

基于 ALBERT 的藏汉神经机器翻译研究

孙义栋¹

SUN Yidong

摘要

藏汉翻译能够极大促进藏汉科技文化交流以及教育文化事业的发展,但一直以来都面临着低资源语种平行语料匮乏的困境,因此通过将预训练语言模型 ALBERT 引入到藏汉翻译中,利用大规模的藏文单语数据集训练 ALBERT 模型来改善藏汉翻译效果。实验表明,引入预训练模型 ALBERT 可以有效提升藏汉神经机器翻译效果。

关键词

藏汉机器翻译; 预训练; ALBERT

doi: 10.3969/j.issn.1672-9528.2024.05.028

0 引言

伴随着互联网技术的普及和发展,各个民族之间的沟通交流与日俱增。我国作为一个统一的多民族国家,一直以来都将促进各民族文化共同发展和融合作为工作重心。藏汉翻译技术的提高对于推进我国藏区信息社会发展、促进各民族文化交流交融、增强社会团结具有重要意义。机器翻译是指通过计算机实现两种自然语言之间的翻译,其发展主要经过三个阶段,分别是基于规则的机器翻译(rule-based system)、统计机器翻译(statistics-based machine translation, SMT)以及目前主流的神经机器翻译(neural machine translation, NMT)。神经机器翻译是指采用深度学习技术,通过端到端的方式实现机器翻译。近年来,由于科技水平的发展,对于英汉等主流大语种而言,高质量、多领域对齐的平行语料的获取成本逐渐降低,使得基于大规模平行语料的神经机器翻译在主流语种翻译上大放异彩。然而对于低资源小语种,受制于市场规模较小、数据标注成本高昂等客观因素,导致高质量平行语料的匮乏,神经机器翻译的潜能尚未完全发挥,距离英汉等主流大语种仍然存在较大的差距。因此,不少学者和研究者逐渐将目光移向预训练语言模型,通过大规模的单语数据训练语言模型,学习低资源语言的语义和语法信息,再将其迁移到下游的机器翻译任务中,以此弥补高质量平行语料不足的缺点,从而提升翻译质量。

1 国内外研究现状

ALBERT^[1]是谷歌提出的一种预训练语言模型,是BERT^[2]的压缩改进版本,所以需要首先对BERT做一个介绍。BERT主模型由三部分构成,分别为嵌入层、编码器和

池化层。嵌入层即词嵌入,是将输入的序列转换为连续分布式表示。它包含三种组件:嵌入变换(embedding)、层标准化(layer normalization)、随机失活(dropout)。编码器是对每个序列进行非线性表示,将其中的特征提取出来。池化层是将[CLS]标记对应的表示提取出来,并且对它做一定的变换以此作为整个序列的表示,并返回所有的标记表示。BERT的特征抽取结构为双向的Transformer^[3],它提出了两种预训练任务来进行预训练。一种是掩码语言模型(masked language model, MLM)的预训练方式,即随机从输入语料上屏蔽掉一些单词,然后通过上下文来预测该单词,形象化理解就是让计算机去做完形填空;另一种是下一句预测任务(next sequence prediction, NSP),即判断句子B是否是句子A的下文,针对每一个训练样本,从语料库中选择句子A和B来组成,50%概率B是A的下一条句子,50%概率B不是A的下一条句子,让BERT进行二分类预测学习句子相关性。ALBERT基本结构框架与BERT几乎一致,不同之处在于使用嵌入层参数因式分解和层参数共享来减少模型中的参数以及用句序预测(sentence order prediction, SOP)替换掉BERT中的NSP任务。嵌入层参数因式分解是将词嵌入矩阵(维度为 $V \times H$)分解为维度 $V \times E$ 和 $E \times H$ 的两个矩阵,形象化理解就是首先将词汇投影到一个低维词嵌入空间,然后再映射到隐藏空间,从而达到大幅降低参数数量的目的。层参数共享则是多个网络层使用相同的参数,具体包括前馈神经网络层的参数以及注意力机制的参数,同样有效减少了模型参数量,并提升了模型的稳定性。在句序预测任务中,其正样本与NSP正样本相同,为正常顺序的两个句子,负样本则是调换顺序后的连续两个句子,通过去除句子话题性干扰只保留学习句子连贯性,有效提升了模型在多种下游任务

1. 马鞍山学院 安徽马鞍山 243000

[基金项目] 安徽省高校自然科学研究重点项目(2023AH052619)

