小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI: 10.20009/j.cnki.21-1106/TP.2020-1084
2022 年 7 月 第 7 期
Vol. 43 No. 7 2022

融入语言差异化特征的汉越神经机器翻译译文质量估计

邹 翔,朱俊国,高盛祥,余正涛,杨福岸

(昆明理工大学 信息工程与自动化学院,昆明 650500)
(昆明理工大学 云南省人工智能重点实验室,昆明 650500)
E-mail: 570188903@qq.com

摘 要: 译文质量估计是机器翻译领域中一个重要的子任务,该任务旨在不依靠参考译文的情况下对机器译文进行质量分析.当前,译文质量估计任务在汉英、英德机器翻译上有较好的表现,技术相对成熟.但是将模型应用到汉-越神经机器翻译中面临较多问题.尤其是译文质量估计模型在汉越平行数据中提取到的语言特征不能够充分地体现汉语与越南语之间的语言特点,加之汉语与越南语之间语序与句法结构也存在明显的差异.针对上述问题,本文采用统计对齐的方法对汉越之间结构差异进行建模,提取汉语与越南语之间的语言差异化特征,以提升汉越译文质量估计的效果.实验结果表明,融入语言差异化特征在汉-越和越-汉两个方向上较基线模型分别提升了 0.52 个百分点和 0.35 个百分点.

关 键 词: 质量估计; 汉越平行数据; 语言特点; 差异化特征; 汉-越神经机器翻译
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2022)07-1413-06

Translation Quality Estimation of Chinese-Vietnamese Neural Machine Translation Incorporating Linguistic Differentiation Features

ZOU Xiang, ZHU Jun-guo, GAO Sheng-xiang, YU Zheng-tao, YANG Fuan
(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)
(Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Quality estimation is an important sub-task in machine translation, which aims to analyze the quality of machine translations without references. At present quality estimation model is well performed in Chinese-English and English-German machine translation, and technology is relatively mature. However, there are still many problems in applying the quality estimation model to Chinese-Vietnamese neural machine translation. In particular, the linguistic features extracted from the Chinese-Vietnamese parallel data by the translation quality estimation model cannot reflect the linguistic characteristics enough between Chinese and Vietnamese. There are obvious differences in word order and syntactic structure between Chinese and Vietnamese. Focus this problem, this paper uses statistical alignment approach modeling structural differences between Chinese and Vietnamese, extracting differentiation features from Chinese and Vietnamese, in order to improve the performance of Chinese-Vietnamese quality estimation model. The experimental results show that, the integration of linguistic features has increased by 0.52% and 0.35% compared with the baseline model in Chinese-Vietnamese and Vietnamese-Chinese task.

Key words: quality estimates; Chinese-Vietnamese parallel data; linguistic characteristics; differentiation features; Chinese-Vietnamese neural machine translation

1 引 言

句子级译文质量估计(Quality Estimation, QE)旨在无需参考译文的情况下,以源语句和翻译系统输出的结果作为输入,对译文的质量进行估计.将可以表示源语言与机器译文的流畅度、忠实度和复杂性度的特征与机器学习方法相结合,以达到训练预测模型的目的.句子级别的译文质量估计不仅可以为终端用户提供一个度量译文可靠性的指标,而且可以减少翻译人员对机器译文进行人工后期编辑的时间[1].当前译文质量估计任务主要关注在一些资源丰富的语言对以及欧洲的资源稀缺型语言上,尚未针对汉越神经机器翻译(NMT)[2-4]开展译文质量估计的相关研究,但是译文质量估计方法对于提升汉越神经机器翻译有一定帮助,所以本文针对汉越译文质量估计展开相关研究.

在汉越神经机器翻译译文质量估计任务中,我们通过分析汉语与越南语之间存在的差异性,将其作为差异化特征融

收稿日期: 2020-12-28 收修改稿日期: 2021-02-19 基金项目: 国家自然科学基金项目(61732005, 61761026, 61672271, 61866020)资助; 国家重点研发计划项目(2019QY1802, 2019QY1801, 2019QY1800)资助; 云南省重大科技专项计划项目(202002AD080001)资助; 云南省人培项目(KKSY201903018)资助. 作者简介: 邹 翔,男,1995 年生,硕士研究生,CCF 会员,研究方向为自然语言处理、机器翻译; 朱俊国,男,1982 年生,博士研究生,讲师,CCF 会员,研究方向为自然语言处理、机器翻译; 高盛祥,女,1977 年生,博士,教授,CCF 会员,研究方向为机器学习、自然语言处理、机器翻译; 余正涛,男,1970 年生,博士,教授,CCF 会员,研究方向为自然语言处理、信息检索、机器翻译; 杨福岸,男,1994 年生,硕士研究生,研究方向为自然语言处理、机器翻译.

