何思兰,左家莉,朱洪坤,等.文言文-现代文神经机器翻译的研究[J].江西师范大学学报(自然科学版),2023,47(5):483-489.
HE Silan,ZUO Jiali,ZHU Hongkun,et.al. The study on the neural machine translation of ancient China-modern Chinese [J].Journal of Jiangxi Normal University(Natural Science),2023,47(5):483-489.

文章编号:1000-5862(2023)05-0483-07

文言文-现代文神经机器翻译的研究

何思兰,左家莉*,朱洪坤,王明文
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要: 中国古典文献汗牛充栋,它们是中国文化的瑰宝,但现代人想要理解这些文献极为困难,人工翻译它们更是不可能完成的任务.因此,该文研究了文言文-现代文的神经机器翻译,通过应用 Seq2Seq 模型和 Transformer 模型,考察了训练语料规模对文言文-现代文翻译性能的影响.研究结果发现:基于现有的训练语料规模,分词与否会极大影响 Seq2Seq 模型的性能.此外,若训练语料和测试语料的文体不同,则模型的性能也会受到影响.

关键词: 文言文-现代文神经机器翻译; Seq2Seq 模型; 翻译
中图分类号: TP 181 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2023.05.07

0 引言

中华文明源远流长,文献典籍汗牛充栋.这些典籍蕴含着中华民族的精神追求,是理解中华文化精神内涵的基础.只有通过这些历史文献回望历史、以史为鉴,才能更好地理解过去并思考未来.然而,由于汉语经过了长时间的历史演变,所以想要阅读、理解这些经典的历史文献并非易事,需要极为深厚的古典文化修养.但若要将这些文献人工翻译为现代文,则其工作量大到无法想象,研究文言文-现代文的机器翻译因而也极具现实意义.

随着深度学习的蓬勃发展,主流的机器翻译模型也由早期的基于规则的机器翻译(rules-based machine translation,RBMT)[1]和统计机器翻译(statistical machine translation,SMT)[2]转为神经机器翻译(neural machine translation,NMT)[3].2013 年,W. Zaremba 等[4]首先提出利用神经网络进行机器翻译.文献[5-8]陆续提出了基于编码器-解码器结构的神经机器翻译模型.2016 年,Google 公司公布了 Google 的神经机器翻译模型[9],该模型通过在层之间引入残差连接以解决深度模型梯度爆炸和梯度消失的问题,这使得机器翻译的水平提升到了一个新的高度.J.Gehring 等[10]提出了基于卷积神经网络的编码器-解码器模型,该模型大幅提升了翻译准确度和翻译速度.2017 年,D. Bahdanau 等[3]提出基于自注意力机制的 Transformer 模型,在模型的训练速度和翻译质量上均取得了较大提升.

如前文所述,最近 10 年神经机器翻译取得了重大的突破,受其鼓舞,该文主要研究文言文-现代文的神经机器翻译模型,探索了 Seq2Seq 模型和 Transformer 模型在不同语料上的翻译表现.研究表明:在目前平行语料匮乏的情况下,在大规模的不同时代的混合语料上训练的模型的性能远比在小规模的单个来源的语料上训练的模型的性能更好.这表明在语料规模有限的情况下因语言变迁而导致的语言鸿沟对翻译质量的影响有限.此外,若训练语料和测试语料的文体不同,则翻译的质量会受到影响.

1 相关工作

下面将介绍文言文-现代文神经机器翻译的国内外相关研究.