中的表现。

ALBERT 在文本分类、命名实体识别、机器翻译等多种下游自然语言处理 (natural language processing, NLP) 任务上都显示出了强大的能力。文献 [4] 将自注意力机制与 ALBERT 相融合进行命名实体识别; 文献 [5] 将 ALBERT 与注意力特征分割融合网络相结合进行文本情感分析。目前将 ALBERT 有效应用于藏汉神经机器翻译还缺乏足够的研究, 因此, 本文旨在探究如何将 ALBERT 预训练模型应用于藏汉神经机器翻译中, 同时也对其他低资源语种翻译起到一定的启发作用。

2 语料库的构建

2.1 语料库的来源

文章数据来源主要分为两部分, 一部分来自第十七届全国机器翻译大会 (CCMT 2021) 所提供的 15 万条左右藏汉平行语句, 用于训练藏汉神经机器翻译模型; 另一部分利用 Python 网络爬虫获取到约 1 GB 左右的藏文单语数据, 用于训练藏文 ALBERT 模型。

2.2 数据预处理

首先进行数据过滤, 包括长度过滤、长度比限制、语种识别、去重。其次进行符号标准化 (对数据中的字符表示或者大小写等进行统一, 包括全角转半角、大小写转换和中文的简繁体转化等), 在本文中, 使用 mosesdecoder^[6] 进行符号标准化。最终得到清洗后约 14 万条数据集。接着进行分词, 在本文中, 汉文使用北大 pkuseg^[7] 分词, 并在此基础上使用字节对编码 (byte pair encoding, BPE)^[8]。藏文使用 TIP-LAS 分词^[9], 同样在 TIP-LAS 基础上使用 BPE。最后划分训练集语句约 13 万条, 验证集与测试集各 5000 条。

3 实验设计

3.1 基线模型架构与参数

本文使用基于百度飞桨的 PaddleNLP 框架, 设置两个基线模型, 即 Transformer-base 和 Transformer-big, 方便对照实验。Transformer-base 的模型参数设定如下: 词向量维度为 512, 编码器和解码器层数设置为 8 层, 使用 8 个注意头和正弦位置嵌入, 前馈网络中隐藏层大小为 2048, 句子过滤长度限制为 256, 为防止过拟合, Dropout 参数值设置为 0.1, 训练中使用 Adam 优化算法, 学习率初始值设置为 2.0, 使用 Vaswani 文献 [3] 中的学习率衰减策略, 模型训练中使用参数共享。

Transformer-big 与 Transformer-base 的模型结构类似, 只不过加深了网络层数。具体不同参数设置为: 词向量维度为

1024, 编码器和解码器层数设置为 16 层, 每个输出大小为 1024 个隐藏单元, 使用 16 个注意头和正弦位置嵌入, 前馈网络中隐藏层大小为 4096, Dropout 参数值设置为 0.3。其余参数设置与 Transformer-base 相同。

所有模型在解码时均使用束搜索算法生成目标语句, beamsize 设置为 5, batch-size 设置为 32。采用 BLEU-4 值作为评测指标, BLEU 值计算如公式 (1) 所示。其原理是通过采用 n-gram 匹配的方式来评定翻译结果与参考译文之间的相似度, 即机器翻译的结果越接近参考译文就认定它的质量越高。所有实验均在 Ubuntu 操作系统下使用 RTX 2080Ti 显卡完成。

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N W_n \cdot \log P_n\right) \quad (1)$$

式中: BP 代表长度惩罚因子, 避免短句子得到更高的值; P_n 表示修正后的 n-gram 精度得分; W_n 表示权重值。

3.2 ALBERT 模型架构与参数

ALBERT 使用遮蔽语言模型 MLM 和 Transformer 的编码器来生成深度的双向语言特征向量。ALBERT 应用主要分为两个任务阶段, 分别为预训练阶段和下游任务微调阶段。在预训练阶段, 主要利用大规模藏文单语语料来训练 ALBERT 模型, 其模型结构如图 1 所示。

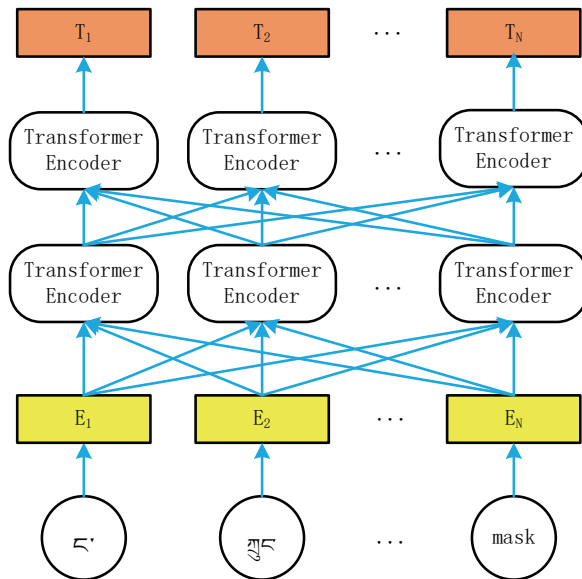


图 1 ALBERT 模型结构

本文在预训练阶段使用基于 PyTorch 框架的 ALBERT 来训练藏文 ALBERT 模型, 最大句子长度为 512, 使用 12 个隐藏层, 隐藏层神经元个数为 768, 采用 12 头注意力机制, 隐藏层激活函数为 GELU, 同时隐藏层去除 dropout, 具体实现过程如下。