入到译文质量估计模型中,以缓解模型对特征抽取不够充分的问题. 另外,为了降低汉越平行数据稀疏问题对本任务带来的负面干扰,我们通过回译的方式对特征提取模型使用的训练集进行了一定规模的扩充,更严谨的验证语言差异化特征对于汉越神经机器翻译译文质量估计任务的影响.

2 相关工作

早期对于译文质量估计研究,将其视为有监督的回归或分类问题,QuEst 是其中最具代表性的译文质量估计框架,QuEst 通过对特征的抽取与选择对机器译文的质量进行估计. 主要抽取的特征有流畅度、忠实度、复杂度等角度提取的反映译文质量的特征;通过网格搜索对特征权重进行学习,再利用支持向量机[7]、逻辑回归[8]、条件随机场[9]等方法学习到特征与译文质量之间的映射关系.

近几年,由于深度学习在自然语言的相关任务中取得较大成功,越来越多的研究人员开始将循环神经网络[10,11]构建的语言模型应用于译文质量估计任务上. 对于译文质量估计而言,深度神经网络有强大的特征学习能力且模型对平行的双语数据有较好的感知能力,可以有效的学习到数据中的上下文信息. Shah[12]等和陈志明[13,14]等利用词语的分布式表达和循环神经网络语言模型等方法抽取特征提升译文质量估计模型的性能. Zhu[15]等提出通过学习双语句子的特征表示来建立机器翻译质量估计模型,让模型针对于翻译过程出现的正例与反例情况进行学习,在一定程度上缓解了训练语料不足的问题. Kim[16]等提出了一种"两阶段"的译文质量估计模型,其中第 1 部分的模型是在 Bahdanau 提出的 NMT 模型的基础上将解码器部分改为了双向的长短期记忆网络,这样可以充分利用目标词左右两侧的信息,该阶段的输入为源语句与机器译文,输出则是包含对应目标词位置翻译质量的序列(Quality Vector). 第 2 部分的模型为单向的长短期记忆网络,上一阶段的输出作为该阶段的输入,最后输出句子的质量分数. 李茂西[17]等提出将"预测器-估计器"中两个子网格组合成一个整体的端到端的联合神经网络质量估计模型(unified neural network for quality estimation,UNQE),该方法有效的对整个神经网络进行联合学习与优化. Fan[18]等提出的"双语专家"模型是一种基于自注意力机制和多头注意力机制的双向 Transformer[19]结构,用于在大规模双语数据基础上进行预训练的语言模型,除利用模型本身提取的特征外,研究人员还设计了一种 4 维的错误匹配特征用于衡量"双语专家"模型所学习到的先验知识与翻译输出之间的差异.

除了基于特征提取的相关 QE 研究外,Okabe[20]等探索了结合文本与视觉形态的多模态译文质量估计,使用多模态的方式提升 QE 系统的性能. 另外,针对带有质量标签 QE 数据稀缺的问题,Rubino[21]等对句子的编码器采用自监督学习的方式,该方法不依赖于 QE 数据,是对基于预训练句子编码器和领域自适应方法的一种补充. Marina[22]等设计了一种无监督学习的 QE 方法,除了训练 NMT 系统所需要的双语数据外,无需额外数据,仅从 NMT 系统中获取有用信息,通过采用不确定性量化的方法可以与人类对质量的判断进行关联,其结果可以媲美效果最佳的有监督 QE 模型.

3 双语专家模型

本研究基于"双语专家"模型基础上开展相关工作. 在汉越译文质量估计任务中,本文使用汉越平行数据(s,t)训练一个特征提取模型,汉越神经机器翻译模型$p(t|s)$ $p(z|s)$是未知的,其中,隐变量z的后验概率可能包含源语句和目标语句之间的浅层语义信息并且有利于下游任务. 根据贝叶斯法则,可以将隐变量z的后验分布表示见公式(1):

$$p(z|t,s)=\frac{p(t|s)\,p(z|s)}{p(t|s)} \tag{1}$$

其中积分$p(t|s)=\int p(t|z)\,p(z|s)\,dz$通常难以解决. 模型中使用一种变分分布$q(z|t,s)$通过 Kullback-Leibler(KL)散度去逼近真实后验. 表示见公式(2):

$$\min D_{KL}(q(z|t,s)\,\|\,p(z|t,s)) \tag{2}$$

除了优化上述目标函数外,我们还可以等效最大化下列目标函数,表示见公式(3):

