基于跨语言预训练模型的朝汉翻译质量评估

赵亚慧，李飞雨，崔荣一，金国哲，张振国，李 德，金小峰

（延边大学 计算机科学与技术学院，吉林 延吉 133002）

摘 要：针对主流翻译质量评估框架在低资源语料上表现较差，句子嵌入策略单一的问题，提出了一个基于跨语言预训练模型的朝汉翻译质量评估模型。首先，借鉴注意力思想提出一种融合跨层信息和词项位置的句子嵌入方法；其次，将跨语言预训练模型引入翻译质量评估任务中，缓解朝鲜语低资源环境带来的数据稀疏问题；最后，对句向量进行回归，实现机器翻译质量评估任务。实验结果表明：该模型能有效提升朝汉翻译质量评估任务性能，与质量评估任务领域主流模型QuEst++、Bilingual Expert、TransQuest相比，皮尔逊相关系数分别提升了0.226、0.156、0.034，斯皮尔曼相关系数分别提升了0.123、0.038、0.026。

关键词：计算机应用；翻译质量评估；跨语言预训练模型；句子嵌入

中图分类号：TP391.1 **文献标志码**：A **文章编号**：1671-5497(2023)08-2371-09

DOI：10.13229/j.cnki.jdxbgxb.20220005

Korean-Chinese translation quality estimation based on cross-lingual pretraining model

ZHAO Ya-hui, LI Fei-yu, CUI Rong-yi, JIN Guo-zhe, ZHANG Zhen-guo, LI De, JIN Xiao-feng

(*Department of Computer Science & Technology, Yanbian University, Yanji* 133002, *China*)

Abstract: On low-resource corpus, the mainstream translation quality estimation models have poor performance. Meanwhile, the sentence embedding strategy is naive. In view of reasons mentioned above, a Korean-Chinese translation quality estimation based on cross-lingual pretraining model is proposed. Firstly, a cross-lingual sentence embedding method is proposed by drawing on the idea of attention. The method can effectively fuse the cross-layer information and token positions of the pre-trained model. Second, a cross-lingual pretraining model is introduced to the task as a way to alleviate the few-shot caused by the low-resource of Korean. Finally, the regression is performed on the sentence embedding vectors, so that the Korean-Chinese translation quality estimation can be completed. Experimental results show that the method can effectively improve the performance of the Korean-Chinese translation quality estimation task. Compared with QuEst++, Bilingual Expert, and TransQuest, the dominant models for quality estimation tasks, Pearson correlation coefficients improved by 0.226, 0.156, and 0.034, and Spearman correlation coefficients improved by 0.123, 0.038, and 0.026, respectively.

收稿日期：2022-01-04.

基金项目：国家社会科学基金重大项目(22&ZD305)；国家自然科学基金项目(62062064)；国家语委“十三五”科研项目(YB135-76)；延边大学外国语语言文学一流学科建设项目(18YLPY13，18YLPY14)；延边大学2020年度校企合作项目(延大科合字[2020]15号).

作者简介：赵亚慧(1974-)，女，教授，硕士.研究方向：自然语言处理.E-mail：yhzhao@ybu.edu.cn

Key words: computer application technology; translation quality estimation; cross-lingual pretraining model; sentence embedding

0 引 言

机器翻译译文质量评估(Quality estimation, QE)作为机器翻译的子任务,由于其在机器翻译的系统评价、译后优化和语料筛选等方面的重要作用,受到学术界和工业界的广泛关注。与常见机器翻译评价指标不同,QE任务能在有监督训练后自动地评估机器生成译文的翻译质量,且在预测阶段不依赖任何参考译文。随着神经网络机器翻译模型的迅速发展,诸多学者开始将深度学习方法引入到翻译质量评估任务中,这使翻译质量评估领域取得了巨大进展。

主流质量评估任务采用预测器-评估器框架实现[1],该方法在特征提取阶段所使用的深度网络需要大规模平行语料作为数据支撑。然而朝鲜语属于低资源语言,针对朝鲜语的各项自然语言处理任务,语料均比较匮乏,尤其在朝汉语言对之间缺乏大规模平行语料,小样本的现状加剧了朝汉机器翻译质量评估任务的难度。目前预训练语言模型在各项任务中表现出了强大的表征能力。预训练模型通过在大规模语料中进行训练学习到语言先验知识,再迁移至不同下游任务中达到提升效果的目的,尤其是跨语言预训练模型为解决各项低资源语言任务带来极大帮助。

跨语言预训练模型可根据上下文生成远距离依赖的词向量,但如何生成高质量的句子表示仍是一个待解决问题。目前大多数方法采用CLS策略[2]或池化操作来生成句子向量[3],但这两种方案都无法包含模型所学全部信息。有大量研究工作表明,预训练模型在不同隐藏层中学习了不同语言学属性的信息[4,5],这为融合预训练模型跨层信息提供了良好理论基础。虽然中间层编码信息具有可转移性,但其在信息编码过程中抛弃掉的低层信息对下游任务可能具有额外帮助,因此使用跨层融合的句子表示能提供更丰富语言信息。