收稿日期:2022-10-19
基金项目:国家自然科学基金(61866018,62266023)资助项目.
通信作者:左家莉(1982—),女,江西宜春人,副教授,博士,主要从事自然语言处理的研究.E-mail: zjl@jxnu.edu.cn

自 2013 年以来,众多学者就已经开始致力于构建纯粹的神经机器翻译模型的工作.早期的基于循环神经网络的机器翻译模型[6-7]将输入压缩为一个固定长度的向量表示,这极大地限制了模型的学习能力,且存在模型很难解决长期依赖的问题.D. Bahdanau等[8]试图引入注意力机制解决上述问题,同时使用双向的循环神经网络来学习上下文信息.文献[9-11]均使用多层神经网络进行翻译任务.在 Transformer[11]中至关重要的是自注意力机制的使用,由于其在翻译任务上优异的表现,所以后续提出了一系列基于 Transformer 的改进方法.如 Feng Yang 等[12]在传统 Transformer 神经机器翻译模型中引入译文流畅度和忠实度的评估模块用于指导训练.进一步,文献[13]提出 Seer Forcing 用于解决神经机器翻译中的 Teacher Forcing 问题.Chen-kehai 等[14]认为在源句子中的内容词(content words)比功能词(function words)能表达更多含义,利用词频信息区分出内容词且基于内容词可用于改善翻译结果.

文言文-现代文的机器翻译自有特点,因为文言文与现代文是历经时间的演变,所以不同时期同一语言产生不同变种.按时间先后顺序汉语可划分为上古汉语、中古汉语、近古汉语和现代汉语 4 个大类[15].上古汉语、中古汉语和近古汉语统称为古代汉语.类似地,英语的历史发展也可以分为 3 个阶段,它们分别是古英语(old english)、中世纪英语(middle English)和现代英语(modern English)[16].H. Jhamtani 等[17]构建了带有丰富复制(copy-enriched)机制的 Seq2Seq 模型,研究了如何将现代英语文本翻译成莎士比亚时期风格英语.K. Carlson 等[18]收集了不同时期各种版本的《圣经》作为训练语料,研究了不同版本的《圣经》之间的风格迁移,实质上也可以看作是不同版本《圣经》之间的翻译.

对于文言文-现代文的神经机器翻译模型的研究,近年来已有一些相关的工作.例如 Yang Zhichao 等[19]提出了白话文-古诗的翻译任务,使用无监督机器学习的方法以应对翻译不足(under-translation)和过度翻译(over-translation)问题.为了缓解语言变迁引发的语言鸿沟问题,E. Chang 等[20]在文言文-现代文的翻译任务中引用了年代预测作为辅助任务.文献[21-22]指出大规模平行语料库的缺乏限制了对古今汉语机器翻译的研究,而预训练模型能在一定程度上有效缓解数据匮乏问题.V. Dankers 等[23]发现 NMT 模型通常无法准确翻译成语,并且会过度生成组合的直译问题,并根据实验结果将成语处理为组合表达的倾向有助于成语的直译.Yang Zinong 等[24]使用预训练 Guwen-BERT 对文言文的翻译进行微调,结果表明预训练模型能有效改善翻译结果.

文言文-现代文的神经机器翻译模型的研究方兴未艾,总的来说,其发展受限于文言文-现代文平行语料的匮乏,而且如前述所言,古代汉语不是一种一成不变的语言,它经过了漫长的发展,语言的变迁所产生的语言鸿沟也使得文言文-现代文的翻译更为困难.最近东北大学在网络上分享了一个他们收集整理的文言文(古文)-现代文平行语料这(https://github.com/NiuTrans/Classical-Modern),内容涵盖了大部分中国经典古籍著作,形成共计约 96 万文言文-现代文句对.这在一定程度上缓解了平行语料匮乏的问题,也是展开本文研究的数据基础.

