基于 ELECTRA 的翻译质量估计建模

孟庆晔，杨沐昀*，李岳旸，赵铁军，朱聪慧，曹海龙

（哈尔滨工业大学计算机科学与技术学院，哈尔滨 黑龙江 150006）

摘要：针对翻译质量估计任务存在的数据稀缺问题，从模型和数据两个方面进行改进。模型方面，在基于“预测器-估计器”结构的模型基础上，使用同一时刻能够进行双向语义表示的 ELECTRA 预训练模型代替原有同一时刻只能单向语义表示的预测器，并在质量估计阶段使预测器继续参与训练，采取微调的方法联合估计器参数共同更新；数据方面，使用 ELECTRA 的生成器对输入句子中部分词进行替换改写的方法构造翻译质量估计伪数据，扩大了训练数据的规模。实验表明上述措施均提升了翻译质量估计模型性能，并在多个数据集上超越了基线模型的结果。

关键词：翻译质量估计；预训练语言模型；ELECTRA；伪数据

中图分类号:TP 391　　**文献标志码:**A

1　引言

机器翻译质量估计技术（Quality Estimation, QE）作为一种不需要参考译文的机器译文自动评价技术，在仅使用源语言和机器翻译译文的情况下即可对译文的质量进行评估，因此具有广泛的应用价值。翻译质量估计模型训练一般面临着数据集普遍规模较小、数据稀缺的问题。这是由于翻译质量估计任务的数据标注需要请专业的翻译人员进行人工标注，即对机器翻译译文进行后编辑，这一过程需要消耗大量的时间和人力。

翻译质量估计技术经历了从单纯的利用统计机器学习来进行特征提取，到利用神经网络结合特征提取，再到完全利用深度学习模型的过程。随着近几年来深度神经网络模型的快速发展，对翻译质量估计的促进作用极大地激发了研究者们的热情，研究者们不再把注意力放在传统特征的提取和选择上，而是更多地利用深度神经网络模型的优势挖掘待质量评估文本的内在信息，相应的完全基于深度神经网络的翻译质量估计技术相继出现。期间主要工作包括：Kreutzer[1]等人在 2015 年提出了基于上下文窗口的翻译质量估计模型 QUETC；Martins[2]等人对 QUETC 模型进行改进提出了多层神经网络模型；Kim[3]等人在 2016 年针对句子级的质量估计提出了基于 Predictor-Estimator 结构的循环神经网路翻译质量估计模型，这个尝试开创性的使用了大规模训练预料作为输入信息，在很大程度上解决了翻译质量估计任务领域的训练集稀少的缺陷，同时使模型不再局限于特定的机器翻译系统的评价，取得了当年最好的研究效果。随着 Transformer[4]的提出和兴起，研究者们利用 Transformer 提出了一系列翻译质量估计模型[5-6]，其中阿里巴巴 Kai[5]等人在 2018 年提出的 Bilingual Expert（双语专家模型）。其在 Kim 等人的研究基础上，在预训练阶段采用表达能力更强的基于注意力机制的 Transformer 替换原有的循环神经网络结构，在 WMT2017/2018 翻译质量估计任务的大多数公共可用数据集中均达到了最先进的水平。此后，随着 BERT 这一双向预训练语言模型的横空出世，将预训练语言模型与翻译质量估计领域结合成为了人们研究的热点[7]，研究者们在 WMT2019 上采用 BERT 和 XLM[8]等预训练语言模型，并采用新的融合技术组合句子级和词级的预测，取得了更好的效果[9-10]。

基于“预测器-估计器”结构的 Predictor-Estimator 模型以及 Bilingual Expert 模型在预测阶段虽然可以从正向和反向进行词预测，但是这种单向编码的机制使得模型只能在同一时刻关注句子中的部分信息，不能结合上下文的语义信息进行整体预测从而导致预测器与训练阶段学习到的语义表征不够充分，从而影响后续质量估计阶段的准确性。另外，这种两阶段式的训练方法会存在训练数据和训练目标的差异性。由于预测阶段使用的是不带有翻译错误的完全正确的平行语料而质量估计阶段会引入带有错误信息的 QE 数据，而上述两个模型在预训练阶段结束后将预测器模型参数冻结，仅作为质量估计阶段的特征抽取器不继续参与模型训练，这样会导致预测器的模型不能很好兼容于下游翻译质量估计任务。

