基于多策略优化的译文质量评估

雷涛，叶恒，蒋宇龙，姜云卓，徐文斌，贡正仙*

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要：针对 CCMT 2022 中英译文质量评估任务（QE），本文探索了基于“预训练-评估器”模型的 QE 系统性能提升的多策略优化方法。本文以 Transquest 模型作为 QE 的基础系统架构。第一个策略是改进 QE 系统依赖的 XLM-R 预训练模型，使用 CCMT 2022 官方提供的中英平行语料，利用 TLM 掩码策略对 XLM-R 预训练模型进一步训练，增强 XLM-R 的双语关联能力。其次，在 QE 模型训练过程中，分别探索了在输入端扩展伪 PE 数据的策略和在 XLM-R 输出端融入依存句法特征的策略。实验表明，相对基准系统，这些策略都能不同程度地提升 QE 系统的预测能力。

关键词：译文质量评估；跨语言预训练模型；依存句法树；伪 PE

0 引言

机器翻译是指无需人工参与，仅通过机器就可以将一种自然语言自动转换成另一种自然语言的过程[1,2,3]，虽然随着神经网络模型的发展，机器翻译的模型越来越多，机器翻译的译文质量也越来越好，但从始至终都存在一个问题，即机器翻译的译文质量一直无法得到保证，所以需要对机器翻译的译文进行评估。机器翻译质量评估（Quality Estimation，QE）是指在给定原文以及机器翻译系统产生的译文时直接预测译文翻译质量的机器翻译评价方法，该方法并不依赖参考译文，其中句子级的质量标签采用的是 HTER[4]（Human-targeted Translation Edit Rate）。

随着神经网络的发展和在自然语言处理领域中的成功应用，在 QE 领域人们开始利用神经网络来提取句子对中存在的各种特征，比较著名的是 Kim[5]等人提出的“预测器-评估器”模型，随后 Fan 等人提出“双语专家”模型[6,7]，利用双向 Transformer 来代替 Kim 等人使用的循环神经网络作为预测器的基本结构，Kepler 等人也尝试使用预训练后的屏蔽语言模型（Masked Language Model，MLM）来作为预测器基本结构[8,9]。像 Kepler 他们这种基于知识迁移的机器翻译质量评估模型逐渐成为主流，目前在预测器结构中用的比较多的是一些大型

基金项目：国家自然科学基金 (61976148)

* **通讯作者：**zhxgong@suda.edu.cn

跨语言预训练模型，比方说 mBert[10]、XLM-Roberta[11]，本文使用的框架模型 Transquest [12]也是利用的 XLM-Roberta 模型，并且该框架模型在 WMT 2020 DA 任务中获得了第一名。但是使用这种基于知识迁移的机器翻译质量评估模型也存在一些局限性，比如说 XLM-Roberta 在训练时使用的是多语言的单语语料，双语关联能力较差，所以还需要进一步微调[11]。

本文针对 QE 系统依赖的跨语言预训练模型（本文指 XLM-R）存在双语关联能力较差的情况，利用掩码策略增强 XLM-R 的双语关联能力，并在此基础上融入了依存句法树，帮助 QE 模型学习一些额外的句法特征进行更有效的译文质量评估。此外针对跨语言预训练模型训练时使用的语料质量比较高，但是在进行质量评估的时候，需要评估的译文质量参差不齐的情况，本文使用通过谷歌翻译得到的伪 PE（Post-Edit）数据作为输入数据的一部分，用来保证 QE 系统评估时的相对一致性和稳定性。

1 技术路线

本文以 Transquest 模型作为参赛系统的基础模型。如图 1 大虚线框所示，它采用一个单独的 XLM-RoBERTa（XLM-R）预训练模型作为预测器，将原文和译文的拼接作为输入直接输入到 XLM-R 预训练模型，原文和译文之间用<sep>特殊标签分隔。Transquest 模型以一个 Softmax 层作为评估器，将预测器中 XLM-R 预训练模型的输出通过池化操作（本文采用[CLS]池化，即选用第一个子词<s>的隐层表示作为整个句子的特征向量）输入到评估器，用来预测 HTER 分数。

参赛系统以 Transquest 模型为基础，探索了如图 1 所示的三个部分的改进策略。第一个改进策略对应于图 1 所示的“Finetuned XLM-R”部分，根据 CCMT 2021 的技术报告[13]可知，采用掩码策略，使用双语平行语料进一步训练 XLM-R 预训练模型能够提升 XLM-R 模型的双语关联能力。第二个改进策略对应于“src 扩展”部分，核心思想是利用伪 PE 数据辅助 QE 模型学习。 Transquest 模型的输入是原文和译文的拼接，而改进的系统首先通过谷歌翻译引擎获得原文对应的谷歌翻译，然后将模型的输入扩展成原文、谷歌翻译和译文三者的顺序拼接。第三个改进在 XLM-R 的输出部分，系统利用图注意力网络 （Graph Attention Networks，GAT）来聚合依存句法的信息，将其与 XLM-R 输出的[CLS]表示相融合，然后再将它们传递给后面的评估模块。三个改进策略的详细描述请分别参见本文第 2 部分中的相应内容。