2 基准模型

2.1 基于 Seq2Seq 的神经机器翻译模型

2014 年,W.Zaremba 等[4]提出用神经网络将序列映射到序列(Seq2Seq)的模型,并成功将其应用于机器翻译任务.Seq2Seq 由编码器(encoder)和解码器(decoder)2 个部分组成,编码器模块首先对输入序列进行语义编码和特征抽取映射到语义向量,解码器模块根据该语义向量生成预测结果.早期 Seq2Seq[4-6]模型用于机器翻译主要以 RNN 为基础,但是此类传统 RNN Seq2Seq 模型在处理长序列文本时还存在不足,极其容易产生梯度消失和梯度爆炸问题.为解决上述问题,学者们提出如 LSTM(long short time memory)[25]和 GRU(gated recurrent unit)[8]等解决办法.D. Bahdanau 等[8]在 Seq2Seq 模型框架中引入注意力机制,可以捕捉全局信息,突破了由长序列而导致的模型限制.带注意力机制的 Seq2Seq 模型框架图如图 1 所示.

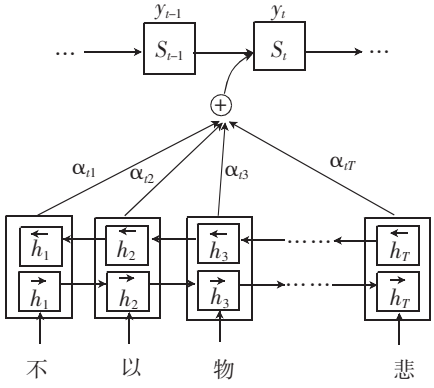

图 1 Seq2Seq 模型框架图

整个计算过程为先用编码器读取输入源句子

$\boldsymbol{x}=(x_1, x_2, \cdots, x_t)$ 和上一个时间步的隐藏状态 h_{t-1}，再计算第 t 个时间步的隐藏状态 h_t:

$$h_t = f_{\text{enc}}(x_t, h_{t-1}).$$

解码器按顺序处理 $t-1$ 时间步预测词 y_{t-1}和隐藏状态 s_{t-1}，得到隐藏状态 s_t:

$$s_t = f_{\text{dec}}(y_{t-1}, s_{t-1}).$$

在注意力机制下，各个元素按其重要程度加权求和得到 c_i，即 $c_i = \sum_{j=1}^{t} \alpha_{ij} h_j$.

公式 $e_{ij}=f_{\text{score}}(s_i, h_j)$ 通过计算解码器的隐藏状态与编码器的隐藏状态得到一个分数 e_{ij}. e_{ij}表示待编码词与该句中其他词之间的相关性，若相关性越大，则给予该词更大的权重，即 α_{ij}值也就越大. α_{ij}使用 softmax 来表示，其计算公式为 $\alpha_{ij} = \exp(e_{ij}) / \sum_{k=1}^{t} \exp(e_{ik})$. $\hat{s}_t = \tan h(W_c [c_t; s_t])$ 将 c_i和解码器的 s_t拼接起来. 最后通过 softmax 函数计算预测输出单词的概率，计算公式为 $p(y_t | y<t, \boldsymbol{x}) = \text{softmax}(W_s \hat{s}_t)$. 最终得到输出句子 $\boldsymbol{y} = (y_1, y_2, \cdots, y_t)$.

2.2　基于 Transformer 的神经机器翻译模型

A. Vaswani 等[11] 提出的仅由自注意机制构成的 Transformer 模型，在机器翻译任务上表现突出. Transformer 模型摒弃 RNN 的顺序生成模式，仅使用自注意力机制获取全局特征并能够较好地利用大型并行计算工具 GPU，大幅提高训练速度. 其模型框架图如图 2 所示.

Transformer 模型编码器由 N 个相同的网络层组成. 每个网络层包含多头自注意力模块(multi-head self-attention module)和基于位置的全连接前向反馈神经网络(position-wise fully connected feed forward network) 2 个子层. 注意力具体计算公式为

$$f_{\text{Attention}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} / \sqrt{d_k})\boldsymbol{V},$$

