

doi:10.3969/j.issn.1673-9833.2020.03.002

低资源场景下印尼语 – 汉语机器翻译 模型对比分析研究

李霞^{1,2}, 马骏腾², 肖莉娴³, 王连喜^{1,2}

(1. 广东外语外贸大学 广州市非通用语种智能处理重点实验室, 广东 广州 510006;

2. 广东外语外贸大学 信息科学与技术学院, 广东 广州 510006;

3. 广东外语外贸大学 东方语言文化学院, 广东 广州 510006)

摘要: 低资源场景下神经机器翻译模型的翻译效果通常不如大规模训练数据下的模型翻译效果好。针对这个问题, 选取 RNN-based 和 Transformer-based 两个主流神经机器翻译模型进行研究, 探究在印尼语到汉语翻译在低资源场景下的机器翻译效果, 并通过实验分析和案例分析, 给出两种模型在印尼语到汉语低资源场景下翻译的适应性分析。

关键词: 神经机器翻译; 序列到序列翻译模型; 低资源语言; 印尼语 – 汉语; Transformer

中图分类号: TP391

文献标志码: A

文章编号: 1673-9833(2020)03-0010-07

引文格式: 李霞, 马骏腾, 肖莉娴, 等. 低资源场景下印尼语 – 汉语机器翻译模型对比分析研究 [J]. 湖南工业大学学报, 2020, 34(3): 10-16.

A Comparative Study on Indonesian-Chinese Machine Translation Models in Low-Resource Scenarios

LI Xia^{1,2}, MA Junteng², XIAO Lixian³, WANG Lianxi^{1,2}

(1. Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies,

Guangzhou 510006, China; 2. School of Information Science and Technology, Guangdong University of

Foreign Studies, Guangzhou 510006, China; 3. Faculty of Asian Languages and Cultures,

Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: The results of neural machine translation models in low-resource scenario is generally not as good as that of models under large-scale training data. For this issue, RNN-based and Transformer-based mainstream neural machine translation models are selected for a study on the effect of neural machine translation models in the Indonesian to Chinese low-resource scenario. Several experiments are carried out and through experimental analysis and case studies, as well as an adaptability analysis given to the two models.

Keywords: neural machine translation; sequence to sequence translation model; low resource language; Indonesian to Chinese; Transformer

收稿日期: 2020-04-02

基金项目: 国家自然科学基金资助面上项目 (61976062), 国家社会科学基金资助项目 (17CTQ045), 广东省普通高校“人工智能”重点领域专项基金资助项目 (2019KZDZX1016), 广州市科技计划基金资助项目 (201904010303)

作者简介: 李霞 (1976-), 女, 江西景德镇人, 广东外语外贸大学教授, 博士, 硕士生导师, 主要从事自然语言处理方面的教学与研究, E-mail: xiali@mail.gdufs.edu.cn

通信作者: 肖莉娴 (1977-), 女, 广东江门人, 广东外语外贸大学讲师, 主要研究方向为印度尼西亚语言和文化, E-mail: 173829137@qq.com

1 研究背景

机器翻译(machine translation, MT)是将文本从一种语言自动转换为另一种语言的技术,是自然语言处理的重要研究领域之一。近几十年来,机器翻译技术不断发展,从传统的统计机器翻译方法^[1-3]到基于端到端的神经机器翻译方法^[4-7],翻译模型和翻译效果都取得了显著提升。

经典的神经机器翻译模型主要基于编码器-解码器结构,是一种基于递归神经网络的变体,常用的编码器和解码器网络单元为LSTM(long short term memory, LSTM)^[8]或GRU(gated recurrent unit, GRU)^[9],该类模型的优点是能有效捕捉句子单词的序列信息,较好地处理长序列时梯度消失问题,但它存在不能并行对句子进行编码的缺点。为此, J. Gehring等^[10]提出一种完全基于卷积神经网络^[11]的编码器-解码器结构,使得计算可以并行化,并通过叠加多层来表达较大窗口内的上下文信息,取得了较好的翻译性能。2017年A. Vaswani等^[7]提出了使用Transformer结构来构建神经机器翻译模型,Transformer抛弃了传统的RNN(recurrent neural network)结构,提出和使用自注意力机制,使得它更关注句子本身,从而能够抽取更多的信息,进而取得了很好的性能,后续的工作中也陆续出现基于Transformer结构的不同改进工作^[12-13]。