首先利用爬虫得到 1 GB 左右大小的藏文单语数据, 再

经过 ALBERT 上下句处理后数据量为 10 GB 左右,最后再利用 10 GB 数据训练藏文 ALBERT 模型。最终 loss 值为 1.52, mask_acc 为 73%, sop_acc 为 87%, 对应结果生成 ALBERT 模型、字符词表以及模型参数配置 3 个文件。最后根据藏汉翻译进行下游任务微调,通过生成的字符词表来构建输入藏文句子的词向量,分别替换掉 Transformer-base 和 Transformer-big 模型中训练好的词向量,从而完成引入。

4 实验结果与分析

4.1 实验分析

在 Transformer-base 和 Transformer-big 上分别引入预训练模型 ALBERT 的实验结果如表 1 所示,其中验证集与测试集各 5000 条。

表 1 引入 ALBERT 前后实验效果对比

Ti-Ch	Dev	Test
Transformer-base	28.57	28.55
Transformer-big	30.75	30.76
Base +ALBERT	31.40	31.40
Big +ALBERT	34.03	34.02

由表 1 可知,对于基线模型 Transformer-base 和 Transformer-big,通过应用训练好的藏文 ALBERT 模型都可以较大程度改善藏汉神经机器翻译效果,在 Transformer-base 上引入 ALBERT 后,验证集和测试集平均提升了约 2.84 个 BLEU 值;在 Transformer-big 上引入 ALBERT 后,验证集和测试集平均提升了约 3.27 个 BLEU 值。这是由于 ALBERT 通过将大量藏文单语语料利用起来,将语料中的知识迁移进了预训练模型的 embedding 中,并取代原基线模型训练的词向量的缘故。但实验效果无法达到论文中的提升程度,可能是由于训练的藏文单语数据较少导致的。

4.2 译文分析

为了直观对比藏汉翻译在 Transformer-big 上引入 ALBERT 前后的翻译效果,从 5000 条测试集中随机选取 2 条藏文绘制成表,引入 ALBERT 前后翻译结果对比如表 2 所示。

表 2 引入 ALBERT 前后译文效果对比

藏文源语句	Transformer-big 译文	ALBERT 译文	参考译文
ཁྱེད་ཀྱིས་བྱ་བ་འདི་ནི་གྲགས་པ་དམ་འཛིན་ལག་པོ་བྱེད་དགོས།	你要抓住这份工作的机会。	你需要好好珍惜这个工作机会。	你要好好珍惜这个工作机会。
ཁྱེད་ཀྱིས་རྒྱུ་མར་དམ་པར་དྲ་ཞིབ་མཐུན་དགོས། ད་ནི་མ་ཅིང་།	你下次要仔细,不能之后出错。	你下次一定要细心,不能再出错。	你下次一定要细心,不能再出错了。

在第一句中,Transformer-big 模型将“གྲགས་པ་དམ་འཛིན་ལག་པོ་(好

好珍惜)”翻译成了“抓住”,并不准确,而且存在词语缺失和语序颠倒等问题,但在使用 ALBERT 之后,预测结果基本与参考译文一致。在第二句中,Transformer-big 译文与参考译文在表达上差距较大,对于藏文源语句中的“རྒྱུ་མར”,其本意是“之后”,但是在这句话中间接地表达了“再”的意思。这说明 Transformer-big 并没有很好地抽取共性特征,并且存在一定的漏译问题“དམ་པར་(一定)”。在引入 ALBERT 之后,译文质量大幅提升,基本与参考译文意思一致,在具体表达上也更贴近参考译文。

从总体上来看,在以 Transformer 为基线的神经机器翻译上引入 ALBERT 预训练语言模型,在一定程度上可以有效提升藏汉翻译质量。

5 结语

本文选取经典的神经机器翻译模型 Transformer-base/big 作为基线模型,通过训练预训练语言模型藏文 ALBERT 并迁移到下游基线模型上,实验证明通过预训练语言模型可以有效提升藏汉神经机器翻译水平,在 Transformer-base 上,验证集和测试集平均提升 2.84 个 BLEU 值;在 Transformer-big 上,验证集和测试集平均提升 3.27 个 BLEU 值。其本质是通过预训练语言模型,利用大规模藏文单语数据来学习藏文的语义和语法信息,从而改善高质量藏汉平行语料缺乏的困境,属于另一种形式上的数据增强。通过本研究,希望能对其他低资源语种机器翻译有所启发,促进各民族文化交流创新交融。

参考文献:

[1]LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. (2019-09-26)[2024-02-01].<https://doi.org/10.48550/arXiv.1909.11942>.