其中 $\boldsymbol{Q}$ 是查询矩阵，$\boldsymbol{K}$ 是键矩阵，$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$是点乘，计算对于 $\boldsymbol{Q}$ 在 $\boldsymbol{V}$ 上的注意力权重通过 $\sqrt{d_k}$ 进行缩放. 同理多头注意力公式为

$$f_{\text{MultiHead}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = f_{\text{Concat}}(h_1, h_2, \cdots, h_h)\boldsymbol{W}^O,$$

$$h_i = f_{\text{Attention}}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V),$$

$W_i^Q \in \mathbf{R}^{d_{\text{model}} d_Q}$, $W_i^K \in \mathbf{R}^{d_{\text{model}} d_k}$, $W_i^V \in \mathbf{R}^{d_{\text{model}} d_v}$, $W_i^O \in \mathbf{R}^{d_{\text{model}} d_v}$. d_{model}表示输入输出 token 的向量维度. 基于位置的全连接前向反馈神经网络计算公式为 $f_{\text{FFN}}(\boldsymbol{x}) = \max(0, \boldsymbol{x}\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2$. 2 个子层均进行残差连接和层正则化处理.

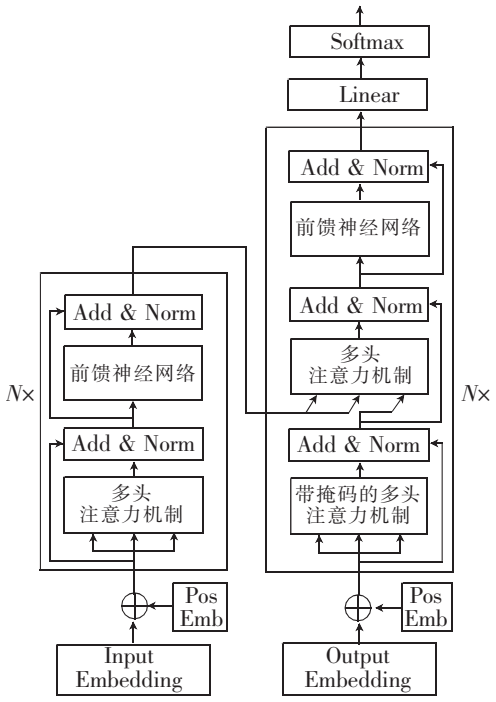

图 2　Transformer 模型框架图

解码器与编码器相似，同样由 N 个相同的网络层构成. 每一网络层包括 3 个子层: 第 1 层是遮罩多头注意力子层，其输入仅包含当前位置之前的词语信息，这样使得解码器输出只能基于已输出的部分按顺序解码; 第 2 层是多头自注意力层; 第 3 层是全连接前向反馈神经网络层. 每个子层同样进行残差连接和层正则化处理.

3　实验

3.1　实验数据

本文的实验主要是基于 3 个语料进行的，它们分别是来自东北大学开源(https://github.com/NiuTrans/Classical-Modern)的短篇章和资治通鉴平行语料(简称为混合语料)、史记平行语料以及从混合语料中分离出的资治通鉴语料. 混合语料由《资治通鉴》大部分内容和《左传》《论语》《吕氏春秋》等多篇短篇章文献混合组成. 众所周知，《史记》是西汉史学家司马迁所撰的纪传体史书，它上迄上古传说中的黄帝时代，下至汉武帝太初四年，历史跨度长达3 000年. 《资治通鉴》为北宋史学家司马光主编的一部多卷本编年体史书，它从周威烈王二十三年(公元前 403 年)起，至五代后周世宗显德六年(公元 959 年)，记载了十六朝 1 362 年的历史. 这 3 个语料的划分如表 1 所示. 其中混合语料 1 和混合语料 2 规模大小相同，仅平行句对的顺序不同.

本文还从东北大学开源的《新唐书》和《徐霞客游记》平行语料中随机抽取了一定数量的平行句对用于测试.《新唐书》是记载唐朝历史的纪传体史书,在北宋仁宗年间由宋祁、欧阳修、范镇和吕夏卿等合撰.《徐霞客游记》是明代地理学家徐霞客所创作的一部散文游记.从《新唐书》中选用了 1 236 个句对,《徐霞客游记》中选用了 2 275 个句对.