近年来,机器翻译工作也聚焦于融合视觉、听觉等信息的多模态机器翻译^[14-15],通过结合图片中的语义信息来辅助文本的语义消歧,进而提升机器翻译的准确率。与此同时,面向资源稀缺的低资源语言机器翻译研究近年来也成为研究热点,如面向尼泊尔语与僧伽罗语^[16]、希伯来语^[17]、印尼语、菲律宾语等资源较少的机器翻译研究。在面向低资源语言的机器翻译场景中,由于训练语料较少,且源端和目标端的语言形态、句子结构差异较大,低资源语言机器翻译也是近年来机器翻译研究的难点之一。

为了解决低资源训练语料少、模型训练难等问题,研究人员提出迁移学习和元学习等技术来实现面向低资源语言的机器翻译模型和方法。如B. Zoph等^[18]提出使用迁移学习的思想用于低资源语言机器翻译研究中,其思想是通过在语料丰富的语言对上训练模型后,再使用低资源语言的数据做微调,从而获得在低资源语言上的翻译模型。Gu J. T.等^[19]使用元学习的方法来解决低资源语言翻译困难的问题。R. Sennrich等^[20]则提出使用易获得的目标语言的大量单语语料,通过现有的翻译工具或者已训练好的机器

翻译模型对其进行反向翻译,从而构造伪平行句对。近几年由于计算能力的提升,研究人员在面向低资源语言的机器翻译研究中从迁移技术的使用^[21]与如何构造更多更好的平行数据^[22-23]角度不断开展低资源语言的机器翻译研究。

虽然神经机器翻译模型相比传统基于统计的翻译模型取得了更好的效果,但是它需要足够规模的训练语料才能更好地调参并获得较好的翻译效果。同时,目标语言与源语言形态的差异、语法结构的差异也会影响模型的性能。例如,属于同一个语系的英语和法语,与属于不同语系的英语和汉语的机器翻译相比,即使在训练数据完全相同的情况下其结果也会有所差异。

本文主要关注主流神经机器翻译模型在低资源语言或在训练数据量少的场景下的表现情况。课题组以印尼语-汉语为语言翻译对象,探究面向资源较为稀缺的低资源语言(印尼语)和在训练数据量较少情况下(低资源场景下),当前主流神经机器翻译模型的翻译表现。课题组尝试从已有的“英语-汉语”数据集构造相同规模的“印尼语-汉语”伪平行数据,使它们具有相同目标端的中文句子,并进行训练和测试,将得到的结果进行对比和分析,探究在相同的低资源场景下,从英语到汉语和从印尼语到汉语的翻译效果。本文工作的主要贡献如下:

- 1) 从已有的“英语-汉语”数据集拓展并构建“印尼语-汉语”伪平行数据集用于模型训练;
- 2) 探索和分析了基于LSTM和基于Transformer两种端到端翻译模型在英语-汉语和印尼语-汉语两个不同语言对上的翻译效果分析;
- 3) 给出了融合注意力机制的基于LSTM和基于Transformer端到端神经机器翻译模型在低资源场景下的适应性分析。