$$\max E_{q(z|t,s)}[p(t|s)]-D_{KL}(q(z|t,s)\,\|\,p(z|s)) \tag{3}$$

如果在优化的过程中使用单样本蒙特卡洛积分,公式(3)中的第一个期望项可以认为是条件自编码器,与大多数变分自编码器相似,预期对数似然通常用实际替代项近似表示见公式(4):

$$E_{q(z|t,s)}[p(t|z)]\approx p(t|\tilde{z}),\tilde{z}\sim q(z|t,s) \tag{4}$$

该模型包含特征提取阶段与质量估计阶段,特征提取阶段依赖于双向 Transformer 模型,其目的用于提取源语句和译文句对的浅层语义特征并结合 4 维的错误匹配特征输入到下游由 Bi-LSTM 构成的质量估计模块中,从而得到句子级任务的得分预测. 特征提取阶段包含 3 个模块:1)针对源语句,采用基于 transformer 自注意力机制的编码器模块;2)针对目标语句,使用带有 masked 机制的前向和后向自注意力编码器模块;3)重构目标语句模块. 前两个模块所提出的后验概率近似为$q(z|s,t)$,第 3 个模块目标句子的重构过程对应$p(t|z)$. 上述过程通过公式(5)、公式(6)进行描述:

$$q(z|s,t)=\prod_k q(\overrightarrow{z_k}|s,t_{<k})\,q(\overleftarrow{z_k}|t_{>k}) \tag{5}$$

假设遵循高斯分布如$q(\cdot|\cdot)\sim N(\mu(s,t),\sigma^2 I)$,隐变量$\overrightarrow{z_k},\overleftarrow{z_k}$分别从$q(\overrightarrow{z_k}|s,t_{<k})$和$q(\overleftarrow{z_k}|s,t_{>k})$中采样. 每一个语言对$(s,t)$将通过共享神经网络生成自己的均值. 通过固定$\sigma$为超参数和 dropout 的方式,有效地增加隐层表示的不确定性,防止过拟合问题的出现.

$$p(t|z)=\prod_k p(t_k|\overrightarrow{z_k},\overleftarrow{z_k}) \tag{6}$$

在公式(6)中z_k的分布被定义为包含来自源语句以及目标语句中第k个单词周围的上下文,第k个单词代表目标语句中出现翻译错误的单词,但是只有源语句和目标语句中除第k个单词之外的所有单词才会输入到最后一层进行预测. 翻译输出中的第k个单词的潜在表征及其反映错误严重程度的错配特征都有利于下游的质量估计阶段. 最后通过对比发现 Bi-LSTM 模型相较于其对应的变体更适用于质量估计阶段,所以本文将特征提取阶段的双向 Transformer 模型和用于质量估计阶段的 Bi-LSTM 模型相结合的"双语专家"模型作为基线模型,如图 1 所示.

4　融入语言差异化特征

4.1　差异化特征建模

越南语是越南的母语,属于南亚语系,其语法信息主要依靠组成单元的顺序来表达. 越南语的主要语言特征有:

1) 音节是越南语的最小组成单元,这些独立的单元又是多音节的组成部分. 音节间的组合大约有 2500 种,书写越南语时用空格隔开每个音节.

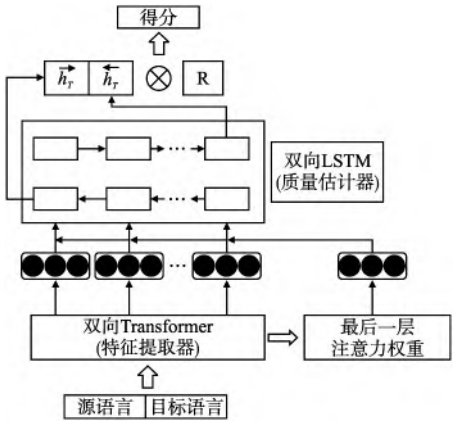

图 1　融入特征模型图

Fig. 1　Incorporate feature model diagram

2) 越南语和汉语都是孤立语,孤立语的典型特征是没有严格意义的形态变化,语法意义大多通过语序和虚词来表达. 具体来说,句子的语义表征会受到单词排序方式的影响. 越南语语句中的词序大体上是一种具体性逐渐增强的过程,位置越靠前的词语其概括性越强,位置越靠后的词语其描述的越具象化. 例如ghế này không ngồi đư ợc bốn người(这个位置最多可容纳 4 个人) .