表 1 训练集、验证集和测试集划分表 万个

语料	训练集	验证集	测试集
史记	1.4	0.17	0.17
资治通鉴	4.8	0.60	0.60
混合语料 1	27.9	3.49	3.49
混合语料 2	27.9	3.49	3.49

本文选择上述数据集主要出于以下考虑: 首先,《史记》和《资治通鉴》作为中国古代最具代表的史书,历朝历代研究者众多,可为后续的研究引为参考; 其次,《史记》成书于西汉武帝太初元年至汉武帝征和二年,大约是公元前 104 年至公元前 91 年,而《资治通鉴》成书于北宋神宗元丰年间的公元 1085 年,它们相距 1 000 多年,选择这 2 个语料可用于考察文言文发展的历时性问题,也就是语言变迁对翻译质量的影响; 再者,《新唐书》不同于《史记》和《资治通鉴》,是一部仅记录唐代历史的断代史,但其修书时间又是在北宋仁宗年间,因此也可以作为考察语言变迁的参考,而《徐霞客游记》的文体则完全不同于其他语料,它是一部散文随笔,可以考察翻译模型在不同文体的语料上的表现.

3.2 实验设置

将甲言分词(https://github.com/jiaeyan/Jiayan)后的词表大小设置为 32 000.Seq2Seq 模型采用带有注意力机制的双向 LSTM,词嵌入大小设置为 256 维,隐藏层大小设置为 512 层,同时设置 dropout 率 e_{dropout} 为 0.1,句子最大长度为 120.Transformer 模型在很大程度上遵循传统 Transformer 的实验设置,编码器和解码器层数 L 设置为 6,多头注意力机制中含有 8 个头,同时设置 e_{dropout} 为 0.3,最大句子长度与 Seq2Seq 一致.另外使用 BLEU[24](Bilingual Evaluation Understudy)值来评估模型的性能.

3.3 实验结果及分析

首先基于表 1 所示的数据集,对 Seq2Seq 模型和 Transformer 模型进行了实验,具体的实验结果如表 2 所示.由于单个的史记和资治通鉴语料规模太小,难以训练 Transformer 模型,所以未列出 Transformer 在这 2 个语料上的结果.使用文言文分词工具甲言对数据集进行了分词,并对比了单字的结果.由表 2 可知: Seq2Seq 在混合语料上的结果比在史记和资治通鉴 2 个语料上的结果好很多.其主要原因是这 2 个语料的规模太小导致模型训练不够.

表 2 Seq2Seq 和 Transformer 模型的翻译结果对比

语料	模型	
	Seq2Seq	Transformer
史记	11.76(J) 18.90(S)	
资治通鉴	8.88(J) 26.05(S)	
混合语料 1	17.92(J) 22.96(J)	23.94(S) 21.28(S)
混合语料 2	18.36(J) 23.67(S)	22.91(J) 21.34(S)

注: 行表示用于模型训练的语料,列表示模型类别,J 表示使用了甲言进行分词,S 表示单字作为输入.下文同.

为了考察文言文的历时变迁是否会影响翻译模型.进一步分别在史记、徐霞客、新唐书和资治通鉴 4 个语料上测试在史记和资治通鉴 2 个语料上训练得到的 Seq2Seq 模型,结果如表 3 所示.实验结果显示: 在史记语料上训练的模型用于翻译其他几个语料结果都不理想,而在资治通鉴语料上训练的模型结果相对更好.其原因可能在于被誉为“史家之绝唱,无韵之离骚”的《史记》相较其他史书,风格更为独特,使得在史记语料上学习的模型不适用于其他语料.另外的原因是,《史记》成书时代早于其他测试语料,《新唐书》和《资治通鉴》包含了大量在《史记》中没有的内容.《资治通鉴》由于较《史记》和《新唐书》更晚,且包含的内容有重合,所以资治通鉴语料模型的结果相对更好.