针对上述问题，本文提出基于 ELECTRA[11]预训练模型的翻译质量估计模型，在预训练阶段采用 ELECTRA 这种双向表示的预训练语言模型，实现了对预训练语料上下文的充分学习，从而获得表征能力更强的预测器来支撑下游的翻译质量估计任务；在翻译质量估计阶段，预训练模型继续参与训练，将 QE 数据作为预测器的输入联合估计器进行共同训练，而不是固定模型参数作为特征向量提取，

预训练模型采取类似微调的联合训练方式有效地减缓了预训练阶段模型和质量估计阶段模型差异性的问题。进一步地，本文还利用 ELECTRA 的生成器作为生成 QE 伪数据的手段从而扩大了数据集的规模，并将得到的伪数据参与训练。实验表明，本文提出的两点改进均有效提升了翻译质量估计模型的性能。

2 基于 ELECTRA 的翻译质量估计模型结构

本文提出的模型分为预训练和翻译质量估计两部分，下面将分别对这两部分进行介绍。

2.1 预训练部分

本文使用基于替换令牌检测(replaced token detection, RTD)机制的 ELECTRA 作为提出模型的预测器参与预训练阶段。ELECTRA 预训练模型采用一种替换令牌检测的新预训练任务形式，该任务在从所有输入词的位置学习语义信息的同时训练了一个基于双向表达的预训练语言模型。ELECTRA 作为基于这种替换令牌检测任务机制的一种预训练语言模型，模型结构如图 1 所示。

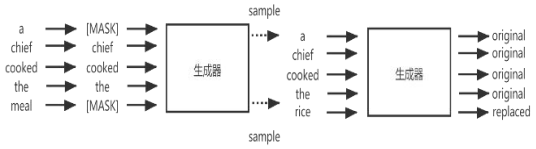

图 1　ELECTRA 模型结构图

模型分成生成器和判别器两部分，每一部分的具体结构都是采用 Transformer 的编码器结构。模型的流程为，首先输入一个句子，以一定的比例随机将句子中的位置替换为[MASK]标识符，生成器的作用是对句子中遮盖掉的部分词进行预测并使用预测的词替换原始位置，从而生成了一个可能带有部分被替换的句子，而判别器的作用则是判断通过生成器输出后的句子中的每个词是原始的还是被替换后的。上述过程描述的就是 ELECTRA 预训练模型的替换令牌检测任务机制，其具体的形式化表示如下：

生成器输入令牌（token）序列 $x=[x_1,...x_n]$ 经过编码得到 $h_G(x)=[h_1^G,...h_n^G]$。于是，生成器在位置 t 上输出 token x 的概率为:

$$P_G\left(x_t \mid x\right)=\text{sotfmax}\left(\text{e}(x)^T h_t^G\right) \tag{1}$$

其中，e 表示 token 的词嵌入表示（embedding）。而对于判别器，输入令牌序列 $x=[x_1,...x_n]$，经过编码得到 $\text{h}_D(x)=[h_1^D,...h_n^D]$。于是判别器预测位置 t 的令牌被替换的概率为:

$$D\left(x,t\right)=\text{sigmoid}\left(w^T h_t^D\right) \tag{2}$$

其中 w 是模型参数。

假设用代表原始输入中随机 token 位置被遮盖掉后的句子，代表生成器生成的替换后的句子。则模型生成器部分的损失函数为：

$$L_{\text{MLM}}\left(\mathbf{x},\theta_G\right)=\text{E}\left(\sum_{i\in\mathbf{m}}-\log p_G\left(x_i \mid x^{\text{masked}}\right)\right) \tag{3}$$

模型判别器部分的损失函数为：

$$\begin{aligned}&L_{Disc}(\boldsymbol{x},\theta_D)\\&=\text{E}\,(\sum_{t=1}^{n}-\text{I}(x_t^{\text{corrupt}}=x_t)\log\text{D}(\boldsymbol{x}^{\text{corrupt}},t)\\&-\text{I}(x_t^{\text{corrupt}}\neq x_t)\log(1-D(\boldsymbol{x}^{\text{corrupt}},t)))\end{aligned} \tag{4}$$

最终模型的优化目标为在一个大型语料库 x 下，最小化生成器损失函数与判别器损失函数的加和。

$$\min_{\theta_G,\theta_D}\sum_{\boldsymbol{x}\in\mathbb{X}}L_{\text{MLM}}\left(\boldsymbol{x},\theta_G\right)+\lambda L_{\text{Disc}}\left(\boldsymbol{x},\theta_D\right) \tag{5}$$

