LAIX CCMT 2022 翻译质量评估任务测评报告

余勇宏[1]，王永杰[1]，王川[1]，李若冰[1*]，林晖[1,2]

（1.上海流利说信息技术有限公司，上海市，200090；2.上海市学习与认知人工智能重点实验室，上海市，200090）

摘要：本报告介绍了我们在 2022 年第十八届全国机器翻译大会（CCMT 2022）机器翻译评测翻译质量评估（Quality Estimation，QE）任务中的中-英和英-中两个赛道中所采用的主要方法。本方法基于现有主流的预测器-评估器双阶段模型的框架，其中预测器我们采用了大规模中英双语平行语料和中英文语法纠错（Grammatical Error Correction，GEC）数据合成的伪 QE 数据对 Cross-lingual Language Model（XLM）[1]、XLM-RoBERTa（XLM-R）[2]、XLM-RoBERTa-Large（XLM-R-LARGE）[2]和 Multilingual BERT（mBERT）[3]等预训练语言模型进行继续训练，并对原文句子、译文句子以及参考翻译进行特征提取，评估器通过输入的上述特征对 HTER（Human-targeted Translation Edit Rate）值进行回归预测。在实验过程中，通过引入 GEC 数据合成的伪 QE 数据，我们在中-英和英-中两个赛道的 QE 任务预测效果都有显著提升。在最终的 CCMT 2022 QE 离线测试结果中，我们的集成系统在中-英赛道排在第二，英-中赛道排在第四。

关键词：语法纠错；翻译质量评估；预训练语言模型

中图分类号：TP302.1　　**文献标识码：**A

0 前言

CCMT 2022 机器翻译评测的离线测评任务包含句子级 QE 任务的中-英和英-中两个赛道。该任务目标为在无参考译文的条件下，预测出译文的翻译质量指标 HTER，并以此近似评估译文质量。本文详细介绍了流利说（LAIX）算法团队在本测评任务中使用的数据处理策略、模型结构、训练流程以及参赛模型在 QE 的中-英和英-中两个语种赛道上的性能表现。因为我们在此次任务中创新地引入了可公开获取的中英文 GEC 数据，该数据目前并未纳入官方大纲指定的数据，所以此次我们参与的任务属性为非受限任务。

1 测评系统

1.1 模型结构

本系统使用了目前 QE 领域主流的预测器-评估器框架[4]，其中预测器用于提取样本特征，评估器根据提取的特征对样本 HTER 值进行回归预测。本文中，我们主要聚焦于句子级别的 QE 任务，但是因为单词级别的学习目标可以为 QE 模型带来有益的细粒度特征，因此我们联合学习了句子级和单词级的标签数据。句子级和单词级的损失函数定义如下：

$$L_{word} = \sum_{s \in D} \sum_{x \in s} -(y_{ok} \log p_{ok} + \lambda y_{bad} \log p_{bad})$$

* 通讯作者：ruobing.li@liulishuo.com

$$L_{sent} = \sum_{s \in D} \| hter_p - hter_l \|$$

s 和 x 分别表示训练数据 D 中的句子和单词，y 是单词级 *OK* 和 *BAD* 标签数据，p 是对单词级的预测变量，$hter_p$ 是预测的句子级 HTER 值，$hter_l$ 是句子级 HTER 标签，λ 是单词级 *BAD* 预测项损失函数的权重因子。句子级和单词级的联合损失函数定义如下：