3) 汉越的主要语法结构排序相同,其结构均为: 主语-谓语-宾语(SVO) . 例如: tôi đọc sách (我读书) ,但是这两种语言的修饰成分在多数情况下不一致,越南语有着较为明显的定语后置、状语后置的特点,如表 1 所示.

表 1　汉越语法结构实例对比

Table 1　Comparison of Chinese and Vietnamese grammatical structure examples

1) 你是(我见过) 的(最善良) 的(女孩) .
1　　3　　中
Cô là (cô gái) (tốt nhất) mà (tôi từng gặp).
中　　3　　1
2) 他是一个(做事干练) 的(办公室) (白领) .
2　　4　　中
Anh ta là một (nhân viên) (văn phòng) (có năng lực).
中　　4　　2

通过对比表 1 中的一些具体例子,可以明显的发现汉越语言定语与中心语之间的排序差异. 其中,1 代表主谓短语; 2 代表动语\介词短语; 3 代表形容词短语及描述性短语; 4 代表描述性名词; 中心语简写为中. 可以看出越语和汉语描述性定语的位置完全不同,但定语修饰中心语的顺序(远近距离) 一致. 汉语描写性多层定语的结构顺序与汉语呈镜像关系. 其中,汉语中描写性定语的顺序是: 1-2-3-4-中心语; 反之,越语的顺序是: 中心语-4-3-2-1.

鉴于上述越南语的语言特点及汉越语法上存在的差异性特点,本文将汉越平行数据中存在的定语后置、状语后置语言情况定义为汉越语言差异化特征,在现有汉语平行数据的基础上利用 GIZA + + 模型进行 IBM Model 的训练,其输入为词汇文件、比特文件与词典文件,词典文件为非必须文件,随着 EM 算法的不断迭代,输出为多个概率表文件,在生成源语言与目标语言之间的翻译概率的同时,也产生了表示源语言与目标语言词对齐关系的文件,我们可以通过获取到的词对齐关系计算出汉越数据中的逆序个数,如表 1 展示的实例 1) 可以清楚的看出,中文“我见过”与对应越南语“tôi từng gặp”,均为其所在句子的主谓短语,中文“女孩”与对应越南语“cô gái”,均为其所在句子的中心语,将同义的两个部分相连接,得到的交点数即为逆序个数.

本文从汉语-越南语方向上获取的源语言与机器译文数据中获取到的逆序个数与目标句句子长度进行一个比值,得到平均逆序数. 以 R 表示平均逆序数,逆序个数表示为 r,目标句子长度表示为 m,公式(7) 为平均逆序数,该值即为本文抽取的汉越语言差异化特征.

$$R = r/m \tag{7}$$

4.2　融入差异化特征

句子级别的分数预测可以表述为具有目标函数的回归问题. 本文将抽取到的特征作为质量估计阶段模型 Bi-LSTM 的最后一个时刻的两个方向的隐状态惩罚因子,见公式(8) :

$$\arg\min \| T - sigmoid(W^{\mathrm{T}} [\overrightarrow{h_T}; \overleftarrow{h_T}] R) \|_2^2 \tag{8}$$

$\overrightarrow{h_T}$和$\overleftarrow{h_T}$(如图 1 所示) 是 Bi-LSTM 输出的最后一个时刻的隐状态. T 代表着译文 TER 分数,W 是一个权重矩阵. 对于句子级译文质量估计的任务,其目的旨在句子层上对机器译文整体质量评分,其本质是预测机器译文的人工后编辑工作(Human-targeted translation edit rate,HTER) ,该值介于 0 ~ 1 之间,但是由于越南语目前尚无公开的人工标注数据集,所以本文采用 TERCOM 工具获取译文句 TER 分数,TER 被定义为将机器译文完全修改为参考句所需要的最小操作数(机器译文与参考句为同一种语言) ,实质上是机器译文和参考句之间的最短编辑距离. 将得到的最小操作数除以参考句中的单词数量:

$$TER = \frac{\#insertions + \#dels + \#subs + \#shifts}{\#referencewords} \tag{9}$$

如公式(9) 所示,定义中的操作包括插入(insertions) 、删除(dels) 、替换(subs) 、移动(shifts) .

5　实　验

5.1　实验数据

实验使用的语料均来自网络爬虫获取的相近领域的汉越对齐数据,为保证模型训练的质量,本文人工去除语料中存在的重复、空行和不规则符号,并过滤了长度大于 80 的句子,源句与目标句长度之比在(1/3-3) 范围内. 最终获得 13.3 万的汉越双语数据; 10 万的汉语、越南语单语数据. 汉语数据平均长度为 17.64,越南语数据平均长度为 24.8.