本文将跨语言预训练模型XLM-R引入至句子级翻译质量评估任务中,并提出了一种融合预训练语言模型跨层信息和词项位置的句子嵌入方法。基于预训练模型的翻译质量评估方法,有效利用预训练模型先验知识缓解了朝鲜语低资源问题,提升了模型在小样本数据上的泛化能力,提高了翻译质量评估任务性能。句子嵌入方法利用注意力思想分别在预训练语言模型的不同隐藏层之间和句子的不同词项位置上进行注意力计算,使模型能有效关注到影响质量评估任务的关键语言学信息和词项信息,极大增强了跨语言句子编码的表征能力,同时为其他句子级任务提供了新的句嵌入思路。

1 相关工作

1.1 机器翻译质量评估

机器翻译质量评估不同于机器翻译评价指标,如BLEU、NIST、METEOR等。它可以在不依赖任何参考翻译的情况下,自动给出机器生成翻译的质量预测。常用的质量评分是人工翻译编辑率(Human-target translation edit rate, HTER)。QuEst是special等[6]提出的用于质量评估任务的模型,该模型作为机器翻译质量评估任务的基线模型,在大多数数据集中可以取得较好效果,但其高度依赖于人工提取的语言学特征,因此面对不同语言时不具备普适性。Kim等[1]首先将机器翻译模型应用到质量评估任务中,提出了一种基于RNN的翻译质量评估模型,该模型虽然克服了人工设计特征的问题,但由于RNN固有的无法并行计算和长距离依赖等问题,质量评估效果仍有待提升;Fan等[7]在预测器-评估器框架基础上将基于RNN的翻译特征提取模块替换为Transformer模型,提出了一种双语专家模型,该模型提高了质量评估的性能和可解释性,但其词预测模块仍高度依赖大规模平行语料,不利于低资源语言训练。

1.2 跨语言预训练模型

预训练模型在多项自然语言处理下游任务上取得了显著效果,开创了自然语言处理领域研究的新范式[8]。其处理流程为先使用大量无监督语料进行语言模型预训练,再使用少量标注语料对预训练好的模型进行微调,以完成具体的自然语言处理任务。跨语言预训练模型在预训练模型基础上延伸,通过学习到多种语言的通用表征,使同一个表征能融入多种语言的相同语义。multiB-

ERT[9]在104种语言任务上效果显著，其训练语料为非平行的维基百科语料；XLM（Cross-lingual language model pretraining）[10]在multiBERT的基础上加入了翻译语言模型预训练任务，使用平行语料进行了对齐目标训练，进一步提升了跨语言性能；XLM-RoBERTa[11]在XLM基础上使用更大规模训练数据，通过多语言标记数据和调整模型参数，在多项跨语言任务中都取得了最佳成绩。

1.3　句子嵌入方法

高质量的句子嵌入方法学习到的语言特征可为下游任务提供丰富的外部信息资源。传统的句子嵌入方法包括独热编码、TF-IDF、LSA、LDA等。随着词嵌入技术的成功，句子嵌入模型开始受到学者关注。Arora等[12]使用完全无监督的方法，通过对词向量进行加权平均生成句向量，该方法仅用简单的算法思路，就在句子相似度等任务上拥有较好的表现，相比神经网络的方法更具高效性；Conneau等[13]通过学习句子对的蕴涵或矛盾关系提出了InferSent模型，并使用BiLSTM完成最大池化操作生成句向量；Logeswaran等[14]将上下文预测问题重建为分类问题，分类器根据上下文将目标句与其余对比句区分从而学习到目标句子向量表示。这些基于神经网络的方法性能明显优于无监督方式，同时还进一步探索了适合句向量泛化的自然语言任务，与基于词袋模型的方法相比能极大程度保留句子的语义信息。SBERT[15]和SBERT-WK[16]则基于BERT预训练模型对句子进行嵌入，该类方法能有效利用先验知识从而提取深层语境特征，但方法仅对预训练模型顶层特征进行处理，丢失了大量低层信息，在一定程度上损害了下游任务的性能。

2　融合跨层信息和词项位置的句子嵌入方法

句子级别翻译质量评估任务需要对源语言和目标语言构成的翻译句对进行回归任务，在此过程中涉及到跨语言句子嵌入问题。目前使用预训练模型进行句子级下游任务时，常用的句子嵌入方法是对最后一层得到的隐层向量进行池化操作，或使用模型的[CLS]标记直接进行编码。然而这两种方式都在一定程度上损害了模型习得的信息。首先[CLS]为句子分类任务而设计，在编码过程中会丢失与分类不相关的特征，所以[CLS]无法包含质量评估所需的全部有效信息。其次多项研究表明[4,5]，预训练模型在获得语义知识的过程中不具备层局部性，即预训练模型学习到的知识编码在每一层，不同隐藏层中包含着不同语言学属性的信息。如词汇等信息编码在预训练模型的较低层，而语法、语义信息编码在模型的中高层等。因此对预训练语言模型的跨层信息和位置信息进行融合，可以使句子编码同时兼具不同层次语言学信息和上下文信息。