表 3 Seq2Seq 模型在不同的训练-测试语料对上的结果对比

训练集	测试集			
	史记	资治通鉴	新唐书	徐霞客游记
史记	11.76(J) 18.90(S)	3.72(J) 7.68(S)	2.58(J) 4.74(S)	2.40(J) 8.00(S)
资治通鉴	8.88(J) 14.66(S)	13.01(J) 26.05(S)	10.84(J) 22.15(S)	5.25(J) 8.27(S)

表 3 还表明: 2 个模型在徐霞客语料上的结果都不好.这其中既有徐霞客游记本身文体不同的原

因，也可能是因为徐霞客语料的时代更晚.资治通鉴语料模型在新唐书语料上的表现优于其在史记测试语料上的表现，既可能是因为《资治通鉴》和《新唐书》成书时代更为接近，语言风格、用词习惯等更为相似，也可能是因为《新唐书》记录的唐朝历史在《资治通鉴》中也有相应的记录.

进一步将在混合数据语料上训练的 Seq2Seq 模型和 Transformer 模型应用于史记、徐霞客游记和新唐书测试语料上，实验结果如表 4 所示(S2S 为 Seq2Seq 缩写，TF 为 Transformer 缩写) .总体上，2 个混合语料的结果也有差别，这表明混合语料的数据规模还不能充分支持模型的训练，因此未来还需要考虑增加训练语料的数量.基于分词的 Seq2Seq 模型的效果远差于基于单字的 Seq2Seq 模型的效果，而在 Transformer 模型上，分词的模型比单字的模型效果更好.其可能的原因是: 对于 Seq2Seq 模型，若分词效果不好则反而影响翻译的效果. 而 Transformer 多层的 Attention 结构可以学习到词与词之间的关系，这在一定程度上可以缓解分词的影响.此外，分词结果也可以看作是某种外部知识，Transformer 的多层架构也可以更好地建模这种知识.

表 4　基于混合语料的模型在不同测试集上的结果

语料		测试集		
		史记	徐霞客游记	新唐书
混合语料 1	S2S	20.87(J) 32.12(S)	13.44(J) 22.95(S)	15.46(J) 31.41(S)
	TF	35.71(J) 31.75(S)	22.12(J) 19.29(S)	25.88(J) 21.66(S)
混合语料 2	S2S	20.87(J) 30.07(S)	13.03(J) 21.57(S)	15.32(J) 32.19(S)
	TF	35.65(J) 32.77(S)	22.28(J) 19.80(S)	25.85(J) 21.39(S)

以往的研究认为影响文言文-现代文翻译质量的主要原因是语言变迁造成的语言鸿沟，但是实验结果显示语料规模的影响更大.由表 2 实验结果可知: 在混合语料上训练的模型明显优于在单个来源语料上训练的模型，即使是在徐霞客这样的不同文体的测试语料上，在混合语料上训练的模型也能取得了不错的结果，以 Seq2Seq 模型结果为例(见表 5) .

