HW-TSC CCMT 2021 翻译质量评估任务测评报告

陈一萌 [1]，苏 畅 [1]，陶仕敏 [1]，杨 浩 [1]，王明涵 [1]， 郭嘉鑫 [1]，张 敏 [1]，

杜纯宁 [1]，王宇侠 [1]，刘宇嘉 [1]

（1. 华为北京研究所研发能力中心，北京市，100095）

摘要：本报告介绍了在2021年CCMT大会机器翻译评测中进行翻译质量评估（Quality Estimation，QE）句子级译后编辑代价评估任务的方法。本方法基于预测器-评估器两阶段模型，其中预测器分别采用了在大规模平行语料上预训练得到的 Transformer、以及 XLM-RoBERTa，对任务中提供的源语言文本、机器翻译文本、以及调用翻译引擎接口得到的伪译后编辑文本进行特征提取；评估器采用全连接层，用提取出的特征对后编辑距离的分数进行回归。在实验过程中，通过引入构造伪后编辑的结果，在中-英、英-中两个方向的 QE 人工译后编辑距离分数预测效果都有明显提升。

中图分类号：TP 302.1　　**文献标志码：** A

0 引言

在第 17 届全国机器翻译大会（China Conference on Machine Translation，简称 CCMT）的离线技术测评任务包含句子级汉英、英汉机器翻译质量评估（Quality Estimation，简称 QE）任务，该任务目标为在无参考译文的条件下，对翻译质量参照译后编辑的结果对待评估译文预测 HTER（Human-targeted Translation Edit Rate）值，用以衡量译文质量。本文详细介绍了华为翻译中心（HW-TSC）在本测评任务中使用的数据处理策略、技术方法、模型结构、以及参赛模型在汉->英、英->汉两个方向质量评估任务上的性能表现。