有效利用模型跨层信息的一个最直接思路就是引入注意力机制，通过借鉴计算机视觉领域对高维数据的处理方式[17,18]，本文提出一个注意力模块对预训练模型的隐藏向量进行句子嵌入。注意力模块包括语言学注意力和词项注意力两部分。语言学注意力通过在预训练语言模型的不同隐藏层之间进行注意力计算，赋予不同隐藏层不同的权重，达到模型能在不同语言学层次上对句对进行翻译质量评估的目的。词项注意力则在不同词项位置上进行注意力计算，使模型更多关注到关键词项信息，同时对语言学注意力进行补充。初始特征$\boldsymbol{X}\in\mathbb{R}^{L\times H\times N}$为跨语言预训练模型在各隐藏层提取到的特征矩阵拼接形成的张量，其中L、H、N分别表示预训练模型的层数、维度和句子对长度。如图1所示，注意力模块根据初始特征$\boldsymbol{X}$分别给出预训练模型不同隐藏层上的语言学注意力矩阵$\boldsymbol{A}_{\mathrm{L}}\in\mathbb{R}^{L\times 1\times 1}$和句子不同位置上的词项注意力矩阵$\boldsymbol{A}_{\mathrm{T}}\in\mathbb{R}^{1\times H\times N}$。整个注意力模块计算过程为：

$$\begin{cases}\boldsymbol{X}'=\mathrm{LangAttn}(\boldsymbol{X})\otimes\boldsymbol{X}\\ \boldsymbol{X}''=\mathrm{TokenAttn}(\boldsymbol{X}')\otimes\boldsymbol{X}'\end{cases}\tag{1}$$

式中：$\otimes$为张量积运算；LangAttn和TokenAttn分别表示语言学注意力和词项注意力的映射运算；$\boldsymbol{X}''$为经过注意力计算后的输出。

在张量积运算过程中，语言学注意力沿着句子词项各位置进行传播。类似地，词项注意力也作用于模型的每一隐藏层。最后通过卷积神经网络完成句子嵌入：

$$\boldsymbol{S}=\mathrm{Flatten}\big(\mathrm{Conv}(\boldsymbol{X}'')\big)\tag{2}$$

2.1　语言学注意力

翻译错误类型繁多是翻译质量评估的难点之一，机器翻译译文与人工翻译相比其错误更加广泛且难以识别。受语言自身特性影响，常见的机器翻译错误包括词汇滥用、语法混乱、内容缺失等类型，涵盖了词汇、语法、语义等多个语言学层次。

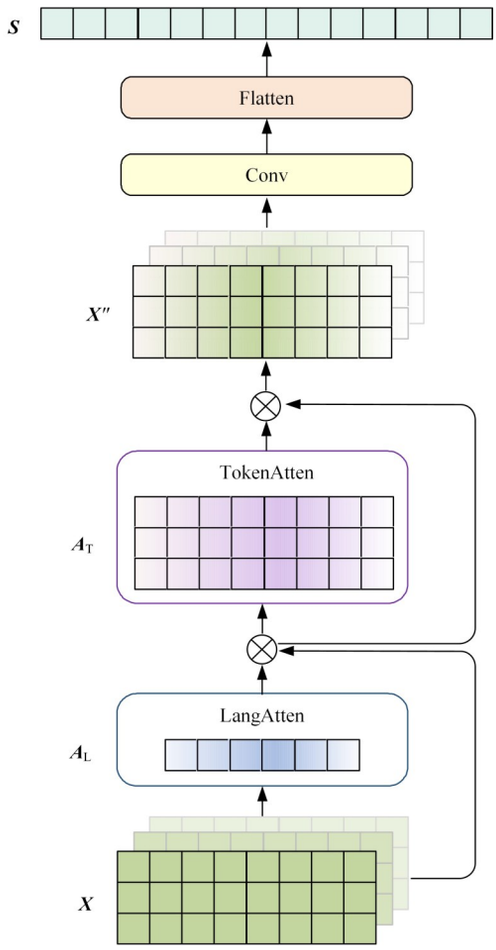

图1 句子嵌入方法示意图

Fig. 1 Schematic diagram of sentence embedding

由于预训练模型的每一隐藏层均表达了不同语言学层面的信息[4,5],因此使用语言学注意力有助于模型更加关注到影响评估任务的关键语言学属性。为融合预训练模型蕴含的多样化的语言学知识,本文利用不同隐藏层之间的特征关系生成语言学注意力矩阵。语言学注意力结构如图2所示。

为了使每一隐藏层特征都包含句子的上下文信息,首先将隐藏层的全局信息进行压缩,分别使用平均池化和最大池化操作聚合句子词项的位置信息,并生成两个不同的上下文特征 $\boldsymbol{X}_{\text{avg}}^{\text{L}}\in\mathbb{R}^{L\times1\times1}$ 和 $\boldsymbol{X}_{\text{max}}^{\text{L}}\in\mathbb{R}^{L\times1\times1}$:

$$\begin{cases}\boldsymbol{X}_{\text{avg}}^{\text{L}}=\text{AvgPool}(\boldsymbol{X})\\ \boldsymbol{X}_{\text{max}}^{\text{L}}=\text{MaxPool}(\boldsymbol{X})\end{cases}\tag{3}$$