本文将获取到的汉越平行数随机划分为两个数据集,规

模分别为 10.2 万对和 31481 对,分别用于特征提取阶段与质量估计阶段. 由于数据规模有限,训练汉越神经机器翻译模型的实验数据与特征提取阶段所使用的实验数据为同一组,验证集和测试集均从从对应数据集中随机抽取,其规模大小均为 2k 对. 表 2 为特征提取阶段实验数据信息,表 3 为质量估计阶段实验数据信息.

表 2 特征提取阶段数据信息

Table 2 Data information in feature extraction stage

	训练集	验证集
数据规模	100k	2k

表 3 质量估计阶段数据信息

Table 3 Data information in quality estimation stage

	训练集	测试集	验证集
数据规模	27481	2k	2k

在汉越方向上,本文将质量估计阶段汉越对齐数据中的中文通过汉-越方向的翻译模型获取到对应的越南语机器译文,在越南语真实数据与译文数据的基础上利用 TERCOM 工具获取越南语数据对应的译文质量 TER 分数,在越汉方向,采用同样的方式获取到中文的译文质量 TER 分数. 最终得到质量估计阶段所需要的由源语句、译文句、TER 质量分数组成的三元组(s,m,T). 表 4 为汉-越方向上获取到的译文质量 TER 分数,表 5 为越-汉方向上获取到的译文质量 TER 分数.

表 4 汉-越方向上的 TER 分数

Table 4 TER score in the Chinese-Vietnamese direction

汉越方向	总操作数	目标句总长度	总 TER 分数
训练集	439053	690200	0.6361
验证集	30759	48787	0.6304
测试集	31570	49721	0.6349

表 5 越-汉方向上的 TER 分数

Table 5 TER score in the Vietnamese-Chinese direction

越汉方向	总操作数	目标句总长度	总 TER 分数
训练集	315699	492591	0.6408
验证集	21961	34551	0.6356
测试集	22327	35442	0.6299

5.2 实验设置

对于汉语数据,使用结巴分词工具对中文语句进行分词,对于越南语数据,使用 tokenizer 切开标点. 利用处理过的数据分别从汉-越、越-汉两个方向训练了翻译模型,模型框架选取了 Transformer-base. 训练翻译模型中使用的词表大小为 32k,Transformer 模型编码器和解码器层数均为 6 层,词向量和隐层单元数为 512,批大小为 2048; 汉越神经机器翻译模型所使用的译文质量评价标准均是基于 4 元组 BLEU(BLEU4)值. 特征提取阶段的双向 Transformer 模型自注意机制编码器和前/后向解码器层数均设置为 2,使用了 8 头注意力机制,前馈子层的神经单元数为 512,进行了多 GPU 训练; 质量估计阶段使用了一层的 Bi-LSTM,进行了单 GPU 训练.

5.3 评价指标

对于句子级译文质量估计系统性能的评价指标有皮尔逊相关系数(Pearson Correlation Coefficients)、斯皮尔曼相关系数(Spearman Correlation Coefficients). 本文选用 Pearson 相关系数作为验证方法有效性的评价指标,其范围介于-1 ~ 1,相关系数绝对值越大,表示两个向量的相关性越强,反之则越弱. 本研究对于 TER 分数进行估计,所以希望模型的输出结果与 TER 分数之间的 Pearson 相关系数越接近 1 越好. 可由公式(10)计算得到:

$$Pr=\frac{\sum_{i=1}^{N}(O_i-\bar{O})(T_i-\bar{T})}{\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2\sum_{i=1}^{N}(T_i-\bar{T})^2}} \tag{10}$$

其中,变量 O 表示质量估计系统输出的分数,变量 T 表示 TER 分数,$\sum_{i=1}^{N}(O_i-\bar{O})(T_i-\bar{T})$ 表示为变量 O 和 T 的协方差,$\sum_{i=1}^{N}(O_i-\bar{O})^2$ 和 $\sum_{i=1}^{N}(T_i-\bar{T})^2$ 和 T 分别是变量 O 和 T 的方差.

5.4 实验结果与分析

5.4.1 汉越数据规模有限的情况下,语言差异化特征对于译文质量估计的影响

本文利用人工处理后的汉越平行数据训练了汉越、越汉这两个方向的翻译模型,将汉语与越南语分别作为源语句得到对应的译文,这样就可以获取到训练双语专家模型所需要的译文句子. 表 6 为两个方向翻译模型的 BLEU4 值,在同等数据规模情况下,越-汉方向翻译模型的 BLEU4 值低于汉-越方向的值.