1 测评系统

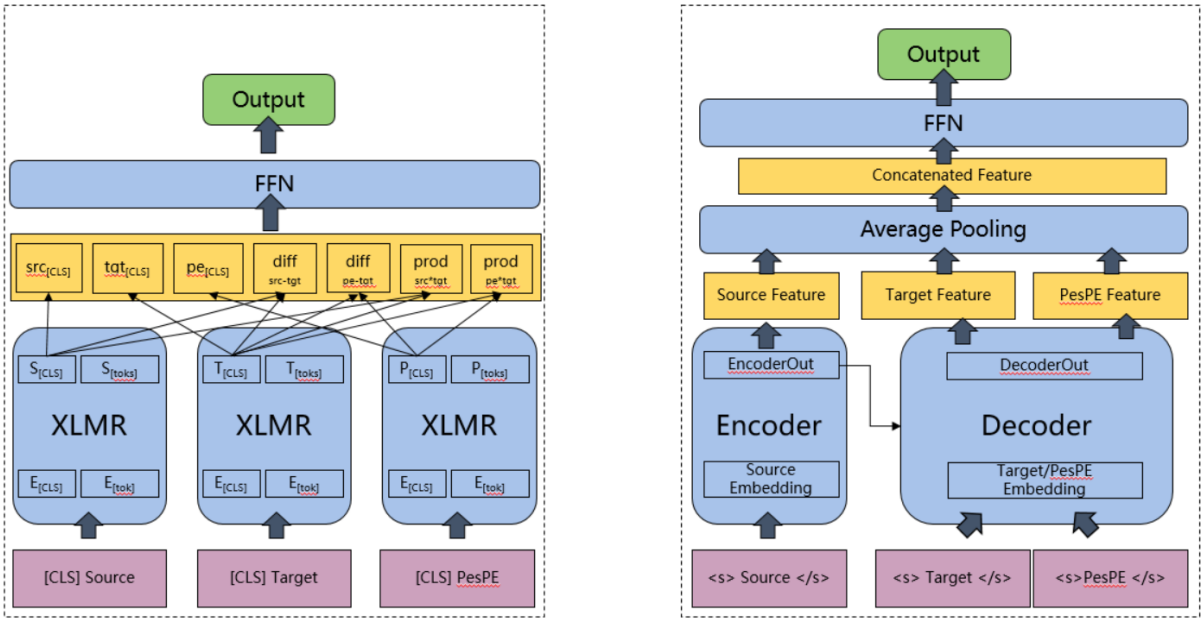
1.1 模型结构

为完成句子级质量评估任务，本系统使用了早期的研究[1]提出的预测器-评估器结构。在系统中使用了两种不同的模型结构：（1）使用语言模型 XLM-RoBERTa$_{Large}$(简称 XLMR)[2]作为预测器，（L=24, H=1024, A=15, 共 550M 参数），用于对原文、译文以及后编辑译文（或伪造后编辑译文）提取语言特征，参考[3]中，在评估器中，将原文句子表征、翻译句子表征、伪译后编辑句子表征、原文与译文句子表征间的差及点乘、伪译后编辑与译文句子表征间的差及点乘进行拼接，将拼接后的整体特征送入由两层全连接层构成的评估器，用于将特征映射到样本标记空间，对 HTER 分数进行回归预测；（2）基于 Transoformer 的神经机器翻译编码器-解码器（Encoder-Decoder，简称 Enc-Dnc）模型结构[4]（如图 1 左），其中编码器用于对源端语句提取特征，解码器用于分别对目标端语言文本及伪造的译后编辑语言进行特征提取（L_{enc}=35, H_{enc}=512, A_{enc}=8, L_{dec}=3, H_{dec}=512, A_{dec}=8,共 353M），之后，分别对（原文，译文，伪后编辑）的特征在句子维度进行平均池化操作，得到三个句子级表征，再由与（1）中相同的评估器对 HTER 的分数进行拟合。基于预测器-评估器结构的 QE 句子级 HTER 分数预测模型见图 1 所示。

1.2 模型集成策略

在最终提交的系统中，使用了模型融合的策略。模型的融合使用了以下几种策略: (1)在单次训练过程中在开发集上表现最好的若干个模型；(2)不同模型结构或通过随机 dropout 得到的若干个模型；

(3)使用装袋（Bagging）策略，将由训练集和开发集所组成的样本集合平均分为 4 份,得到{A_1, A_2, A_3, A_4}，用相同的模型结构重复地进行 4 次训练，在每一次训练中混合其中的 3 份样本集进行训练，剩余的一份数据集作为开发集进行训练，得到 4 个模型。最终，由 4 个模型分别独立地对句子的 HTER 分数进行预测，取每个句子的 4 个 HTER 分数的平均值作为最终的分数。



(a) 基于 XLMR 的 QE 模型结构　　(b) 基于 Enc-Dec 的 QE 模型结构

图 1 基于预测器-评估器结构的 QE 句子级 HTER 分数预测模型

Fig. 1 Predictor-Estimator Based QE Model for Estimating Sentence-Level HTER Score

2 数据

本次测评任务提供了英汉方向的 3043 句原文、14789 句译文以及对应的后编辑结果作为训练数据，2826 句原文、译文及对应的编辑结果作为开发集；汉英方向的 2503 句原文、10070 句译文及对应的编辑结果作为训练集、2528 句原文、译文及对应的编辑结果作为开发集。

除了测评任务中提供的数据外，本系统还使用了 CCMT2021 双语翻译任务中的英汉-汉英平行语料、WMT 2021 双语翻译任务中的英汉-汉英相关平行语料，对基于 Transformer 的神经机器翻译模型进行预训练。XLM-RoBERTa 则直接使用了 Huggingface 提供的开源模型，该模型在 2.5TB 经语料过滤的 CommonCrawl 数据上预训练得到。此外，为了进一步提升模型表现，本系统还使用到了伪造译后编辑句的数据增强方法[5-6]，该数据通过调用华为翻译引擎，对训练集、开发集和测试集中的原文进行翻译得到。

在实验过程中，所采用到的另外一种数据增强的方式是通过调用谷歌、百度、有道等若干翻译引擎分别对原文进行翻译，得到翻译结果作为伪造的译文句，生成额外的（原文，译文）二元组来对 QE 模型进行训练。然而，实验结果表明，通过引入该种数据增强方式，未能提升句子级 QE 分数预测的性能。