该上下文特征的统计信息可表达整个句子,便于之后捕获不同隐藏层之间的相关依赖关系。为使注意力模块能学习到不同隐藏层之间的非线性交互,这两个包含全局词项信息的上下文特征通过共享神经网络生成两个注意力矩阵,最后使用元素求和合并输出最终语言学注意力矩阵

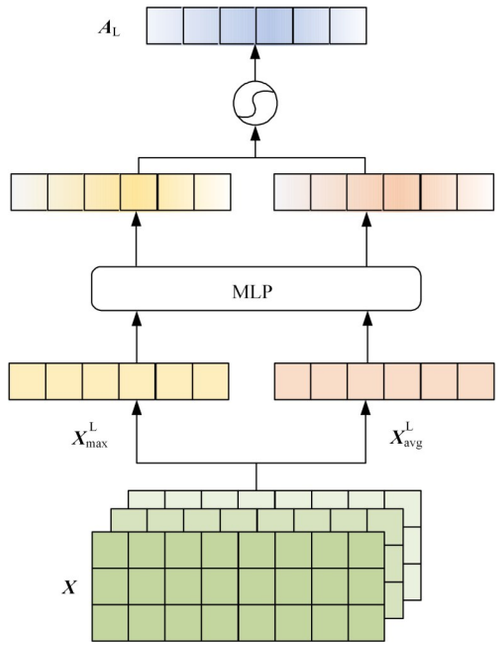

图2 语言学注意力结构示意图

Fig. 2 Schematic diagram of linguistic attention

$\boldsymbol{A}_{\text{L}}\in\mathbb{R}^{L\times1\times1}$,计算过程为:

$$\boldsymbol{A}_{\text{L}}=\text{LangAttn}(\boldsymbol{X})=\sigma\left(\left(\boldsymbol{W}_1\delta\left(\boldsymbol{W}_0\boldsymbol{X}_{\text{avg}}^{\text{L}}\right)\right)+\left(\boldsymbol{W}_1\delta\left(\boldsymbol{W}_0\boldsymbol{X}_{\text{max}}^{\text{L}}\right)\right)\right)\tag{4}$$

式中:σ 为 Sigmoid 函数;δ 为 ReLU 激活函数;$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 为网络共享参数。

2.2 词项注意力

语言学者认为在翻译过程中,语境是比语法、语义等信息更为重要的因素。语境通常指句子范围内使词项产生意义的上下文信息。在机器翻译过程中,要求模型可以根据提取到的上下文特征实现词义消歧、形态变化等任务。在预训练模型中句子分词后进行词嵌入并按词项位置顺序构成矩阵,因此使用词项注意力有助于模型更加关注到输入句子中对评估任务起决定作用的词项。本文使用词项注意力对语言学注意力进行补充,利用待评估句对的位置关系生成词项注意力矩阵。词项注意力结构如图3所示。

为有效凸显不同词项位置重要性,首先将同一词项位置的各层次语言学信息进行压缩,分别沿不同隐藏层方向进行平均池化和最大池化操作并生成 $\boldsymbol{X}_{\text{avg}}^{\text{T}}\in\mathbb{R}^{1\times H\times N}$ 和 $\boldsymbol{X}_{\text{max}}^{\text{T}}\in\mathbb{R}^{1\times H\times N}$,然后将其拼接得到最终特征 $\boldsymbol{X}^{\text{T}}\in\mathbb{R}^{1\times H\times N}$,计算过程为:

$$\boldsymbol{X}^{\text{T}}=\text{Concat}\left(\boldsymbol{X}_{\text{avg}}^{\text{T}},\boldsymbol{X}_{\text{max}}^{\text{T}}\right)=\text{Concat}\left(\text{AvgPool}(\boldsymbol{X}'),\text{MaxPool}(\boldsymbol{X}')\right)\tag{5}$$

为使词项注意力模块能够学习到不同词项之

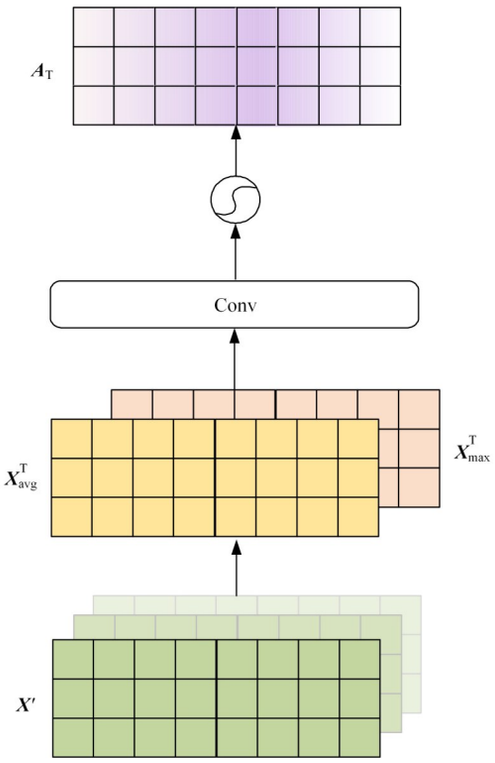

图 3　词项注意力结构示意图

Fig. 3　Schematic diagram of token attention

间的非线性交互并对重要词项位置进行关注，最后应用标准卷积层对 $\boldsymbol{X}^{\mathrm{T}}$ 进行卷积生成词项注意力矩阵 $\boldsymbol{A}_{\mathrm{T}} \in \mathbb{R}^{1\times H\times N}$，计算过程为：