表 6 两个方向上 NMT 模型的 BLEU4 值

Table 6 BLEU value of the nmt model in both directions

数据规模	BLEU4 值(%)	
	汉越方向	越汉方向
100K	20.22	18.16

表 7 为基线模型与融入语言差异化特征的对比结果,这两组实验使用均采用相同的汉-越平行数据,以验证在数据质量、规模相同的情况下,融入语言差异化对于译文质量估计模型性能的影响.

表 7 基线模型与其融入特征的实验结果

Table 7 Experimental results of the baseline model and its integrated features

方法	Pearson 相关系数	
	汉越方向	越汉方向
基线	0.6282	0.5568
基线 + 语言差异化特征	0.6334	0.5603

由表 7 可知,仅使用 100k 汉越平行数据的基线模型,在汉越和越汉两个方向的 Pearson 相关系数分别为 0.6282 与 0.5568,对比试验中,融入语言差异化特征的译文质量估计系统相比于基线模型表现出更好的性能,在汉越方向,较基线模型提升 0.52 个百分点,在越汉方向,较基线模型提升 0.35 个百分点. 这可能是因为汉越平行数据规模有限,特征提取模型学习效果不佳,融入汉越语言差异化特征可以在一定程度上缓解这一问题.

5.4.2 扩充训练数据的情况下,语言差异化特征对于译文质

量估计的影响

汉越平行数据资源稀缺且大部分为跨领域数据，获取难度较大，单语数据相较于汉越平行数据更容易获取，实验为使得双语专家模型尽可能的学到更多翻译知识，包括翻译错误等有利于训练模型的信息，通过回译的方式对汉越训练数据的规模进行扩充. 首先把 100k 的汉语单语数据(mono-zh) 通过回译得到对应的越南语回译数据(syn-vi) ，这样就得到汉语单语数据与越南语回译数据相所构成的 100k 规模的汉越伪平行数据. 以相同的方式，获得了越南语单语数据(mono-vi) 与汉语回译数据(syn-zh) 构成的 100k 规模的汉越伪平行数据. 随后将获得到的两组 100k 规模的汉越伪平行数据分别与特征提取模型所使用的 100k 真实的汉越训练数据以 10k 为基本单位进行结合，最终获得了两组 200k 的汉越合成语料库，分别为添加(mono-zh，syn-vi) 的合成语料库与添加(syn-zh，mono-vi) 的汉越合成语料库. 这样做是为了更有效的扩充我们的训练集，缓解由于数据稀疏带来的负面影响，例如: 特征提取模型训练不佳、特征提取不够充分等问题，并且通过添加伪平行数据，可以更有效地防止过拟合问题.

为了直观的展现添加不同规模的数据对 Pearson 相关系数的影响，本文将实验的结果生成了两组折线图进行比较，图 2 表示使用添加(mono-zh，syn-vi) 的合成语料库训练特征提取模型，图 3 表示使用添加(syn-zh，mono-vi) 的合成语料库训练特征提取模型. 实验结果表明添加合成语料库均对译文质量估计模型的训练产生利好结果，汉越方向的 Pearson 相关系数均优于越汉方向的 Pearson 相关系数.

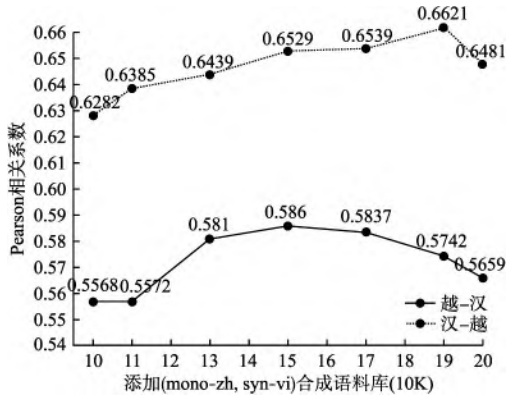

图 2　添加(mono-zh，syn-vi) 合成语料库的实验结果

Fig. 2　Add(mono-zh，syn-vi) synthetic corpus experimental results

在汉越方向上，当(mono-zh，syn-vi) 合成数据规模规模增至 190k，Pearson 相关系数达到 0. 6621 后开始下降，相较于仅使用 100k 汉越平行数据的基线模型，提升了 3. 39 个百分点. 当(syn-zh，mono-vi) 合成数据规模总量增至 200k，Pearson 相关系数达到 0. 6448，相较于基线模型提升了 1. 66 个百分点，而且存在继续上升的趋势.