3 实验

3.1 实验设置

在本次的测评实验中，使用了 Facebook AI Research 开源的神经机器翻译程序 fairseq[7](https://github.com/facebookresearch/fairseq)，以及 Hugging Face 开源的 transformers 程序(https://github.com/huggingface/transformers)。其中，fairseq 主要用于处理英汉、汉英神经机器翻译

模型，即 Enc-Dec 评估器的训练，transformers 则是用于处理基于 XLM-RoBERTa 的领域适应微调以及在 QE 任务上的微调。通过使用在 1.2 小节描述的模型技术策略，使用装袋策略分别训练得到 4 个基于 XLMR 的模型及 4 个基于 Enc-Dec 的模型，通过计算共 8 个模型得到的 HTER 分数的平均分数作为最终的结果。在 QE 模型的训练中的批大小为 8。优化器使用 Adam[8]，学习率为使用 $1e^{-4}$ 进行第一次训练，训练过程中锁定 Enc-Dec 评估器和 XLM-RoBERTa 的学习参数，使用 $1e^{-5}$ 进行第二次训练，训练过程中打开 Enc-Dec 评估器和 XLM-RoBERTa 的训练学习参数。

3.2 实验结果

在本次测评任务中，测评指标使用自动评价的方式，主要评价标准为皮尔森相关系数（Pearson's correlation coefficient）。实验在开发集上的表现见表 1 所示。

表 1 英汉、汉英方向机器翻译质量评估任务的实验结果

语种方向	模型	皮尔森相关系数
英汉（Dev）	XLMR	0.5220
	XLMR+PesudoPE	0.5830
	Enc-Dec	0.4856
	Enc-Dec+PesudoPE	0.5718
	Ensemble（4XLMR+4Enc-Dec with PE）	0.6816
汉英（Dev）	XLMR	0.5247
	XLMR+PesudoPE	0.5893
	Enc-Dec	0.5017
	Enc-Dec+PesudoPE	0.5534
	Ensemble（4XLMR+4Enc-Dec with PE）	0.6291

注：本系统在英汉、汉英开发集上的实验结果

4 总结

本文介绍了华为翻译中心团队在第 17 届全国机器翻译大会中参加翻译质量评估评测任务的情况。在该任务的实验中，主要使用到了预训练的语言模型 XLM-RoBERTa、以及预训练的神经机器翻译模型作为预测器，对 QE 原文、译文、译后编辑句进行特征提取，评估器则是对原文、译文、译后编辑距离的句子特征在经过差、点乘操作后进行拼接，经全连接层对 HTER 分数进行回归拟合。在数据方面，本系统主要是使用华为机器翻译引擎对原文进行翻译得到伪译后编辑结果，作为数据增强的方法，实验证明通过引入伪译后编辑结果，对分数预测的表现有较大的提升。

但由于时间有限，本系统在实验过程中并未对评估器的模型结构进行更仔细的设计和试验，此外，笔者认为通过在有限的 QE 数据集上，如何能够更好的利用原文、译文、译后编辑结果进行数据增广，让生成的数据能够更加接近真实的 QE 数据，如研究[9]，是能够进一步提升 QE 结果的策略之

一，可以在后续的研究与实验中进行更加详细的实验。

参考文献：

[1] 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.

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

[3] REI R, STEWART C, FARINHA A C, et al. COMET: A neural framework for MT evaluation[J]. arXiv preprint arXiv:2009.09025, 2020.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
2017.

[5] WANG M H, YANG H, SHANG H C, et al. HW-WSC's participation at wmt 2020 quality estimation shared task[C]//Proceedings of the Fifth Conference on Machine Translation. 2020: 1056-1061.

[6] KEPLER F, TRENOUS J, TREVISO M, et al. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task[J]. arXiv preprint arXiv: 1907.10352, 2019.

[7] OTT M, EDUNOV S, BAEVSKI A, et al. fairseq: A fast, extensible toolkit for sequence modeling[J]. arXiv preprint arXiv: 1904.01038, 2019.

[8] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[9] CUI Q, HUANG S J, LI J, et al. DirectQE: Direct Pretraining for Machine Translation Quality Estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(14): 12719-12727.