2.2 翻译质量估计部分

句子级别的翻译质量估计任务可看做给定原文及对应机器翻译译文，模型对 HTER 分数拟合的一个过程，因此可以归为一个回归任务。同时，由于 ELECTRA 为单语预训练模型而翻译质量估计任务为跨语言任务的原因，本文提出了如图 2 所示的适用于句子级 QE 任务的基于 ELECTRA 的翻译质量估计模型。

下面介绍本模型的具体实施细节。首先，将序列长度（这里指 token 数）分别为 n 和 n'的源文、机器译文输入到预训练过的 ELECTRA 判别器中，针对每一个 token，取预训练模型最后一层的隐状态（hidden states）输出 h_D 并进行平均池化操作（mean pooling），得到了源文、机器译文对应的隐向量表示 u, v。如公式(6)、(7)所示。

$$u=\text{mean}\left(E\left(src\right)\right)=\frac{1}{n}\sum_{t=1}^{n}h_D(src)_t \tag{6}$$

$$v=\text{mean}\left(E\left(mt\right)\right)=\frac{1}{n'}\sum_{t=1}^{n'}h_D(tgt)_t \tag{7}$$

接着，将 u，v 进行连接（concat）操作并输入到一层全连接层（fully connected layers, FC）中，使用 sigmoid 作为激活函数，最终得到取值在 0-1 之间的输出表示 *output*。在具体 concat 方式上采取 u，v，$|u\text{-}v|$三者进行连接。

$$output=\text{sigmoid}\left(\text{FC}\left(\text{concat}\left(u,v\right)\right)\right) \tag{8}$$

其中：

$$\text{sigmoid}\left(x\right)=\frac{1}{1+e^{-x}}\,\text{R}\ \rightarrow(0,1) \tag{9}$$

最后，句子级别翻译质量估计任务作为回归任务，采用均方误差（Mean Square Error, MSE）作为损失函数：

$$MSE\left(x,\tilde{x}\right)=\frac{1}{n}\sum_{i=1}^{n}(x-\tilde{x})^2 \tag{10}$$

最终模型优化目标为最小化损失函数：

$$L_{\text{sentence\_level}}=\mathrm{E}\left(MSE\left(HTER,output\right)\right) \tag{11}$$

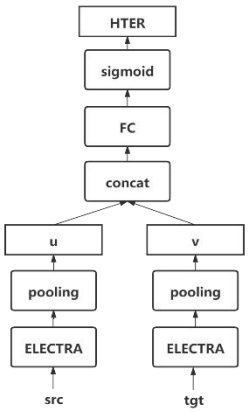

图 2　本文的翻译质量估计模型结构

其中，整个模型框架最底部的预训练语言模型可以自由替换为其他各种单语预训练语言模型，只需按照不同预训练语言模型的要求在文本输入阶段做对应不同的数据预处理即可。因此，本文提出的翻译质量估计模型具有一定的通用性。

3　基于 ELECTRA 的翻译质量估计伪数据生成方法

3.1 方法介绍

本文提出了如图 3 所示的基于 ELECTRA 生成器的句子级翻译质量估计伪数据生成方法。下面以生成一组 QE 伪数据为例，结合示意图介绍本方法的具体操作流程。

首先向 ELECTRA 的生成器输入一行句子，生成器会选择句子中的词进行预测并使用预测到的词替换原词。在这一环节本文取消了对输入句子的 mask 操作，因此生成器会对句子中每个词都预测并替换一遍，而不是只对部分词预测替换。上述过程的形式化表示如下，一段句子 $p=[p_1,...p_n]$，经过已经训练好的生成器 G 编码得到 $h_G(p)=[h_1^G,...h_n^G]$，对于 t 位置预测得到所有词的概率：

$$p_G\left(p_t|p\right)=\mathrm{softmax}\left(\exp\left(p_t\right)\cdot h_G(p)_t\right) \tag{12}$$

最后选择概率最大的词 $\hat{p}$ 替换为该位置的词：

$$\begin{aligned}\hat{p}&=\arg\max_{p_t}\ p_G\left(p_t\mid p\right)\\&=\arg\max_{p_t}\ \mathrm{softmax}\left(\exp\left(p_t\right)\cdot h_G(p)_t\right)\end{aligned} \tag{13}$$