2 神经机器翻译模型 NMT

经典神经机器翻译模型(neural machine translation, NMT)采用基于编码和解码的序列到序列翻译模型框架,输入端为源语言单词序列 $X=(x_1, x_2, \dots, x_M)$,输出端为目标语言单词序列 $Y=(y_1, y_2, \dots, y_N)$, NMT模型希望学习得到 X 翻译为 Y 的概率 $P=(Y|X)$ 最大的模型,从而学习得到训练集数据的条件概率分布。本文使用的NMT模型分别为融入注意力机制的基于LSTM机器翻译模型和基于Transformer的神经机器翻译模型,分别以Att-LSTM-based NMT模型和Transformer-based NMT模型来命名表示该两个模型。

2.1 融入注意力机制的 LSTM-based 端到端翻译模型

2.1.1 编码端

在 LSTM-based 端到端神经机器翻译模型中, 编码器使用双向 LSTM 对源语言句子进行编码, 其中前向、反向 LSTM 的输出序列分别为 $(\bar{h}_1, \bar{h}_2, \dots, \bar{h}_N)$ 和 $(\bar{h}_1, \bar{h}_2, \dots, \bar{h}_N)$, 其计算方法如式(1)~(2)所示:

$$\bar{h}_i = f_{\text{enc}}(\mathbf{W}_x[x_i], \bar{h}_{i-1}), \quad (1)$$

$$\bar{h}_i = f_{\text{enc}}(\mathbf{W}_x[x_i], \bar{h}_{i+1}). \quad (2)$$

式(1)(2)中: \mathbf{W}_x 为源语言单词 x_i 转换为词向量的词向量查找矩阵; f_{enc} 为 LSTM 计算过程的函数。

最终编码器在每个时间步上的输出是每个单词前向和后向隐含向量的拼接, 即 $h_i = [\bar{h}_i; \bar{h}_i]$, 编码端源语言输入序列经过编码得到的输出为 $h = (h_1, h_2, \dots, h_N)$ 。

2.1.2 解码端

注意力机制旨在更好地得到目标单词与源语言句子中单词的对齐关系。参考 M. T. Luong 等^[6]的工作, 本文在模型的解码端使用 LSTM 单元, 并融合注意力机制, 计算方法如下:

首先, 计算解码器 t 时刻的隐含状态 s_t 和编码器的每个输出隐含状态 h_i 的对齐信息 $e_{t,i}$, 如式(3)所示, 其中 s_{t-1} 为解码器的上一个隐含状态, \mathbf{W}_a 为参数矩阵。

然后, 通过 Softmax 函数计算得到解码器 t 时刻隐含状态所对应编码器不同隐含状态 h_i 的权重 $a_{t,i}$, 如式(4)所示。

最后, 进行加权求和得到解码端 t 时刻隐含状态在编码器端的上下文向量 c_t , 如式(5)所示。

最后, 如式(6)所示, 解码器 t 时刻的隐含状态 s_t 的结果由 3 部分计算得到: 解码器的上一个隐含状态 s_{t-1} ; 解码器 $t-1$ 时刻预测输出的目标单词 \tilde{y}_{t-1} (模型训练时 \tilde{y}_{t-1} 取与 x_{t-1} 对应的单词 y_{t-1}); t 时刻解码器隐含状态的上下文向量 c_t , 其中 \mathbf{W}_y 为目标语言单词 y_{t-1} 转换为目标语言单词的词向量查找矩阵, f_{dec} 则表示 LSTM 计算过程的函数。

计算式如(3)~(6)所示:

$$e_{t,i} = h_i^T \mathbf{W}_a s_{t-1}, \quad (3)$$

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^N \exp(e_{t,j})}, \quad (4)$$

$$c_t = \sum_{i=1}^N a_{t,i} h_i, \quad (5)$$

$$s_t = f_{\text{dec}}\left(\mathbf{W}_y \left[\tilde{y}_{t-1}, s_{t-1}, c_t \right]\right). \quad (6)$$