在越汉方向上，当(mono-zh，syn-vi) 合成数据规模达到 150k，Pearson 相关系数达到 0. 586 后开始下降，相较于基线模型提升了 2. 92 个百分点. 当(syn-zh，mono-vi) 合成数据规模增加到 170k，Pearson 相关系数达到 0. 5927，相较于基线模型提升了 3. 59 个百分点. 两组实验表明，对特征提取模型的训练数据进行扩充，可以显著提升译文质量估计的效果，并且提升幅度远远大于在基线上融入语言差异化特征，从另一方面也反映出数据稀疏问题对汉越的 QE 任务有较大影响.

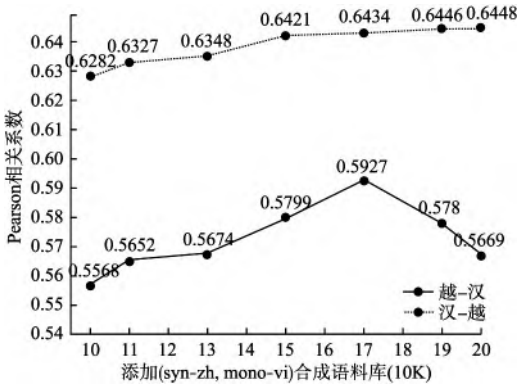

图 3　添加(syn-zh，mono-vi) 合成语料库的实验结果

Fig. 3　Add(syn-zh，mono-vi) synthetic corpus experimental results

为了缓解数据稀疏问题对本任务的干扰，更科学的验证语言差异化特征对于本任务的影响. 我们挑取了汉-越、越-汉两个方向上使用两组合成语料库的最佳实验结果，总计共 4 组. 汉-越方向最佳结果分别为 0. 6621(添加(mono-zh，syn-vi) 合成语料库 190k) 和 0. 6448(添加(syn-zh，mono-vi) 合成语料库 200k) ，越-汉方向上分别为 0. 586(添加(mono-zh，syn-vi) 合成语料库 150k) 和 0. 5927(添加(syn-zh，mono-vi) 合成语料库 170k) . 保留这四组实验参数，在此基础上融入语言差异化特征.

表 8 为汉-越方向上在两组最佳结果的基础上融入语言差异化特征，由实验结果可知，在添加(mono-zh，syn-vi) 合成语料库的最佳结果基础上融入语言差异化特征提升了 0. 32 个百分点，在添加(syn-zh，mono-vi) 合成语料库的最佳结果基础上融入语言差异化特征下降了 0. 45 个百分点.

表 8　汉-越方向上最佳结果与其融入特征的实验结果

Table 8　Best results in the Chinese-Vietnamese direction and the experimental results of its integration characteristics

汉越方向	Pearson 相关系数	
合成语料库	(mono-zh，syn-vi)	(syn-zh，mono-vi)
最佳结果	0. 6621	0. 6448
最佳结果 + 特征	0. 6653	0. 6403

表 9 为越-汉方向上在两组最佳结果的基础上融入语言差异化特征，由实验结果可知，在两组最佳结果的基础上融入语言差异化特征分别提升了 0. 15 个百分点和 0. 24 个百分

表 9　越-汉方向上最佳结果与其融入特征的实验结果

Table 9　Best results in the Vietnamese-Chinese direction and the experimental results of its integration characteristics

越汉方向	Pearson 相关系数	
合成语料库	(mono-zh，syn-vi)	(syn-zh，mono-vi)
最佳结果	0. 586	0. 5927
最佳结果 + 特征	0. 5875	0. 5951

点,但是提升幅度低于越-汉方向在基线基础上融入语言差异化特征.

上述两组实验表明,在训练数据规模增加的情况下,融入语言差异化特征对于译文质量估计系统性能提升能力有限,在汉-越方向上添加(syn-zh,mono-vi)合成语料库 200k 的实验中甚至还产生了负面的影响,这可能是因为加入的伪平行数据规模过大,其中一部分质量较差或者不符合汉越之间的语法规则的数据被特征提取模型所学习,导致译文质量估计系统的性能有所下降,但是从整体的实验情况而言,融入语言差异化特征可以有效提升汉越神经机器翻译译文质量估计任务的表现,尤其是在特征提取阶段的训练数据稀缺这一情况下,效果较为明显.

6 结 论

本文通过分析汉越语言间存在的语言上的差异,对其进行了统计建模,与神经网络本身提取的特征互为补充,在数据规模有限的情况下,本方法有效地缓解了模型对于汉越语言间的特征提取不充分的问题,提升了汉越译文质量估计与机器评价的相关性. 我们也明确了下个阶段的任务: 利用译文质量估计模型对扩充数据进行筛选与修改编辑、结合深度学习的方法去挖掘汉语与越南语之间的语言特性进行更深层次的探索与研究.