$$\boldsymbol{A}_{\mathrm{T}} = \mathrm{TokenAttn}(\boldsymbol{X}') = \sigma\left(\mathrm{Conv}\left(\boldsymbol{X}^{\mathrm{T}}\right)\right) \quad (6)$$

式中：σ 为 Sigmoid 函数。

3　基于跨语言预训练模型的翻译质量评估模型

质量评估任务和机器翻译任务均涉及到两种不同语言，因此模型在某种程度上必须同时理解源语言和目标语言以及它们间的转换关系。大量基于神经网络的特征提取器配合分类器的评估架构取得了较好成绩，但该架构必须使用大规模平行语料作为特征提取器的训练数据输入。朝鲜语受限于资源条件，不仅缺乏大量人工后编辑注释的翻译数据集，同时朝汉语言对之间缺乏大规模平行语料，因此使用主流预测器-评估器框架独立进行朝汉翻译质量评估任务训练效果并不理想。

近年来，跨语言预训练模型在各项跨语言自然语言处理任务中展示出了巨大潜力，为低资源语言任务开辟了新的研究范式和解决思路。因此本文将跨语言预训练模型引入到翻译质量评估任务中，首先基于 XLM-R（Cross-lingual language model pretra-ining RoBERTa）对待评估源语言及翻译译文句对进行句子嵌入，其次直接使用生成的句子向量实现翻译质量评估任务。基于跨语言预训练模型的翻译质量评估模型主要结构如图 4 所示。

模型输入由源语言句子和目标语言的机器译文拼接构成，同时在序列首位和语言之间分别添加特殊嵌入符[CLS]和[SEP]，由于质量评估与翻译语言建模任务在跨语言理解、对齐等方面的相似性，本文认为采用该输入方式有助于模型拥有理解“翻译质量”的能力，而非简单地针对小规模数据分布进行强拟合。不同于以往的做法[2,3]，我们将 XLM-R 模型的每一隐藏层和各词项位置输出均用于翻译质量评估任务，并增添注意力机制帮助模型更加关注有益于质量评估任务的语言学层次和词项位置。隐藏层状态 $\boldsymbol{h}_i = \left(\boldsymbol{h}_{[\mathrm{CLS}]}, \boldsymbol{h}_{\mathrm{src}1}, \cdots, \boldsymbol{h}_{\mathrm{src}m}, \boldsymbol{h}_{[\mathrm{SEP}]}, \boldsymbol{h}_{\mathrm{tar}1} \cdots, \boldsymbol{h}_{\mathrm{tar}n}\right)$ 为预训练模型第 i 层待评估句对和特殊标志位词向量的按序拼接，将模型 24 个隐藏层和词嵌入层 e 特征拼接后得到句子初始评估特征：

$$\boldsymbol{X} = \mathrm{Concat}(e, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{24}) \quad (7)$$

由于预训练模型在大规模语料上进行过翻译语言建模任务，因此初始特征 $\boldsymbol{X}$ 包含翻译评估相关先验知识。为使预训练模型的翻译知识能够有效迁移至后续质量评估任务中，模型对初始特征 $\boldsymbol{X}$ 进行了不同语言学层次和词项位置注意力计

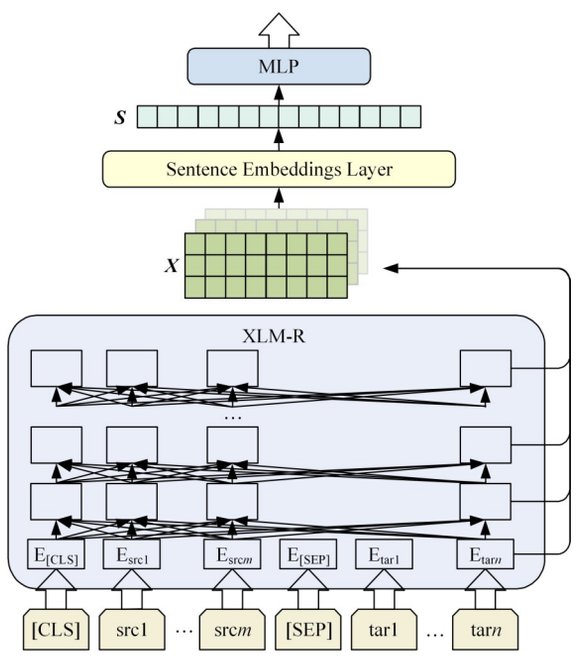

图 4　朝汉翻译质量评估模型

Fig. 4　Architecture of Korean-Chinese translation quality estimation

算,由式(1)(2)所述得到最终句子嵌入$\boldsymbol{S}$。

由于高复杂度神经网络易对低资源语言产生较大负面影响,因此在顶端的回归任务中,模型不依赖复杂的输出层,仅使用简单全连接神经网络计算得到质量得分:

$$y_{\text{score}}=\sigma\left(\boldsymbol{w}^{\mathrm{T}}\left(\tanh\left(\boldsymbol{W}\mathbf{s}\right)\right)\right) \tag{8}$$