2.2 基于 Transformer 端到端翻译模型

2.2.1 编码端

Transformer 的内部结构包含一个自注意力层 (self-attention layer) 和前馈神经网络层 (feed-forward network layer, FFN), 编码器和解码器都可以通过堆叠层数来加深网络的深度, 提高模型的效果。每个编码器层只包含自注意力层和 FFN 层, 在自注意力层中, 每个输入的向量会被 3 个不同的权重矩阵 $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ 转换成 3 个不同的向量, 其中 \mathbf{Q} 为解码器的上一层输出, \mathbf{K} 和 \mathbf{V} 为编码器的输出, 分别为 $\mathbf{Q}(\text{Query})$ 、 $\mathbf{K}(\text{Key})$ 和 $\mathbf{V}(\text{Value})$ 。自注意力机制使用缩放点积 (scaled dot-product) 作为相似度计算函数, 如式(7)所示,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \quad (7)$$

式中 d 为 \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 的维度。

由于 Transformer 与 RNN 不同, 它并不能捕捉句子的序列信息, 因此在输入的词向量中加入了位置编码给每个单词提供位置信息。为了使自注意力层有更好的表现, A. Vaswani 等^[7]还将输入的 \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 投影到多个子空间中, 即多头机制 (Multi-Head), 它将每个“头”得到的输出拼接后, 输入一个全连接层, 如式(8)所示,

$$\begin{cases} \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \\ \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)\mathbf{W}^O, \\ \text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \end{cases} \quad (8)$$

式中: $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d}$; $\mathbf{W}^O \in \mathbb{R}^{hd \times d_{\text{model}}}$, 为将每个头的结果拼接后转化为模型维度 d_{model} 的全连接权重矩阵, 其中, d_{model} 为模型维度, h 为头的数量。

2.2.2 解码端

解码器层在自注意力层和 FFN 层之间多了一个编码器-解码器注意力层 (encoder-decoder layer), 接收来自编码器的最终输入。它采用了和 2.2.1 节描述的编码端多头注意力机制一致的计算方法。

编码器和解码器中的每个子层的输出都使用了残差连接, 防止网络加深时出现退化的梯度消散问题。并使用了层正则化^[23], 使得正则化后的数据尽可能地保持原始表达能力, 更适应模型的训练。

3 实验

3.1 实验数据

本文所用 IWSLT 2015 (<https://sites.google.com/>)

site/iwslt-evaluation2015/mt-track) 机器翻译评测任务所提供的评测数据集, 该数据收集的领域是 TED (<https://www.ted.com>) 中的演讲内容。为了对比神经机器翻译模型在不同语种上的翻译效果, 保证数据的一致性, 课题组选取 IWSLT 2015 英语-汉语数据集作为基础并进行扩展, 得到印尼语-汉语伪平行数据集。扩展的方法是使用 Google Translation API (<https://translate.google.cn/>, 翻译时间为 2020-01) 将训练集中的英语句子自动翻译为相应的印尼语, 选择用英语翻译的原因是考虑到印尼语与英语都是使用 26 个英文字母所构成的单词, 语言形态相似, 得到的数据可能质量更好。实验所采用的校验集和测试集则由印尼语专业教师进行人工翻译和标注, 校验集用于翻译模型的调参, 测试集用于模型的结果评测。详细数据集信息如表 1 所示。

表 1 实验数据细节 (单位: 句对)

Table 1 Details of experimental data (unit: sentence pair)

翻译语言对	训练集	开发集	测试集
英语-中文	209 940	887	1 080
印尼语-中文	209 940	887	1 080

3.2 实验设置

3.2.1 数据预处理

在数据预处理阶段, 对于英语语言使用 Moses^[24] 统计机器翻译库中的预处理脚本对语言进行预处理操作, 其中包括分词、标点规范化、字母大小写处理 (truecase) 等。对于印尼语课题组则直接将所有字母小写并进行分词, 对于中文分词则采用的是 Jieba (<https://github.com/fxsjy/jieba>) 分词工具。实验中, 课题组控制训练集的句子长度在 1~50 个词之间, 并且将源语言及目标语言的词汇表大小设定为 50 000 个。在训练过程中使用校验集对模型进行调参和模型选择, 整个实验采用 BLEU^[25] 作为评价指标和结果汇总。实验所使用的硬件设备为 1 块 GeForce RTX 2080GPU 显卡。