表 5　一些翻译结果的实例

测试语料	类型	文本
史记	原文	乌孙发二千骑往，持两端，不肯前.
	参考译文	乌孙出动二千骑兵前往大宛，但却采取骑墙态度，观望不前.
	资治通鉴模型	乌孙可汗派遣两名骑兵前往，双方观望，不肯前进.
资治通鉴	原文	逖妻，柳之姊也，固谏不从.
	参考译文	祖逖的妻子是许柳的姐姐，一再劝谏，祖约不听.
	资治通鉴模型	何低的妻子，是何低的姐姐，何低不服从.
新唐书	原文	乃亲谒宗庙，赠父玄贞上洛郡王.
	参考译文	于是韦氏亲自拜谒祖庙，追赠亡父韦玄贞上洛郡王封号.
	史记模型	就住在宗庙里的宗庙里，何低父亲何低上的何低.
	资治通鉴模型	于是亲自去见宗庙，并追赠父亲刘洛为洛郡王.
	混合语料模型	于是亲自拜谒宗庙，追赠父亲玄贞上洛郡王.
徐霞客	原文	有长节枝弱不繁者，潇洒而颇细;
	参考译文	有种竹子节长枝柔不繁杂，样子潇洒而且很细;
	史记模型	有一个长子，弱小的国家弱弱的人，确实被欺骗;
	资治通鉴模型	有一位长者的节度使的官员，有长时间不能繁琐繁衍;
	混合语料模型	有长节的枝节不繁杂的，潇洒却很细;

尽管更大规模的语料在各类测试语料上整体表现良好，但也并不足以说明语言变迁的影响不存在.表 5 列出了一些翻译结果实例，在史记语料上训练的模型用于新唐书语料的翻译，译文出现了严重的语义不全，而在资治通鉴语料上训练的模型虽然错翻了人名，但翻译结果也明显更好.这在一定程度

上支持了语言变迁会影响翻译模型的表现,只是在目前同时代平行语料严重匮乏的情况下,这个问题难以评估.在混合语料上训练的模型在徐霞客语料上的翻译结果较其他几个模型明显更优,除上述的数据规模的影响之外,其可能原因也在于混合语料包含了《资治通鉴》这类史料之外的其他文体的古籍,对于游记的翻译有所帮助,这也表明在文言文-现代文的翻译任务中还要考虑不同文体的翻译.

4 总结

本文研究了文言文-现代文的机器翻译问题,考察了 Seq2Seq 和 Transformer 这 2 个模型在文言文-现代文翻译上的效果.在这 2 个混合语料上的实验结果表明:目前训练语料的规模还不足以支持训练性能稳定的翻译模型,未来需要构建更大规模的高质量的文言文-现代文平行语料.国内学术界近年来的古籍信息化提供了大量的无标签的古代文献数据,也需要研究如何基于这些数据进行预训练.此外,文言文的句子短小精悍,存在大量的缩写和简写,也可以考虑引入篇章级别的翻译,以更好地学习上下文,帮助提高翻译质量.

5 参考文献

[1] FORCADA M L, GINESTÍ-ROSELL M, NORDFALK J, et al. Apertium: a free/open-source platform for rule-based machine translation [J]. Machine Translation, 2011, 25(2): 127-144.

[2] BROWN P F, DELLA PIETRA V J, DELLA PIETRA S A, et al. The mathematics of statistical machine translation: parameter estimation [J]. Computational Linguistics, 1993, 19(2): 263-311.

[3] MANNING C D. Human language understanding & reasoning [J]. Daedalus, 2022, 151(2): 127-138.

[4] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization [EB/OL]. [2022-01-19]. http://de.arxiv.org/pdf/1409.2329.

[5] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models [EB/OL]. [2022-01-11]. https://www.researchgate.net/publication/289758666_Recurrent_continuous_translation_models.

[6] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches [EB/OL]. [2022-01-17]. http://de.arxiv.org/pdf/1409.1259.

[7] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2022-02-10]. https://arxiv.org/pdf/1406.1078.pdf.

[8] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2022-02-16]. https://arxiv.org/pdf/1409.0473.pdf.

[9] WU Yonghui, SCHUSTER M, CHEN Zhifeng, et al. Google's neural machine translation system: Bridging the gap between human and machine translation [EB/OL]. [2022-02-18]. https://arxiv.org/pdf/1609.08144.pdf.

[10] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning [EB/OL]. [2022-02-22]. https://arxiv.org/abs/1705.03122.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2022-01-22]. https://doi.org/10.48550/arxiv.1706.03762.