式中:σ为Sigmoid函数;tanh为双曲正切函数;$\boldsymbol{w}$、$\boldsymbol{W}$为全连接网络参数。

模型通过最小化均方误差进行训练,损失函数为:

$$\text{loss}=\left\| y_{\text{score}}-\hat{y}_{\text{score}}\right\|_2^2 \tag{9}$$

计算得到的值y_{score}即为翻译质量评估模型对译文的评分。

4 实验结果及分析

4.1 数据集及实验设置

实验使用数据集来自实验室承担"中韩科技信息加工综合平台"项目构建的中-英-韩语料库[19]。原始语料库包含32 688篇科技文献,涵盖生物技术、海洋环境、航空航天等13个领域。经分句对齐后,朝汉平行句对数量为163 449对。朝鲜语使用Kkma工具按词素为单位对句子进行划分,分析得到词素总数为32 414。

质量评估得分使用人工后编辑距离(Human-target translation edit rate, HTER)。HTER是一种基于编辑距离的度量方法,它根据将一个翻译转换为另一个翻译所需的修改数量捕获机器译文和参考翻译之间的距离,因此质量评估模型需合理预测该分数。HTER得分由TERCOM工具自动计算得到。本文选取XLM-RoBERTa-large版本[20]作为实验所用跨语言预训练模型,包含24个隐藏层和一个词嵌入层,各隐藏层维度大小为1024。评估模型句子padding长度为80,梯度优化算法使用Adam优化器,学习率设置为2×10^{-5}。

4.2 翻译质量评估实验

为验证模型评估性能,本文在相同硬件条件和语料规模下与代表性翻译质量评估模型进行朝汉机器翻译质量评估任务对比实验。其中QuEst++[6]为WMT2013-WMT2019官方基线系统;Bilingual Expert[11]为预测器-评估器框架下的一个先进结果;TransQuest[3]为基于跨语言预训练模型的评估模型,其在WMT2020多语言直接评估中表现最优。分别计算各模型在测试集上预测值与真实值之间的相关性,结果如表2所示。

从表2中可以看出,本文模型的预测评分与人工评分的相关性超过了所有基线模型,皮尔逊相关系数分别提升了0.226、0.156、0.034,斯皮尔曼相关系数分别提升了0.123、0.038、0.026。这表明通过融合不同层次语言学知识和词项位置信息能有效提升翻译质量评估任务性能。此外在实验中发现,本文方法和基于预训练模型的TransQuest方法所得到的均方误差值和均方根误差与其他模型相比结果稍差。这是由于主流预测器-评估器方法在分类模块使用了复杂神经网络,该策略实际是针对小规模数据进行的强拟合,缺乏真正理解"翻译质量"的能力。而预训练模型使用的大规模数据使得提取特征不局限于表层信息,具备了强通用性和弱针对性,因此虽然误差表现略差但相关性水平更高。

为验证模型的有效性及通用性,本文同时在WMT202任务1提供的DA数据集上进行了实验。实验分别在中低资源语言到英语语向上进行,包括罗日尼亚语(Ro)、爱沙尼亚语(Et)、尼泊尔语(Ne)和僧伽罗语(Si)。分别计算各语向上模型评估结果与人类打分的皮尔逊相关系数,实验结果如表3所示,可以看出,本文模型的预测评分与人工评分的相关性超过了TransQuest模型,表明了本方法的有效性。

表2 各模型相关系数得分

Table 2 Correlation coefficient of each model

模型	Pearson	Spearman
QuEst++	0.397	0.471
Bilingual Expert	0.476	0.556
TransQuest	0.589	0.568
本文	**0.623**	**0.594**

表3 WMT2020皮尔逊相关系数得分

Table 3 Pearson correlation coefficient in WMT2020

模型	Ro-En	Et-En	Ne-En	Si-En
TransQuest	0.898	0.775	0.791	0.652
本文	0.902	0.820	0.827	0.705

4.3 句子嵌入方法对比实验

由于模型使用了融合跨层信息的句子嵌入方法对待评估句对进行句子嵌入,为验证该策略的合理性和有效性,本文使用不同句子嵌入方式进行翻译质量评估任务实验。其中Last[CLS]为仅使用XLM-R模型顶层[CLS]标签向量作为句子嵌入;Last+GRU表示将XLM-R模型顶层矩阵