3.2.2 基于 LSTM 的端到端模型实验设置

编码器端使用双向 LSTM 作为编码单元, 隐含状态向量的维度为 512 维, 源语言和目标语言的词向量维度为 512 维。解码器端使用单向 LSTM, 隐含状态向量的维度为 512 维。模型采用 Adam 优化器^[26], 初始学习率为 0.001。batch 类型为句子, batch size 设置为 64, dropout 概率设置为 0.3。在训练时每 2 000 个训练迭代在校验集上测试, 如果连续 4 次模型在校验集上的困惑度值 (perplexity) 没有下降, 则执行一次学习率衰减, 衰减为原来的二分之一。实验中, 如果这种情况出现 8 次, 则使用早期停止策略 (early

stop) 结束训练, beam search 的大小设置为 5。

3.2.3 基于 Transformer 的端到端模型实验设置

文中 Transformer 模型使用 OpenNMT-py^[27] 开源框架进行实验, 其中编码器与解码器均使用 6 层 8 个头, 其中隐含状态向量和词向量维度均为 512 维。词向量加入了位置编码, FFN 维度设置为 2 048 维。优化器选用 Adam, 初始学习率为 0.1。实验使用 Noam 学习率下降方案^[4], 其中 $\beta_1=0.9$, $\beta_2=0.998$, warm up steps 为 16 000。batch 的类型为单词, batch size 设置为 4 098, 训练 200 000 个 training iteration, 选择在校验集上表现最好的模型。dropout 概率设置为 0.3, beam search 大小设置为 5。

3.3 实验结果与分析

使用所给出的实验设置和评价指标, 融入注意力机制 LSTM-based 的神经机器翻译模型 (Att-LSTM-based NMT) 和基于 Transformer 的神经机器翻译模型 (Transformer-based NMT) 的详细实验结果如表 2 所示。

表 2 不同语言对的实验结果

Table 2 Experimental results of different models with different language pairs

翻译模型	源语言	目标语言	BLEU4($\mu \pm \sigma$)
Att-LSTM-based NMT	英语	汉语	15.35 \pm 0.37
	印尼语	汉语	13.92 \pm 0.73
Transformer-based NMT	英语	汉语	14.41 \pm 0.14
	印尼语	汉语	13.59 \pm 0.15

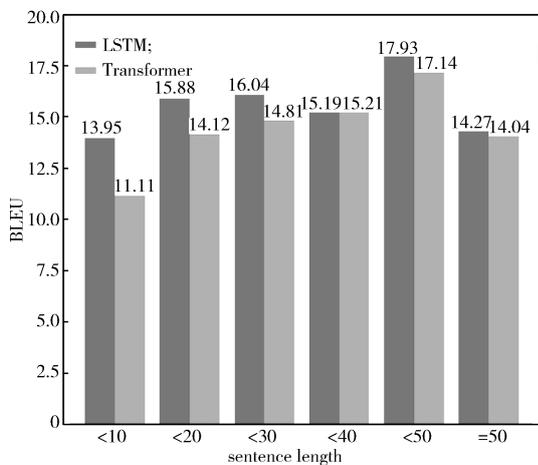
由表 2 可知, 在英语-汉语和印尼语-汉语两个语言翻译对中, Att-LSTM-based NMT 模型均比 Transformer-based NMT 模型略好, 其中 Att-LSTM-based NMT 模型在英语-汉语语言对上的 BLEU 值为 15.35, 在印尼语-汉语语言对上的 BLEU 值为 13.92, 相比而言, 在印尼语-汉语语言对上的结果要比在英语-汉语语言对上的翻译结果低 1.43 个 BLEU 值, 这表明模型在低资源语言以及针对印尼语-汉语这种语言形态差异较大的语言对上的翻译结果相对较低。