[12] FENG Yang, XIE Wanying, GU Shuhao, et al. Modeling fluency and faithfulness for diverse neural machine translation [EB/OL]. [2022-03-10]. https://arxiv.org/abs/1912.00178.

[13] FENG Yang, GU Shuhao, GUO Dengji, et al. Guiding teacher forcing with seer forcing for neural machine translation [EB/OL]. [2022-02-13]. https://arxiv.org/abs/2106.06751v1.

[14] CHEN Kehai, WANG Rui, UTIYAMA M, et al. Content word aware neural machine translation [EB/OL]. [2022-03-11]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 358.

[15] QIU Xipeng, PEI Hengzhi, YAN Hang, et al. A concise model for multi-criteria Chinese word segmentation with transformer encoder [EB/OL]. [2022-03-19]. https://aclanthology.org/2020.findings-emnlp.260/.

[16] BAUGH A, CABLE T. A history of the English language [M]. London: Routledge, 1993.

[17] JHAMTANI H, GANGAL V, HOVY E, et al. Shakespearizing modern language using copy-enriched sequence-to-sequence models [EB/OL]. [2022-04-11]. https://arxiv.org/pdf/

1707.01161.pdf.

[18] CARLSON K, RIDDELL A, ROCKMORE D. Evaluating prose style transfer with the Bible [J]. Royal Society open science, 2018, 5(10): 171920.

[19] YANG Zhichao, CAI Pengshan, FENG Yansong, et al. Generating classical Chinese poems from vernacular Chinese [EB/OL]. [2022-04-15]. https://arxiv.org/abs/1909.00279.

[20] CHANG E, SHIUE Y T, YEH H S, et al. Time-aware ancient Chinese text translation and inference [EB/OL]. [2022-02-17]. https://arxiv.org/abs/2107.03179v1.

[21] LIU Dayiheng, YANG Kexin, QU Qian, et al. Ancient-modern chinese translation with a new large training dataset [J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2019, 19(1): 1-13.

[22] ZHANG Zhiyuan, LI Wei, SU Qi. Automatic translating between ancient Chinese and contemporary Chinese with limited aligned corpora [EB/OL]. [2022-03-19]. https://arxiv.org/abs/1803.01557.

[23] DANKERS V, LUCAS C G, TITOV I. Can transformer be too compositional? analysing idiom processing in neural machine translation [EB/OL]. [2022-02-19]. https://arxiv.org/abs/2205.15301.

[24] YANG Zinong, CHEN Kejia, CHEN Jingqiang. Guwen-UNILM: machine translation between ancient and modern Chinese based on pre-trained models [EB/OL]. [2022-04-10]. https://link.springer.com/chapter/10.1007/978-3-030-88480-2_10.

[25] SUTSKEVER I, VINYALS O, LE Quoc V. Sequence to sequence learning with neural networks [EB/OL]. [2022-02-18]. https://www.docin.com/p-2163732294.html.

[26] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation [EB/OL]. [2022-02-15]. https://aclanthology.org/P02-1040.pdf.

The Study on the Neural Machine Translation of Ancient Chinese-Modern Chinese

HE Silan, ZUO Jiali *, ZHU Hongkun, WANG Mingwen

(School of Computer and Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: There are a large number of Chinese ancient documents, which are the treasures of Chinese civilization. However, it is extremely difficult for modern people to understand these documents, and it is also impossible to translate them manually. Therefore, the Neural Machine Translation of ancient Chinese-modern Chinese is studied. By applying the Seq2Seq model and Transformer model, the impact of the size of training corpus on the translation performance of Ancient Chinese-Modern Chinese is investigated. It is also found that based on the existing training corpus of this scale, word segmentation will greatly affect the performance of Seq2Seq model. In addition, if the style of training corpus and test corpus is different, the performance of the model will also be affected.

Key words: ancient Chinese-modern Chinese neural machine translation; Seq2Seq model; transformer

(责任编辑: 冉小晓)