通过GRU网络得到句子向量；Conv为使用XLM-R模型不同隐藏层直接进行卷积而不进行注意力计算进行句子嵌入；Attention为使用不同隐藏层信息进行注意力计算得到句子嵌入。分别计算各句子嵌入方法下质量评估模型性能，具体结果如表4所示。

从表4中可以看出，使用XLM-R全部层次进行卷积效果最差，仅适用顶层[CLS]标签向量次之。这是由于使用全部隐藏层信息直接进行卷积时，模型原始特征中包含大量与翻译质量评估任务无关的冗余信息，不加以筛选则为后续下游任务引入过多噪声进而影响了评估性能。仅使用顶层[CLS]标签则丢失了过多预训练模型习得的低层信息，而低层信息包含多种质量评估任务关键的词汇级别语言特征，因此该嵌入方法效果较差。在模型顶层特征矩阵后添加GRU网络虽在均方根误差上取得了最好结果，但本文认为这同样是针对数据进行强拟合的体现，对于质量评估任务本质无较大意义，故其相关性结果表现一般。因此，本文提出的句子嵌入方法更利于模型进行质量评估任务。

表4 不同句子嵌入方法模型性能

Table 4　Performance of different sentence embedding

句子嵌入方法	Pearson↑	Spearman↓	MAE↓	RMSE↓
Last[CLS]	0.589	0.568	0.160	0.204
Last+GRU	0.597	0.571	0.145	0.180
Conv(0-24)	0.554	0.544	0.153	0.191
Conv(20-24)	0.605	0.582	0.150	0.188
Attention(1-24)	0.609	0.591	0.146	0.185
Attention(0-24)	0.623	0.594	0.144	0.183

4.4　注意力对比实验

为有效融合模型跨层信息和不同位置词项信息，本文在句子嵌入方法中引入了注意力机制。由于注意力机制包括语言学注意力和词项注意力两个模块，因此本节对两个注意力模块进行消融实验和结合顺序实验，以此验证本文注意力机制设计的合理性和有效性。实验结果如表5所示，可以发现，模型采用两种顺序组合方式注意力模块性能均优于单独使用一个注意力模块，这表明本文提出的两个注意力模块均在质量评估任务中发挥了作用。同时，在第一阶段使用语言学注意力比使用词项注意力表现稍好，因此本文在句子嵌入过程中使用该排列策略。

表5 不同注意力顺序模型性能

Table 5　Performance of different attention order

注意力方法	Pearson	Spearman	MAE	RMSE
LangAtten	0.592	0.570	0.159	0.190
TokenAtten	0.600	0.572	0.145	0.189
TokenAtten+LangAtten	0.613	0.592	0.149	0.187
LangAtten+TokenAtten	0.623	0.594	0.144	**0.183**

4.5　实例分析

考虑到注意力机制在本模型实现中的重要性，本文从开发集中抽取部分实例对训练过程中的语言学注意力矩阵进行可视化，结果如图5所

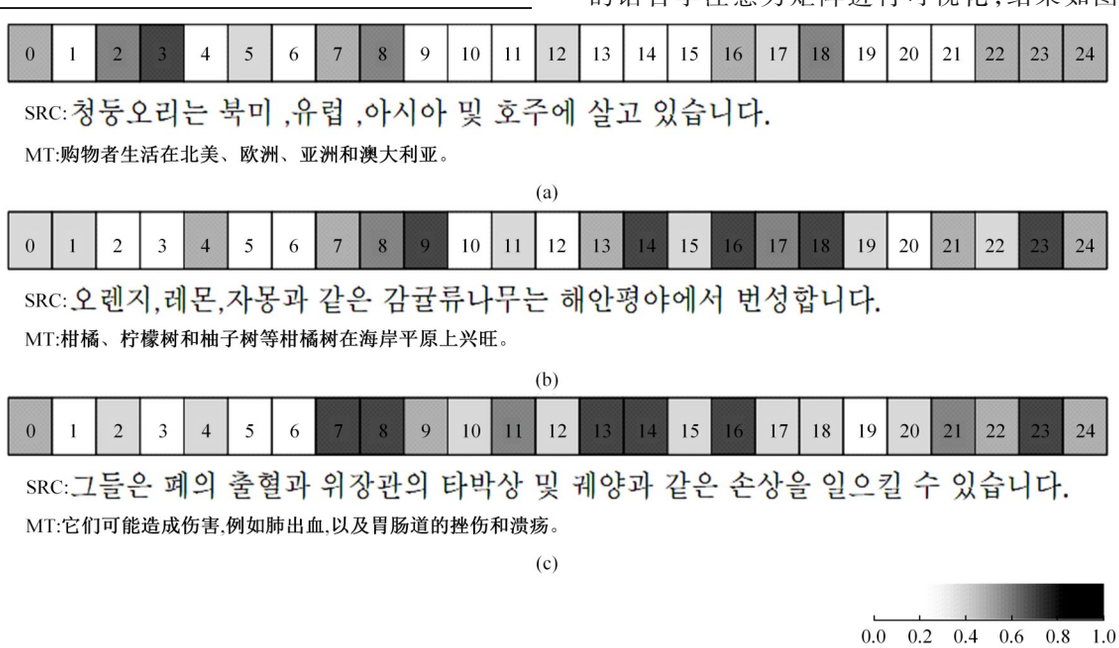

图5　结果示例

Fig. 5　Examples of result

示。图5展示了本模型习得的语言学注意力在25个隐藏层(含词嵌入层)上的权重分布,SRC和MT分别表示待评估句对的源语言句子和目标语言机器译文。通过该实验结果可进一步定性分析本文提出的翻译质量评估模型的效果。