实验结果还表明, Transformer-based NMT 模型在两个相同平行语言对上翻译的结果, 均略低于 Att-LSTM-based NMT 模型。课题组分析, 可能的原因是在低资源的情景中 (数据集的数量较少), Transformer-based NMT 模型在训练数据不够的情况下, 翻译效果可能不如 RNN 结构的机器翻译, 因为 Transformer 的网络深度更深, 在这种量级数据上容易过拟合。同时, Transformer-based NMT 在参数相同的情况下, 运行出来的结果相对更为稳定, Transformer-based NMT 模型在英语-汉语和印尼语-

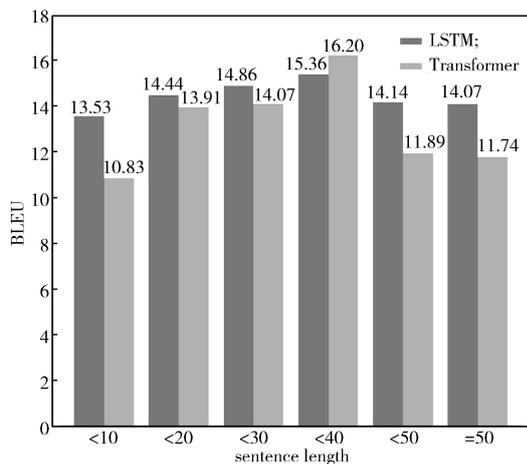
汉语的翻译 BLEU 值标准方差分别为 0.14 和 0.15, 相比于 Att-LSTM-based NMT 模型的 0.37 和 0.73 更为稳定。

3.4 模型的适应性分析

本研究针对印尼语-汉语低资源语言对, 以及翻译训练数据较小的情况下, Att-LSTM-based NMT 模型和 Transformer-based NMT 模型的翻译适应性进行分析。课题组分别分析了两个模型在不同语言对上在不同句子长度时的翻译效果, 结果如图 1 所示。其中图 1a 为句子长度不同时两个模型在英语-汉语对上的翻译结果对比分析, 图 1b 为句子长度不同时两个模型在印尼语-汉语对上的翻译结果对比分析。



a) 英语-汉语语言对翻译结果



b) 印尼语-汉语语言对翻译结果

图 1 模型在不同长度句子上的 BLEU 值比较分析

Fig. 1 A comparative analysis of two models with BLEU metrics with different length of sentences

如图 1 所示, 在大多情况下, Att-LSTM-based NMT 模型结果相对更好, 且两个模型在长度为 30~50 时, 翻译的结果最好。但在句子长度为 30~40 时, Transformer-based NMT 模型的翻译效果都要稍好于 Att-LSTM-based NMT 模型。同时课题组发现,

当句子长度越长, Transformer-based NMT 模型的性能下降最大, 这可能是因为 Transformer 尽管通过加入位置编码来解决单词位置信息的问题, 但是其仍然在建立更长距离依赖方面效果较差, 这与位置编码的方式选择有一定的关系。

同时, 课题组还对 Att-LSTM-based NMT 模型和 Transformer-based NMT 模型的训练时间进行了对比分析, 两个模型以 3.2.2 节和 3.2.3 节参数设置情况下的训练时间对比如表 3 所示。

表 3 模型的训练时间对比

Table 3 Comparison of training time of the models

模 型	时间/h
Att-LSTM-based NMT	16.5
Transformer-based NMT	26.8

由表 3 可得, Transformer-based NMT 模型所需要的平均训练时长要比 Att-LSTM-based NMT 模型更长。以实验结果来看, 在 3.2.2 和 3.2.3 节使用的参数和相同硬件情况下, Transformer-based NMT 模型训练所用时间为 26.8 h, 而 Att-LSTM-based NMT 模型训练时间为 16.5 h, 低于 Transformer-based NMT 模型训练时间。这也验证了 Transformer-based NMT 模型更为复杂, 需要更多的大规模训练数据进行调参才能达到较好的效果。