因此经过 ELECTRA 生成器处理后输出的句子与原句相比，可看成带有部分翻译错误的新生成的机器译文。接着使用通过 ELECTRA 生成器改写过的新机器译文与人工后编辑译文通过 TERCOM 工具包计算出对应的 HTER 分数。最终源文、经生成器改写的新机器译文、人工后编辑译文以及重新计算得到的新 HTER 分数这四部分构成了一条新的句子级翻译质量估计伪数据四元组。

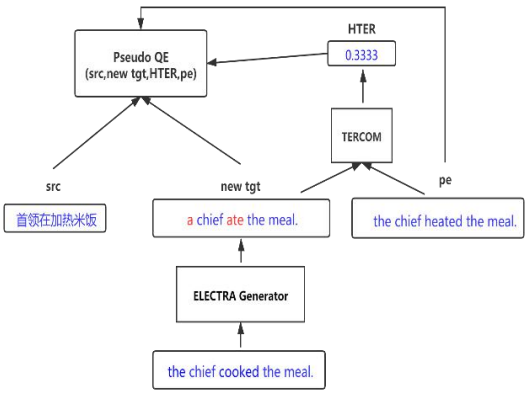

图 3　翻译质量估计伪数据生成方法示意图

3.2 不同来源的伪数据生成策略

本文分别使用待扩充的 QE 数据集中目标语言方向的人工后编辑译文和机器翻译译文作为伪数据生成的母本，根据本文提出的方法进行了伪数据生成操作，得到了两组数据分布特征不同的翻译质量估计任务伪数据集。

为了更好的了解生成的不同两种伪数据集的数据分布特点，本文以 CCMT2019 EN-ZH 数据集为例使用 TERCOM 对生成的数据集进行了错误统计。统计结果如表 1 所示。从整体的错误统计结果来看，由机器译文作为输入母本生成的伪数据集在错误分布上与原 QE 数据集基本一致，且数据集的总体错误比例相近。而由人工后编辑译文作为输入母本生成的伪数据集与原 QE 数据集相比，错误分布差异较大，替换错误占据大部分比例，且数据集总体错误比例偏小。

4　实验

4.1 数据集

表 1　基于不同输入来源生成的翻译质量估计伪数据错误统计

数据集	插入错误	删除错误	替换错误	调序错误	总计	平均 TER
原始数据集	18887（16.3%）	36268（31.3%）	49110（42.3%）	11706（10.1%）	115971	15.631%
人工后编辑译文生成的伪数据集	2461（7.9%）	4465（14.3%）	24227（77.6%）	70（0.2%）	31223	4.208%
机器译文生成的伪数据集	18764（13.3%）	41225（29.3%）	69245（49.2%）	11600（8.2%）	140834	18.982%

本节提出的翻译质量估计模型分为预训练阶段和翻译质量估计阶段，所以需要使用两种不同的数据集。预训练语料上，本文选取了中-英、英-德两个方向上与翻译质量估计任务尽可能相似的平行语料。其中中-英平行语料来自 CCMT2020 中英新闻领域机器翻译任务和 CCMT2020 翻译质量估计任务源语言句子与人工后编辑译文组成的平行句对，而英-德平行语料来自 WMT17 英德新闻机器翻译任务、WMT2017 翻译质量估计任务源语言句子与人工后编辑译文组成的平行句对、TED2020 v1 数据集[12]以及 QED[13]数据集。另外，本文对收集到的平行语料进行如下操作，过滤掉句子长度大于 70 的平行句对并控制双语句对之间的长度比例位于 1/3 至 3 之间，以确保平行语料的质量，最终得到的预训练阶段平行语料规模如表 2 所示。

表 2　预训练阶段平行语料规模

语言方向	平行语料对数
中-英	9023453
英-德	8702358

翻译质量估计数据集上，本文分别选取了 CCMT2019 中↔英 (ZH-EN) 方向以及 WMT2017 英↔德 (EN-DE) 方向的翻译质量估计相关的句子级数据集，具体数据集规模如表 3 所示。

表 3　翻译质量估计数据集规模

数据集	训练集	开发集	测试集
WMT2017 EN-DE	23000	1000	2000
WMT2017 DE-EN	25000	1000	2000
CCMT2019 ZH-EN	10070	1143	1385
CCMT2019 EN-ZH	14789	1381	1445