$$L_{joint} = \sum_{s \in D} (L_{sent} + \eta \sum_{x \in s} L_{word})$$

μ 因子用来平衡句子级和单词级的损失函数项。

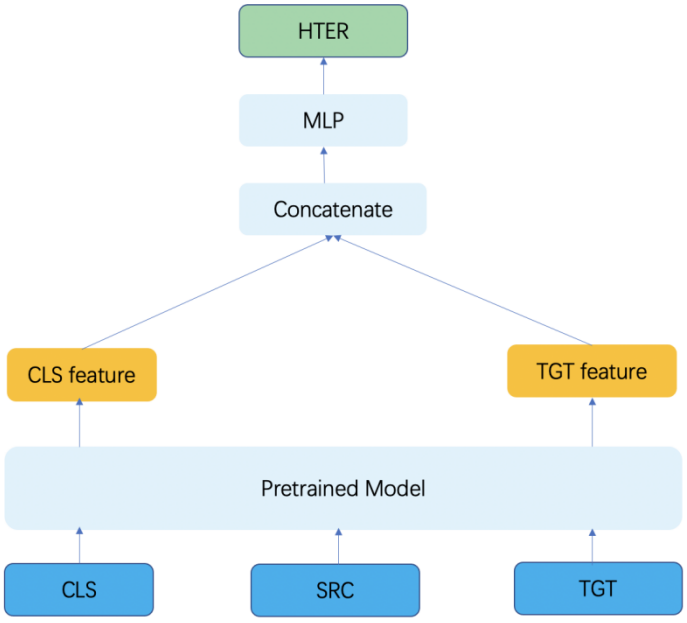

图 1 QE 模型结构 1，拼接原文句子和译文句子作为输入

Fig.1 QE model structure 1, concatenate the original sentence and the translation sentence as input

本评测系统采用了 2 种模型结构，模型结构 1 从结构上较模型结构 2 简单一些。如图 1 所示，模型结构 1 拼接原文句子和译文句子后输入预训练语言模型，获得对应的 CLS 位置输出作为 CLS 特征，将原文句子中每个 token 在所有 Transformer 的 block 层的输出 embedding 通过注意力池化（Attention Pooling）操作得到 token 特征，并将译文句子中所有的 token 特征通过平均池化（Average Pooling）操作得到译文句子特征 TGT feature。随后，我们将 CLS feature 和 TGT feature 的拼接结果输入评估器对 HTER 值进行回归预测。在模型训练阶段，我们对 HTER 值、译文词汇和译文词汇 GAP 这三个目标进行联合学习（Joint Learning）。在模型结构 1 中，预测器中的预训练语言模型 Pretrained Model 使用了 XLM。

如图 2 所示，模型结构 2 拼接了原文句子、译文句子和参考翻译作为预训练语言模型的输入。其中，参考翻译由官方大纲指定的中英双语平行语料训练所得机器翻译系统对原文句子进行翻译得到。该做法参考了[5]的相关工作，但使用了不同的预测器特征种类。首先，我们通过预训练语言模型获得 CLS 特征、原文句子 token 特征、译文句子 token 特征以及参考翻译 token 特征，获得 token 特征的方式同模型结构 1；其次，我们将句子中的 token 特征通过平均池化操作获得原文句子特征 SRC feature、译文句子特征 TGT feature 和参考翻译特征 REF feature；随后，我们将 SRC feature 和 TGT feature 表征间的差取绝对值，作为原文句子和译文句子的组合特征，并通过同样的方式得到参考翻译和译文句子的组合特征；最后，我们将这两种组合特征通过最大池化（Max Pooling）操作得

到全局组合特征 Combination feature，并将 CLS feature 和 Combination feature 拼接后输入评估器对 HTER 值进行回归预测。类似模型结构 1，在训练阶段，我们对 HTER 值、译文词汇和译文词汇 GAP 进行联合学习。在模型结构 2 中，预测器中的预训练语言模型 Pretrained Model 我们分别尝试了 mBERT、XLM-R 和 XLM-R-LARGE。