3.5 案例分析

本研究还对两个模型的翻译结果进行案例分析, 比较两种 NMT 翻译的效果。课题组选取了两个印尼语-汉语的翻译结果进行对比分析, 如表 4 和表 5 所示。所选取的两个印尼语句子长度都在 30 个单词以上, 选择了两种模型表现最好的句子长度的情景下结果。

表 4 案例分析 (1)

Table 4 Case study (1)

源印尼语	Saya belajar bahwa itu ide jenius untuk menggunakan penjepit barbekyu untuk mengambil barang-barang yang anda jatuhkan. Saya belajar trik yang bagus di mana anda dapat mengisi baterai ponsel anda dari baterai kursi anda.
参考译文	我学到的是, 用烧烤棍把你掉的东西捡起来是个相当不错的主意。我觉得用电动轮椅的电池给手机充电也是个绝妙的点子。
Att-LSTM-based NMT	我学会了, 这是一个天才, 用烧烤烧烤的东西来提取你的东西。我学到了一个不错的戏法, 你可以用你的电池电池充电。
Transformer-based NMT	我了解到这是个天才的想法, 用一个风车来把你的东西拿出来。我学会了一个很好的戏法。我知道你可以用你的电池充电。

表5 案例分析(2)
Table 5 Case study(2)

源印尼语	Saya sudah lupa berapa kali saya didekati oleh orang asing yang ingin memberitahukan kepada saya bahwa mereka pikir saya berani atau inspiratif. Dan ini terjadi jauh sebelum saya memiliki pekerjaan sebagai profil publik apa pun.
参考译文	我已经记不清有几次,有陌生人向我走过来,想要告诉我他们认为我很勇敢,很励志,这还是在我没有成为公众人物之前的事。
Att-LSTM-based NMT	我已经忘记了我和陌生人的接触,他们想告诉我,他们认为我是有勇气或鼓舞人心的,这发生在我工作之前,在我工作之前。
Transformer-based NMT	我忘记了有多少次被陌生人告诉我,他们觉得我很勇敢,或者鼓舞人心,而在我拥有任何公共资料之前,我才知道。

以表4所示的案例为例,Att-LSTM-based NMT模型结果中,在句子的后半部分出现了单词重复的情况,如“你可以用你的电池电池充电”,而Transformer-based NMT模型的翻译结果更加符合源句子真实翻译结果。而如表5所示,在翻译的结果中,同样出现了“这发生在我工作之前,在我工作之前”这种重复翻译的情况,并且相比而言,Transformer-based NMT模型的翻译效果更好。这表明,即使在小数据集的训练场景下,Transformer-based NMT模型依然可以取得较好的翻译稳定性。

4 结语

课题组探究了在低资源语言和数据量较小的场景下,基于Transformer的神经机器翻译模型和基于LSTM神经机器翻译模型的翻译结果和适应性对比分析。实验中课题组发现,在低资源的情景下,融入注意力机制的基于LSTM的神经机器翻译模型无论是翻译效果或者训练时间上,都略好于Transformer,这可能是由于基于Transformer的神经机器翻译模型更深,参数量大,需要在更大规模的数据中才能凸显其优势,并且Transformer对超参数的设定更加敏感,调参所需代价较大,因此在低资源场景下使用RNN结构模型的性能会比更高。同时在句子长度中等的情况下,基于Transformer的NMT翻译效果会略胜一筹。另外,总体而言,尽管在小数据集量情况下,基于Transformer神经机器翻译模型的稳定性更好。

参考文献:

[1] BROWN P F, COCKE J, PIETRA S A D, et al.