4.2 模型训练方式及细节

预训练阶段：预训练模型在大规模无监督语料中学习到了通用的知识表示，但是由于训练语料多为通用文本，且长文本较多，与翻译质量估计数据集之间的数据分布存在着一定差距。因此，为了减缓预训练阶段的通用训练数据与下游翻译质量估计任务数据的差异性，我们在预训练阶段采用了在公开的已训练好的 ELECTRA 预训练模型基础上使用与翻译质量估计数据分布近似的语料继续预训练。具体的操作为，将收集到的与翻译质量估计任务相似的平行语料拆分成两份单语语料用于训练两个不同语言方向的单语 ELECTRA 模型。

翻译质量估计阶段：本模型在翻译质量估计阶段，为了使预训练语言模型更好的适应翻译质量估计任务，不再固定作为预测器的 ELECTRA 模型的参数，使其继续参与翻译质量估计阶段的训练，ELECTRA 采取微调的方式和估计器联合训练，共同进行模型参数的更新。

伪数据参与训练：在具体的结合伪数据的模型训练策略上，本文采取先使用人工后编辑译文生成的伪数据对模型进行初次训练再使用机器译文生成的伪数据与原数据混合后的数据集进行二次微调的策略。

4.3 实验参数设置

实验整体在基于 Pytorch 的环境中进行。预训练阶段，首先从 Hugging face[14]上加载开源的经过通用语料训练得到的英、中、德三个语言版本的 ELECTRA 预训练语言模型及对应的词表文件，具体模型规模为 12 层 Transformer-encoder、隐层大小 768 及 12 个注意力头。接下来使用准备好的语料进行二次预训练，采用 AdamW 作为优化器，sequence_length 统一设置为 128，batch size 设置为 32。在采用衰减的学习率设置，初始学习率设置为 2e-4 的情况下，使用 4 张 2080Ti 显卡进行训练直至收敛需要花费 6 天的时间。

翻译质量估计阶段，预训练模型继续参与训练，本文仍选取 AdamW 作为优化器，在 batch size 设置为 16，学习率设置为 5e-5 的条件下，使用 2 张 2080Ti 训练不到 1 小时模型即可完成训练。

4.4 实验评价指标

在句子级翻译质量估计任务中，使用斯皮尔曼相关系数（Spearman's Rank Correlation Coefficient, Spearman）、皮尔森相关系数（Pearson Correlation Coefficient, Pearson）、平均绝对误差（Mean Absolute Error, MAE）以及均方根误差（Root Mean Squared Error, RMSE）这四个指标来衡量一个翻译质量估计模型性能的好坏。斯皮尔曼相关系数及皮尔森相关系数作为主要衡量指标，两者值越高说明该翻译质量估计模型性能越好，平均绝对误差及均方根误差作为参考指标，两者值越低模型性能越好。

4.5 实验结果及分析

本文选取 WMT2017 EN-DE、WMT2017 DE-EN、CCMT2019 EN-ZH 及 CCMT2019 ZH-EN 四个数据集进行了相关实验，并选取 Bilingual Expert 模型作为基线模型进行对比，该模型在 WMT2017、WMT2018 翻译质量估计任务的各项子任务中均取得最佳效果。由于 Bilingual Expert 的论文中未公布在中↔英方向数据集下的模型结果，因此本文依据论文提供的代码使用相同预训练语料进行复现，并将复现的模型结果作为基线。实验结果分别如表 4、5 所示。

表 4　WMT2017 句子级翻译质量估计任务实验结果

	Models	Pearson ↑	Spearman ↑	MAE ↓	RMSE ↓
EN-DE	Bilingual Expert paper result	0.684	0.709	**0.100**	0.144
	Our Model with ELECTRA pretrained	0.689	0.72	0.109	**0.141**
	Our Model with ELECTRA pretrained+pseudo data	**0.705**	**0.736**	0.112	0.147
DE-EN	Bilingual Expert Paper Result	0.710	0.642	**0.093**	**0.139**
	Our Model with ELECTRA pretrained	0.731	0.65	0.117	0.149
	Our Model with ELECTRA pretrained+pseudo data	**0.737**	**0.663**	0.115	0.142