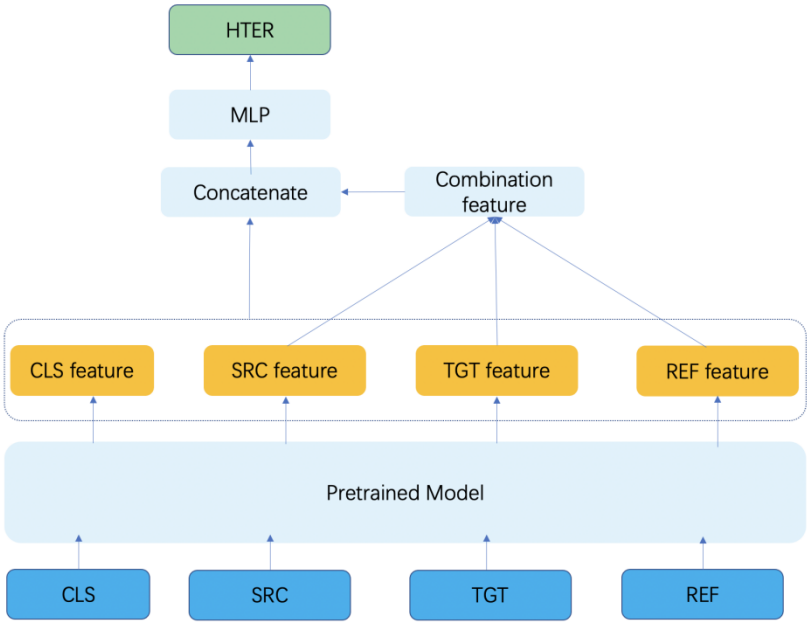

图 2 QE 模型结构 2，拼接原文句子、译文句子和参考翻译作为输入

Fig.2 QE model structure 2, concatenate the original sentence, translation sentence and reference translation as input

1.2 模型集成策略

模型集成可以充分利用多种不同模型的多样性来降低预测时的扰动，从而提升预测的效果，因此我们两个赛道的提交系统均采用了模型集成的策略。

在精调阶段，我们将训练集做了 5 折的划分。对每种 QE 预训练模型，将每 4 折数据进行合并，并在合并数据上精调模型，同时以剩余 1 折数据作为开发集，用来选取最终单模型的 checkpoint。

在预测阶段，所有训练得到的模型分别对某个测试样本进行预测，并将每个模型在该样本上预测得到的 HTER 值求平均作为该样本最终的 HTER 值。

特别地，为了筛选出用来集成的效果较好的单模型，我们以 CCMT 2019 测试集作为模型集成阶段的开发集，过滤掉效果较差、对最终集成效果没有帮助或起到负向作用的单模型。

2 数据

表 1 CCMT 2019 QE 数据描述

Tab.1 CCMT 2019 QE data description

数据划分	中-英（句对）	英-中（句对）
训练集	10070	14789
验证集	1143	1381
测试集	1385	1445

本次测评任务包括句子级翻译质量评估的中-英和英-中两个赛道，需要通过输入一对原文和译文来预测 HTER 值。由于 CCMT 2021 测试集标签数据尚未公布，我们使用了 CCMT 2019 的测试集进行效果验证。两个赛道的 QE 数据集基本统计结果如表 1 所示。

为了使模型具备更多的翻译知识，本系统使用了 CCMT 2022 中英双语平行语料来进一步预训练已有的预训练语言模型。

同时，我们观察到 QE 数据中的译文句子存在大量类似语法错误的数据，因此我们创新地引入了中英文 GEC 数据来合成量级较大、质量较高的伪 QE 数据。具体来说，我们将 GEC 的原文句子（即含有语法错误的句子）作为 QE 数据的译文句子，将 GEC 的目标句子（即进行语法纠错之后的正确句子）进行机器翻译后得到的翻译结果作为 QE 数据的原文句子，然后将这样生成的句子对作为合成的伪 QE 数据。因为目前机器翻译模型在解码器端文法生成方面的优良性，我们不需要过多担心机器翻译后的原文句子引入语法错误数据的影响。在模型进行真实的 QE 数据精调训练之前，我们生成的伪 QE 数据将用来对我们的模型进行预训练。GEC 数据的基本统计结果如表 2 所示。

表 2 GEC 数据描述

Tab.2 GEC data description

语言	数据集	句子数量
英文	FCE[6]	29015
英文	Lang-8[6]	969586
英文	NUCLE[6]	57151
英文	W&I+LOCNESS[6]	34304
中文	NLPCC 2018[1]	717241

3 实验

3.1 配置

本系统共使用了 4 种预训练语言模型，其中每种预训练语言模型对应的 QE 模型结构和训练流程如表 3 所示。在训练流程的符号表示中，“L”表示使用中英双语平行语料预训练预测器的预训练语言模型，“G”表示使用由 GEC 数据合成的伪 QE 数据来预训练 QE 模型，“F”表示使用 QE 数据精调 QE 模型。

表 3 每种预训练模型对应的 QE 模型结构和训练流程

Tab.3 The QE model structure and training pipeline corresponding to each pre-trained model