- A Statistical Approach to Machine Translation[J]. Computational Linguistics, 1990, 16(2): 79-85.
- [2] BROWN P F, PIETRA S A D, PIETRA V J D, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.
- [3] CHIANG D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2): 201-228.
- [4] SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C]//Advances in Neural Information Processing Systems. Montréal: NIPS, 2014: 3104-3112.
- [5] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[EB/OL]. [2020-03-24]. <https://arxiv.org/abs/1409.0473>.
- [6] LUONG M T, PHAM H, MANNING C D. Effective Approaches to Attention-Based Neural Machine Translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 1412-1421.
- [7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[C]//Advances in Neural Information Processing Systems 30. Long Beach: NIPS, 2017: 5998-6008.
- [8] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [9] CHUNG J, GULCEHRE C, CHO K, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[EB/OL]. [2020-03-24]. <https://arxiv.org/pdf/1412.3555.pdf>.
- [10] GEHRING J, AULI M, GRANGIER D, et al. Convolutional Sequence to Sequence Learning[C]// Proceeding ICML'17 Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017: 1243-1252.
- [11] LECUN Y, HAFNER P, BOTTOU L, et al. Object Recognition with Gradient-Based Learning[M]. Berlin: Springer Berlin Heidelberg, 1999: 319-345.
- [12] MA J T, QIN S H, SU L, et al. Fusion of Image-Text Attention for Transformer-Based Multimodal Machine Translation[C]// 2019 International Conference on Asian Language Processing (IALP). Shanghai: IEEE, 2019: 199-204.
- [13] ZHANG B, TITOV I, SENNRICH R. Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention[EB/OL]. [2020-03-24]. <https://arxiv.org/pdf/1908.11365.pdf>.
- [14] CALIXTO I, LIU Q. Incorporating Global Visual

- Features into Attention-Based Neural Machine Translation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 992-1003.
- [15] MA J T, QIN S H, CHEN M P, et al. Independent Fusion of Words and Image for Multimodal Machine Translation[C]// Communications in Computer and Information Science. Singapore: Springer Singapore, 2019: 35-46.
- [16] GUZMÁN F, CHEN P J, OTT M, et al. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). HongKong: Association for Computational Linguistics, 2019: 6098-6111.
- [17] GAO F, ZHU J H, WU L J, et al. Soft Contextual Data Augmentation for Neural Machine Translation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 5539-5544.
- [18] ZOPH B, YURET D, MAY J, et al. Transfer Learning for Low-Resource Neural Machine Translation[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 1568-1575.
- [19] GU J T, WANG Y, CHEN Y, et al. Meta-Learning for Low-Resource Neural Machine Translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 3622-3631.
- [20] SENNRICH R, HADDOW B, BIRCH A. Improving Neural Machine Translation Models with Monolingual Data[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 86-96.
- [21] KIM Y, GAO Y B, NEY H. Effective Cross-Lingual Transfer of Neural Machine Translation Models Without Shared Vocabularies[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence: Association for Computational Linguistics, 2019: 1246-1257.
- [22] XIA M Z, KONG X, ANASTASOPOULOS A, et al. Generalized Data Augmentation for Low-Resource Translation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence: Association for Computational Linguistics, 2019: 5786-5796.
- [23] BA J L, KIROS J R, HINTON G E. Layer Normalization. [EB/OL]. [2020-03-24]. <https://arxiv.org/abs/1607.06450>.pdf.
- [24] 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. Prague: Association for Computational Linguistics, 2007: 177-180.
- [25] PAPIENI K, ROUKOS S, WARD T, et al. Bleu: a Method for Automatic Evaluation of Machine Translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311-318.
- [26] KINGMA D P, BA J L. Adam: A Method for Stochastic Optimization[EB/OL]. [2020-03-24]. <https://arxiv.org/abs/1412.6980>.pdf.
- [27] KLEIN G, KIM Y, DENG Y et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation[C]// Proceedings of ACL 2017, System Demonstrations. Vancouver: Association for Computational Linguistics, 2017: 67-72.

(责任编辑: 申 剑)