表 5　CCMT2019 句子级翻译质量估计任务实验结果

	Models	Pearson ↑	Spearman ↑	MAE ↓	RMSE ↓
EN-ZH	Bilingual Expert paper result	0.296	0.215	0.116	0.152
	Our Model with ELECTRA pretrained	0.385	0.353	**0.101**	**0.135**
	Our Model with ELECTRA pretrained+pseudo data	**0.411**	**0.366**	0.107	0.143
ZH-EN	Bilingual Expert Paper Result	0.441	0.453	0.104	0.136
	Our Model with ELECTRA pretrained	0.508	0.511	**0.0892**	**0.132**
	Our Model with ELECTRA pretrained+pseudo data	**0.52**	**0.524**	0.0906	0.135

实验结果表明，本文提出的基于 ELECTRA 的翻译质量估计模型在实验选取的四个数据集中主要评价指标的结果上均优于 Bilingual Expert 基线模型，并且当使用本文提出的伪数据生成方法构造的 QE 伪数据参与训练后，模型的性能进一步得到提升。

另外由于本文提出的模型具有通用性，兼容大多数单语预训练模型，本文选取模型大小规模相同的 BERT、ALBERT[15]、RoBERTa[16]预训练模型作为翻译质量估计模型的预训练部分，使用相同的预训练语料在 WMT2017 EN-DE 数据集进行了实验。实验结果如表 6 所示，实验表明在本文提出的翻译质量估计模型中，使用 ELECTRA 预训练模型作为模型预训练部分相较于使用其他单语预训练模型能获得更好的结果。

本文提出的基于 ELECTRA 的翻译质量估计模型作为类似“预测器-估计器”的两阶段式模型，与先前模型相比进行了如下改变：（1）使用同一时刻能够双向表示的预训练语言模型代替原有 Transfomer、RNN 等同一时刻只能单向语义表示的模型。（2）在翻译质量估计阶段，不再固定预测器的模型参数，而是采取微调的形式使预测器和估计器同时进行模型参数更新。（3）在预训练模型的训练方式上，本文采取了在已训练好的 ELECTRA 模型上使用与下游翻译质量估计任务分布相似的语料二次预训练的方法。（4）针对二次预训练所需单语语料，采取了使用平行语料拆分得到两份不同语言方向的单语语料的方法。为了证明提出的以上改动均具备有效性，本文在 CCMT2019 EN-ZH 句子级 QE 数据集上做了如表 7 所示的消融实验。

表 6　不同单语预训练模型作为模型预训练部分的实验结果

	Models	Pearson ↑	Spearman ↑	MAE ↓	RMSE ↓
EN-DE	Our Model with BERT pretrained	0.654	0.695	0.114	0.148
	Our Model with ALBERT pretrained	0.656	0.699	0.112	0.146
	Our Model with RoBERTa pretrained	0.669	0.709	0.118	0.148
	Our Model with ELECTRA pretrained	**0.689**	**0.72**	**0.109**	**0.141**
DE-EN	Our Model with BERT pretrained	0.704	0.626	**0.107**	**0.145**
	Our Model with ALBERT pretrained	0.702	0.631	0.110	0.146
	Our Model with RoBERTa pretrained	0.716	0.641	0.110	0.148
	Our Model with ELECTRA pretrained	**0.731**	**0.65**	0.117	0.149

其中，模型（1）在预训练阶段使用 GPT-2[17] 代替 ELECTRA 参与训练。GPT-2 是 OpenAI 发布的一个基于 Transformer 解码器的自回归预训练语言模型。作为单向语言模型，GPT-2 不具有双向语义表示的能力。本文利用 Hugging face 加载相同模型规模的预训练好的中文及英文 GPT-2 得到了模型（1）的实验结果。通过对比模型（1）和模型（5），有效证明了使用双向预训练语言模型相比单向预训练语言模型在翻译质量估计任务上的必要性。通过对比模型（2）和模型（5），充分证明了使用与翻译质量估计任务相似的语料对预训练语言模型二次预训练有助于模型效果的提升。模型（3）为使用不相关的中文和英文单语语料去分别训练两个语言方向的 ELECTRA 模型得到的结果。具体操作为，中文 ELECTRA 模型使用来自本文收集的中-英平行语料的中文部分进行二次预训练，而英文 ELECTRA 模型则使用英-德平行语料的英文部分进行二次预训练。通过比较模型（3）和模型（5），可以看到将平行语料拆分成单语语料进行二次预训练的做法的确会提升模型的部分性能。最后，对比模型（4）和模型（5），证明了翻译质量估计阶段预测器与估计器共同更新参数的有效性。

