平行语料库。

参考文献

Zhang, Xiang and Zhao, Junbo and LeCun, Yann. 2015. *Character-level convolutional networks for text classification.* Advances in neural information processing systems.

Jiao, Xiaoqi and Yin, Yichun and Shang, Lifeng and Jiang, Xin and Chen, Xiao and Li, Linlin and Wang, Fang and Liu, Qun. 2019. *Tinybert: Distilling bert for natural language understanding* arXiv preprint arXiv:1909.10351

Gao, Fei and Zhu, Jinhua and Wu, Lijun and Xia, Yingce and Qin, Tao and Cheng, Xueqi and Zhou, Wengang and Liu, Tie-Yan. 2019. *Soft Contextual Data Augmentation for Neural Machine Translation* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

Li, Zhenhao and Specia, Lucia. 2019. *Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back Translation* arXiv preprint arXiv:1910.03009

Nag, Sreyashi and Kale, Mihir and Lakshminarasimhan, Varun and Singhavi, Swapnil. 2020. *Incorporating Bilingual Dictionaries for Low Resource Semi-Supervised Neural Machine Translation.* arXiv preprint arXiv:2004.02071.

(c) 2020 中国中文信息学会计算语言学专业委员会

Nishikawa, Sosuke and Ri, Ryokan and Tsuruoka, Yoshimasa. 2020. *Data Augmentation for Learning Bilingual Word Embeddings with Unsupervised Machine Translation*. arXiv preprint arXiv:2006.00262.

Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, Christopher Ré 2017. *Learning to Compose Domain-Specific Transformations for Data Augmentation*

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le. 2019. *AutoAugment: Learning Augmentation Strategies From Data*. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 113-123.

Wei, Jason W and Zou, Kai. 2019. *Eda: Easy data augmentation techniques for boosting performance on text classification tasks*. arXiv preprint arXiv:1901.11196.

Christian Buck and Philipp Koehn. 2016. *Findings of the wmt 2016 bilingual document alignment shared task*. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 554–563.

Lu ıs Gomes and Gabriel Pereira Lopes. 2016. *First steps towards coverage-based document alignment*. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 697–702.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. *Sentence boundary detection: A long solved problem?*. In Proceedings of COLING 2012: Posters, pages 985–994.

Sagar Shivaji Salunke. 2014. *Selenium Webdriver in Python: Learn with Examples*. CreateSpace Independent Publishing Platform.

Kingma, Diederik P and Ba, Jimmy. 2014. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980。

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. *OpenNMT: Opensource toolkit for neural machine translation*. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. *The united nations parallel corpus v1. 0*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3530–3534.

Liu B, Huang L. 2020. *NEJM-enzh: A Parallel Corpus for English-Chinese Translation in the Biomedical Domain*. arXiv preprint arXiv:2005.09133, 2020.

William A Gale and Kenneth W Church. 1992. *A program for aligning sentences in bilingual corpora*. Computational linguistics, 19(1):75–102.

Robert C Moore 2002. *Fast and accurate sentence alignment of bilingual corpora*. In Conference of the Association for Machine Translation in the Americas, pages 135–144. Springer.

Rico Sennrich and Martin Volk. 2010. *Mt-based sentence alignment for ocr-generated parallel texts*. In The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010).

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

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

沙九; 冯冲; 张天夫; 郭宇航; 刘芳2020. 多策略切分粒度的藏汉双向神经机器翻译研究. 厦门大学学报(自然科学版)

Sun J 2012. *Jieba chinese word segmentation tool*. Accessed: Jun, 2012, 25: 2018.

李亚超, 江静, 加羊吉, 等. 2015. *TIP-LAS:* 一个开源的藏文分词词性标注系统. 中文信息学报, 2015, 29(6): 203-207.

Sennrich R, Haddow B, Birch A. 2015. *Neural machine translation of rare words with subword units*. arXiv preprint arXiv:1508.07909, 2015.

(c) 2020 中国中文信息学会计算语言学专业委员会

Vaswani A, Shazeer N, Parmar N, et al. 2017. *Attention is all you need*. Advances in neural information processing systems. 2017: 5998-6008.

Guo J, Tan X, Xu L, et al. 2019. *Fine-Tuning by Curriculum Learning for Non-Autoregressive Neural Machine Translation*. arXiv preprint arXiv:1911.08717, 2019.

Weng R, Yu H, Huang S, et al. 2019. *Acquiring Knowledge from Pre-trained Model to Neural Machine Translation*. arXiv preprint arXiv:1912.01774, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018.

Lample G, Conneau A. 2019. *Cross-lingual Language Model Pretraining*. arXiv preprint arXiv:1901.07291, 2019.

Han L, Jones G J F, Smeaton A F. 2020. *MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora*. arXiv preprint arXiv:2005.10583, 2020.

CCL2020

(c) 2020 中国中文信息学会计算语言学专业委员会