预测器中的预训练模型	QE 模型结构	训练流程
XLM	1	G->F
XLM-R	2	L->G->F
XLM-R-LARGE	2	G->F
mBERT	2	G->F

在使用中英双语平行语料预训练预测器的预训练语言模型时，我们将 CCMT 2022 中英双语平行语料中的原文句子和译文句子进行拼接，对拼接后的文本进行随机掩码，掩码比率设为 15%。我们采用和原预训练语言模型相同的掩码语言模型学习目标对其进行训练。训练 epoch 数设置为 20, 初始学习率为 5e-05, 由于显存容量限制，通过累积梯度更新的方法，将 batch size 设置为 8，每 16 个 batch 更新一次梯度。

[1] https://github.com/zhaoyyoo/NLPCC2018_GEC

QE 模型的预训练和精调均使用了 OpenKiwi 框架[7]。优化器使用 AdamW[8]，在默认参数的基础上，freeze step 数设置为 1 个 epoch，warmup step 数设置为 1.5 个 epoch，training step 数设置为 20 个 epoch，validation step 数设置为 0.5 个 epoch, early stop patience 设置为 10 或 15，预训练阶段学习率为 1e-05, 精调阶段学习率为 3e-06。因显存容量的限制，当预测器使用 XLM-R-LARGE 时，batch size 设置为 8，其他情况下 batch size 设置为 32，预训练阶段采用累积梯度更新的方法，每 8 个 batch 更新一次梯度。

在精调阶段，如 1.2 节模型集成策略部分所述，在 5 折数据划分后的每组数据上均进行模型精调，并将精调后的模型用于模型集成。

本系统在中-英赛道上优化了基于 XLM、XLM-R、XLM-R-LARGE 和 mBERT 的 QE 模型效果。由于时间关系，在英-中赛道上主要优化了基于 XLM 和 XLM-R-LARGE 的 QE 模型效果。

3.2 结果

本系统在英-中和中-英两个赛道的 CCMT 2019 测试集结果如表 4 所示。单模型效果表示本系统在 5 折数据划分上精调后的最优单模型效果，集成模型表示模型集成策略后的效果。

表 4 英-中、中-英两个赛道的 CCMT 2019 QE 测试集实验结果

Tab.4 Experimental results of CCMT 2019 QE test set on En-Zh and Zh-En tracks

赛道	模型	皮尔森相关系数
英-中	XLM	0.44539
	XLM-R-LARGE	0.51816
	集成模型	0.54659
中-英	XLM	0.57577
	XLM-R	0.58132
	XLM-R-LARGE	0.58760
	mBERT	0.56243
	集成模型	0.62693

特别地，我们验证了 GEC 数据对 QE 任务有显著的效果提升。如表 5 所示，我们选取了一组和 GEC 数据有关的消融实验结果。在中-英赛道上，我们对比 XLM-R 和 mBERT 在 CCMT 2019 开发集上不同流程的结果，加上 GEC 数据预训练的流程为 L->G->F，未加上 GEC 数据预训练的训练流程为 L->F。实验表明，GEC 数据预训练环节带来的效果提升是非常显著的。

表 5 GEC 数据在 XLM-R 和 mBERT 的消融实验结果

Tab.5 Ablation test results of GEC data on XLM-R and mBERT model

模型	皮尔森相关系数
XLM-R	0.50420
XLM-R + GEC 数据预训练	0.58236
mBERT	0.50316
mBERT + GEC 数据预训练	0.55839

CCMT 2022 QE 任务最终公布的离线测试结果中，同时包含了 CCMT 2021 和 CCMT 2022 QE 测试集的结果，以期作为阶段性的测试，给研究者带来模型提升有效策略验证的辅助信息。在没有复杂集成策略的基础上，我们的系统充分验证了通过 GEC 数据合成伪 QE 数据在 QE 任务上非常稳定的有效性。如表 6 所示，我们的系统在最近 2 年的 CCMT QE 中-英和英-中任务上，均为表现较为稳定且效果较好的系统之一。在本次评测任务中，官方以皮尔森相关系数（Pearson）为关键评测指标。