表 7 消融实验结果

Model	Pearson ↑	Spearman ↑
（1）Our model with GPT-2 (second pre-training by multi-lingual, unfixed)	0.328	0.307
（2）Our model with ELECTRA (no second pre-training, unfixed)	0.336	0.314
（3）Our model with ELECTRA(second pre-training by monolingual, unfixed)	0.379	0.349
（4）Our model with ELECTRA(second pre-training by multi-lingual, fixed)	0.352	0.338
（5）Our model with ELECTRA(second pre-training by multi-lingual, unfixed)	**0.385**	**0.353**

4　结论

本文围绕翻译质量估计任务，从模型和数据两个方面进行了改进。模型方面提出了基于 ELECTRA 的翻译质量估计模型，该模型作为一种两阶段式模型分为预训练阶段和翻译质量估计阶段。与基于“预测器-估计器”的基线模型相比，本模型引入能够双向表示的 ELECTRA 预训练语言模型作为预测器，使模型在预训练阶段学习到更好的语义表示应用到翻译质量估计阶段，且预训练模型继续参与翻译质量估计阶段的训练，采取微调的方式联合估计器共同进行模型参数更新。数据方面提出了基于 ELECTRA 的翻译质量估计伪数据生成方法。本方法通过 ELECTRA 生成器对输入句子中部分词进行替换改写的方式构造了翻译质量估计伪数据集，从而扩大了数据集的规模。实验表结果明，本文提出的两点改进均有效提升了翻译质量估计模型的性能。

参考文献：

[1] Kreutzer J, Schamoni S, Riezler S. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015: 316-322.

[2] Martins A F T, Astudillo R, Hokamp C, et al. Unbabel’s participation in the wmt16 word-level translation quality estimation shared task[C]//Proceedings of the First

Conference on Machine Translation: Volume 2, Shared Task Papers. 2016: 806-811.

[3] Kim H, Lee J H. A recurrent neural network approach for estimating the quality of machine translation output[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Linguistics:Human Language Techologies. StroudsBurg, PA: ACL ,2016:494-498.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[5] Fan K, Wang J, Li B, et al. “Bilingual Expert” Can Find Translation Errors[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6367-6374.

[6] Mikhail Mosyagin, Varvara Logacheva. MIPT System for World-Level Quality Estimation[C]// Proceedings of the Fourth Conference on Machine Translation .Italy: 2019:92–96.

[7] Kim H, Lim J H, Kim H K, et al. QE BERT: bilingual BERT using multi-task learning for neural quality estimation[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019: 85-89.

[8] Conneau A, Lample G. Cross-lingual language model pretraining[J]. 2019.

[9] Kepler F, Trénous J, Treviso M, et al. Unbabel’s Participation in the WMT19 Translation Quality Estimation Shared Task[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019: 78-84.

[10] Yankovskaya E, Tättar A, Fishel M. Quality estimation and translation metrics via pre-trained word and sentence embeddings[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019: 101-105.

[11] Clark K, Luong M T, Le Q V, et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators[C]//International Conference on Learning Representations. 2019.

[12] Reimers N, Gurevych I. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4512-4525.

[13] Abdelali A, Guzman F, Sajjad H, et al. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain[C]//LREC. 2014, 14: 1044-1054.

[14] Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art natural language processing[J]. arXiv preprint arXiv:1910.03771, 2019.

[15] Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[C]//International Conference on Learning Representations. 2019.

[16] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019.

[17] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

ELECTRA Based MT Quality Estimation Model

Meng Qingye, Yang Muyun*, Li Yueyang, Zhao Tiejun, Zhu Conghui, Cao Hailong

（School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150006, China)

Abstract: This paper proposed an Electra based MT quality estimation method. In the aspect of model, on the basis of "Predictor-Estimator" structure, a pre-trained language model based on bidirectional semantic representation is applied to replace the original predictor via ELECTRA. In the quality estimation stage, the predictor is updated together with the estimator parameters in a fine-tuning method. In the aspect of data, ELECTRA generator is used to replace and rewrite some words in the input sentence to construct pseudo data as an enhancement to the training data. Experiments show that the proposed method improves the performance of the quality estimation model, out-performing the Bilingual Expert baseline model on several datasets.

Keywords: quality estimation; pre-trained language models; ELECTRA; pseudo data