[2]DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2024-02-01].<https://doi.org/10.48550/arXiv.1810.04805>.

[3]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].(2017-06-12)[2024-02-01].<https://doi.org/10.48550/arXiv.1706.03762>.

[4]游乐圻,裴忠民,罗章凯.融合自注意力的 ALBERT 中文命名实体识别方法[J].计算机工程与设计,2023,44(2):605-611.

[5]叶星鑫,徐杨,罗梦诗.基于 ALBERT-AFSFN 的中文短文本情感分析[J].计算机工程与应用,2022,58(12):170-176.

一种基于 FPGA 的时序控制优化方法

禹 芳¹
YU Fang

摘 要

同一种控制硬件模块处在项目不同层级中, 外接设备待控制模块型号和数量各不相同, 发挥不同的时序控制作用, 且不同层级中的控制硬件模块对外所使用的 RS422 接口和光纤接口使用重合率基本上达到了 80% 以上, 然而在 FPGA 这个硬件平台上做软件设计这块, 涉及不同层级中的控制硬件模块管脚复用问题, 很难在一个程序上实现所有的时序关系, 进而要对不同的程序维护, 这会增加很大的工作量。针对这个问题, 提出一种在软件上统型设计的方法, 不仅能防止人为固化程序出差错, 还有利于后期维护。

关键词

时序控制; FPGA; 控制模块; 程序维护; 统型设计

doi: 10.3969/j.issn.1672-9528.2024.05.029

0 引言

随着现在系统的集成度越来越高, 处理速度也要求越来越快。现场可编程门阵列 (field programmable gate array, FPGA) 技术应运而生, FPGA^[1-2] 具有丰富的布线资源、可重复编程、集成度高及可并行处理等特点, 受到广大设计者的欢迎。FPGA 的输入输出引脚多, 且配合上驱动控制芯片后, 可灵活实现输入输出方向的控制选择以及输入输出的形式——差分或者单端。单端信号形式比如 TTL, 可作为短距离传输, 且它占用更少的低频连接器引脚, 可在有限的空间内排布更多; 差分信号比如 RS422^[3-4], 它传输距离远, 抗干扰能力强, 传输速率高, 可达兆比特级别, 但占用连接器的空间是 TTL 的 2 倍。FPGA 的高速口速率可达到吉比特, 在

大数据传输时可很大程度上缩减时间, 配合上光电转换模块, 有助于高速数据的长距离稳定传输。

基于以上特性, 使用 FPGA 做主处理器, 搭建控制模块硬件平台, 是当前的设计主流。基于 FPGA 的硬件平台——控制模块, 做时序处理^[5], 不仅处理速度快, 而且并行处理的特性完美解决了时序对齐问题。针对频点多、通道多的情况, 需要进行高时效性的校正, 用 FPGA 做处理完全契合。

赛灵思的 FPGA 资源包括可配置逻辑模块 CLB、DSP、嵌入式 RAM 及时钟管理等。它不仅资源丰富, 而且软件设计上使用的 vivado 软件平台, 对于初学者来说, 简单易上手。采用的硬件描述语言为 verilog 语言或者 VHDL 语言, verilog 语言类似于 C 语言, 简单直接^[6-8]。

文章针对控制模块在某对抗项目中多个且发挥不同作用的情况, 提出一种在程序上做优化设计, 减少程序版本的方法。

1. 中国电子科技集团公司第三十八研究所 安徽合肥 230088

[6]KOEHN P, HOANG H, BIRCH A, et al.Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Companion Volume. Stroudsburg, PA: Association for Computational Linguistics, 2007:177-180.

[7]LUO R, XU J, ZHANG Y, et al. Pkuseg: a toolkit for multi-domain chinese word segmentation[EB/OL].(2019-06-27)[2024-03-01].<https://doi.org/10.48550/arXiv.1906.11455>.

[8]SENNRICH R, HADDOW B, BIRCH A.Neural machine translation of rare words with subword units[C]//54th Annual

meeting of the Association for Computational Linguistics,vol. 3.long papers.Stroudsburg,PA:Association for Computational Linguistics, 2016: 1715-1725.

[9]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):203-207.

【作者简介】

孙义栋 (1997—), 男, 安徽马鞍山人, 硕士, 助教, 研究方向: 机器翻译。

(收稿日期: 2024-03-04)