表 6 CCMT 2022 离线测试结果

Tab.6 CCMT 2022 offline test results

排名	CCMT 2021				CCMT 2022			
	英-中		中-英		英-中		中-英	
	队伍	Pearson	队伍	Pearson	队伍	Pearson	队伍	Pearson
1	HIT	0.4194	HIT	0.5405	HIT	0.4761	SUDA	0.5623
2	LAIX	0.4022	LAIX	0.5372	SUDA	0.4554	LAIX	0.5476
3	NJU	0.3977	SUDA	0.5156	NJU	0.4541	HNU	0.5279
4	SUDA	0.3834	HNU	0.5106	LAIX	0.4137	HIT	0.5216
5	HW	0.3500	HW	0.4823	HW	0.3704	HW	0.4850

4 总结

本文主要介绍了流利说算法团队在 CCMT 2022 机器翻译评测中参加翻译质量评估任务的情况。在该任务的实验中，我们尝试了多个不同的预训练语言模型，并首次将 GEC 数据引入了 QE 任务，同时提出了几种有效的基于 GEC 数据合成的伪 QE 数据预训练流程。我们在实验中发现 GEC 数据合成的伪 QE 数据在对 QE 模型进行预训练后，能显著提升模型的预测效果。在几种不同的基线预训练大模型基础上，伪 QE 数据增强的 QE 模型取得了一致且显著的提升效果，并在 CCMT 2021 和 CCMT 2022 QE 中-英和英-中赛道中均获得了较好的结果。我们希望该方法可以进一步促进 QE 领域数据增强方向的研究。

由于时间有限，我们在不同赛道的不同模型上，并没有尝试出统一的最优训练流程。同时，将 GEC 领域的合成数据引入本任务时，也没有得到比较正面的结果。对此，我们仍然在继续该方面的实验和探索，我们希望通过 GEC 数据增强 QE 数据的启发，来找到更有效的逼近真实 QE 数据的数据增强方法。通过引入量级更大的 GEC 合成数据，是否能进一步提升 QE 模型的效果？我们在后续会进行更多更细致的实验。

参考文献：

[1] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv:1901.07291, 2019.

[2] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[J]. arXiv preprint arXiv:1911.02116, 2019.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[4] 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 Second Conference on Machine Translation. 2017: 562-568.

[5] CHEN Y, SU C, ZHANG Y, et al. HW-TSC's Participation at WMT 2021 Quality Estimation Shared Task[C]//Proceedings of the Sixth Conference on Machine Translation. 2021: 890-896.

[6] CHRISTOPHER B, MARIANO F, ØISTEIN E A, et al. The BEA-2019 Shared Task on Grammatical Error Correction[C]//In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2019.
[7] KEPLER F, TRÉNOUS J, TREVISO M, et al. OpenKiwi: An open source framework for quality estimation[J]. arXiv preprint arXiv:1902.08646, 2019.
[8] LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in adam[J]. 2018.

The LAIX Systems in the CCMT 2022 Quality Estimation Shared Task

YU Yonghong[1], WANG Yongjie[1], WANG Chuan[1], LI Ruobing[1*], LIN Hui[1,2]

（1. LAIX Inc., Shanghai 200090, China; 2. Shanghai Key Laboratory of Artificial Intelligence in Learning and Cognitive Science, Shanghai 200090, China）

Abstract: This paper presents the main methods we adopted in the Chinese-English (Zh-En) and English-Chinese (En-Zh) tracks of the Quality Estimation (QE) shared task of CCMT 2022. Our methods are based on the mainstream predictor-estimator framework. We used the pseudo QE data synthesized from large-scale Chinese and English bilingual corpus and grammatical error correction (GEC) data to pre-train the Cross-lingual Language Model (XLM), XLM-RoBERTa (XLM-R), XLM-RoBERTa-Large (XLM-R-LARGE) and Multilingual BERT (mBERT). Features are then extracted from the original sentence, translation sentence and reference translation. The estimator uses the features to predict the Human-targeted Translation Edit Rate (HTER). Experimental results show that the pseudo QE data produced by GEC data can significantly improve the performance for both Zh-En and En-Zh QE tasks. In the final CCMT 2022 QE test, our ensemble system ranks second in the Zh-En track and fourth in the En-Zh track.

Key words: Grammatical Error Correction; Translation Quality Estimation; Pre-trained Language Model