References:

[1] Specia L. Exploiting objective annotations for measuring translation post-editing effort [C] //Proceedings of the 15th Conference of the European Association for Machine Translation,2011: 73-80.

[2] Bahdanau Dzmitry,Kyunghyun Cho,Yoshua Bengio. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409.0473,2014.

[3] Sennrich R,Firat O,Cho K,et al. Nematus: a toolkit for neural machine translation [J]. arXiv preprint arXiv: 1703.04357,2017.

[4] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need [C] //Advances in Neural Information Processing Systems,2017: 5998-6008.

[5] Li Ya-chao,Xiong De-yi,Zhang Min. A survey of neural machine translation [J]. Chinese Journal of Computers,2018,41(12): 100-121.

[6] Liu Yang. Recent advances in neural machine translation [J]. Journal of Computer Research and Development,2017,54(6): 1144-1149.

[7] Specia L,Shah K,De Souza J G,et al. QuEst-a translation quality estimation framework [C] //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations,Sofia: ACL,2013: 79-84.

[8] Almaghout H,Specia L. A CCG-based quality estimation metric for statistical machine translation [C] //Proceedings of MT Summit XIV,Langhorne: AMTA,2013: 223-230.

[9] Aaron L F Han. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling [C] //Proceedings of the 8th Workshop on Statistical Machine Translation,Stroudsburg: ACL,2013: 365-372.

[10] Mikolov T,Karaflat M,Burget L,et al. Recurrent neural network based language model [C] //Eleventh Annual Conference of the International Speech Communication Association,Makuhari: DBLP,2010: 1045-1048.

[11] Shi Lei,Zhang Xin-qian,Tao Yong-cai,et al. Sentiment analysis model with the combination of self-attention and tree-LSTM [J]. Journal of Chinese Computer Systems,2019,40(7): 1486-1490.

[12] Shah K,NG Ravmond W M,Bougares F,et al. Investigating continuous space language models for machine translation quality estimation [C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing,Lisbon: EMNLP,2015: 1073-1078.

[13] Chen Z,Tan Y,Zhang C,et al. Improving machine translation quality estimation with neural network features [C] //Proceedings of the 2nd Conference on Machine Translation,Copenhagen: EMNLP,2017: 551-555.

[14] Chen Zhi-ming,Li Mao-xi,Wang Ming-wen. Sentence-level machine translation quality estimation based on neural network features [J]. Journal of Computer Research and Development,2017,54(8): 1804-1812.

[15] Zhu J,Yang M,Li S,et al. Learning bilingual sentence representations for quality estimation of machine translation [C] //China Workshop on Machine Translation,Springer,Singapore,2016: 35-42.

[16] Kim H,Lee J H,Na S H. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation [C] //Proceedings of the 2nd Conference on Machine Translation,2017: 562-568.

[17] Li M X,Xiang Q Y,Chen Z M,et al. A unified neural network for quality estimation of machine translation [J]. IEICE Transactions on Information and Systems,2018,101(9): 2417-2421.

[18] Fan K,Wang J,Li B,et al. Bilingual expert can find translation errors [C] //Proceedings of the AAAI Conference on Artificial Intelligence,Hawaii: AAAI,2019: 6367-6374.

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

[20] Okabe S,Blain F,Specia L. Multimodal quality estimation for machine translation [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020: 1233-1240.

[21] Rubino R,Sumita E. Intermediate self-supervised learning for machine translation quality estimation [C] //Proceedings of the 28th International Conference on Computational Linguistics,2020: 4355-4360.

[22] Fomicheva M,Sun S,Yankovskaya L,et al. Unsupervised quality estimation for neural machine translation [J]. arXiv preprint arXiv: 2005.10608,2020.

附中文参考文献:

[5] 李亚超,熊德意,张 民. 神经机器翻译综述 [J]. 计算机学报,2018,41(12): 100-121.

[6] 刘 洋. 神经机器翻译前沿进展 [J]. 计算机研究与发展,2017,54(6): 1144-1149.

[11] 石 磊,张鑫倩,陶永才,等. 结合自注意力机制和 Tree-LSTM 的情感分析模型 [J]. 小型微型计算机系统,2019,40(7): 1486-1490.

[14] 陈志明,李茂西,王明文. 基于神经网络特征的句子级别译文质量估计 [J]. 计算机研究与发展,2017,54(8): 1804-1812.