如图5所示,机器译文错误类型广泛,主要体现在词汇滥用、语序混乱、内容缺失等不同属性语言学层面。图5(a)中机器译文误将"청둥오리"译为"购物者",这属于典型词汇滥用现象,因此针对其中待评估句对,模型主要关注于预训练语言模型的低层次词汇信息,与此类似的现象还存在于图5(b)对"오렌지"和"자몽"的翻译中。但相比于简单的词汇误译,更为棘手的是针对一词多义现象的翻译质量评估问题。结合上下文可知图5(b)中"번성하다"的词义为"茂盛"而非"兴旺",因此本文模型针对图5(b)句对除关注到低层次词汇信息外,还较多关注到了高层次的语义信息。图5(c)中译文错误主要集中在语序混乱、可读性差、缺少宾语等语法层面的问题,因此模型针对该句关注较多的部分为预训练模型的中高层次。综上,本文认为语言学注意力模块可以迫使模型在句嵌入时理解不同语言学属性与最终质量判断的相关关系,同时关注到有助于评估翻译质量的关键信息,有效提升了翻译质量评估任务的水平。

5 结束语

机器翻译译文质量评估是机器翻译领域的重要研究子课题,目前主流方法基于预测器-评估器框架,在朝鲜语低资源环境下提升朝汉翻译质量评估性能面临诸多挑战。本文将跨语言预训练模型引入翻译质量评估任务,并提出了一种融合预训练模型跨层信息的句子嵌入方案。实验结果表明:该方法能有效提升朝汉翻译质量评估任务性能。在下一步工作中,将结合中文和朝鲜语的语言特点探索更具语言适配性的翻译评估模型,进一步改善朝鲜语数据稀疏问题。

参考文献:

[1] Kim H, Lee J. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation[J]. ACM Transactions on Asian and Low-resource Language Information Processing, 2017, 2: 562-568.

[2] Takahashi K, Sudoh K, Nakamura S. Automatic machine translation evaluation using source language inputs and cross-lingual language model[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, Washington, United States, 2020: 3553-3558.

[3] Ranasinghe T, Orasan C, Mitkov R. TransQuest: translation quality estimation with cross-lingual transformers [C]//Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 2020: 5070-5081.

[4] Jawahar G, Sagot B, Eddah D. What does BERT learn about the structure of language? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 3651-3657.

[5] Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 4996-5001.

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

[7] Fan K, Wang J Y, Li B, et al. "Bilingual Expert" can find translation errors[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Phoenix, United States, 2019: 6367-6374.

[8] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized bert pretraining approach[C]//Proceedings of the 2020 International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020: 1-15.

[9] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, United States, 2019: 4171-4186.

[10] Conneau A, Lample G. Cross-lingual language model pretraining[J]. Advances in Neural Information Processing Systems, 2019, 32: 7059-7069.

[11] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, Washington, United States, 2020: 8840-8451.

[12] Arora S, Liang Y Y, Ma T Y. A simple but tough-to-beat baseline for sentence embeddings[C]//Proceed-

ings of the 5th International Conference on Learning Representations, Toulon, France, 2017: 1-16.

[13] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017: 670-680.

[14] Logeswaran L, Lee H. An efficient framework for learning sentence representations[C]//Proceedings of the 6th International Conference on Learning Representations, Pennsylvania, United States, 2018: 1-16.

[15] Reimers N, Gurevych I, Reimers N, et al. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 2019: 3982-3992.

[16] Wang B, Kuo J. SBERT-WK: a sentence embedding method by dissecting BERT-based word models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2146-2157.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, United States, 2018: 7132-7141.

[18] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision, Antibes, France, 2018: 3-19.

[19] 赵亚慧, 杨飞扬, 张振国, 等. 基于强化学习和注意力机制的朝鲜语文本结构发现[J]. 吉林大学学报: 工学版, 2021, 51(4): 1387-1395.
Zhao Ya-hui, Yang Fei-yang, Zhang Zhen-guo, et al. Korean text structure discovery based on reinforcement learning and attention mechanism[J]. Journal of Jilin University(Engineering and Technology Edition), 2021, 51(4): 1387-1395.

[20] 李健, 熊琦, 胡雅婷, 等. 基于 Transformer 和隐马尔科夫模型的中文命名实体识别方法[J]. 吉林大学学报: 工学版,2023, 53(5): 1427-1434.
Li Jian, Xiong Qi, Hu Ya-ting, et al. Chinese named entity recognition method based on transformer and hidden markov model[J]. Journal of Jilin University(Engineering and Technology Edition), 2023, 53(5): 1427-1434.