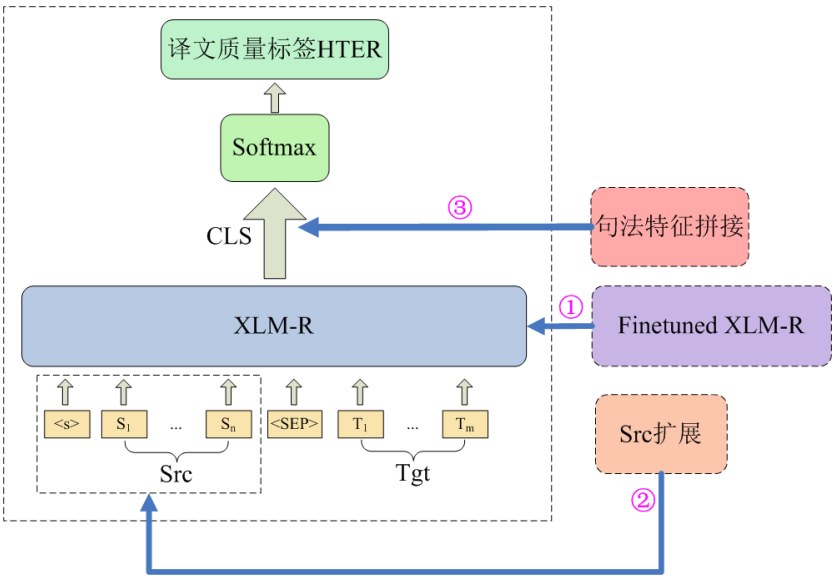

图 1 QE 模型多策略优化方案

Fig.1 QE system with Multi-strategy Optimization

2. 方法

2.1 预训练模型改进策略

XLM-R 虽然在跨语言任务上获得了相当好的结果，但是它仅在多语言的单语语料上进行训练，没有任何语言指示符，因此缺少双语关联能力。本文利用 CCMT 2022 所提供的中英平行语料对 XLM-R 进一步的预训练以增加它对双语关联的能力。具体方法采用如图 2 所示的翻译语言模型（Translation Language Modeling，TLM）[14]的训练过程，将拼接后的平行句对作为输入，然后在句对上随机选取 15%的 token，将选取的 token 中的 80%用特殊符号“<mask>”代替，10%用词表中任意一个词替换，剩下 10%不做任何处理，通过这样的方式提升 XLM-R 模型的双语关联能力[13]。

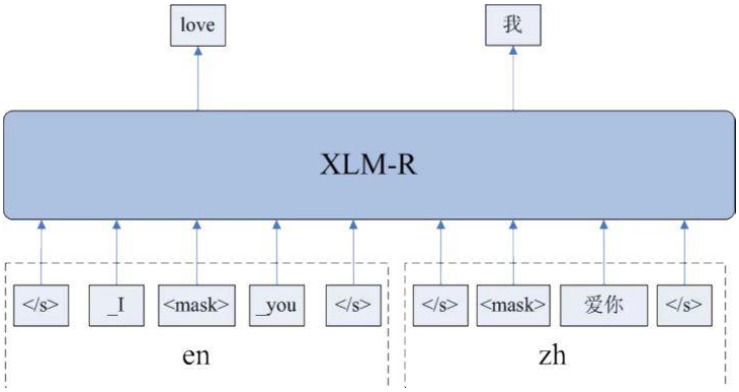

图 2 对 XLM-R 预训练采用 TLM 训练策略

Fig.2 TLM training strategy is used for XLM-R pre-training

2.2 伪 PE 辅助策略

在标准的翻译任务中，可以使用反向翻译等各种数据增强技术[15]来提升系统性能，但是由于 QE 任务需要额外公正且高质量的 PE（Post-Edit）数据来生成标签和 HTER 分数，所以 QE 系统很难通过一般的数据扩充方法来改善性能。华为在 WMT20 的工作[16]表明利用伪 PE 数据可以帮助提升系统性能，核心思想是质量较优的机器翻译文本可以近似当作 PE 数据（伪 PE），将这些伪 PE 数据参与编码与特征提取可以给评估器提供更有效的参考信息。该工作的伪 PE 数据通过 Google 翻译引擎翻译原文而获得， QE 模型采用了“预测器-评估器”结构，其中预测器是一个标准的 Transformer，因此可以将伪 PE 数据放置到 Transformer 的编码器和解码器的任意底层输入中然后再进行特征的提取。与该工作不同，本文的 QE 模型是“预训练-评估器”结构，预训练模型 XLM-R 可视作为 QE 的编码器，并无额外的解码器，所以仅将伪 PE 数据嵌入到编码器端，即先将原文使用 Google 翻译引擎进行翻译，再将得到的翻译数据作为伪 PE 数据，与原文拼接后替代系统中原始的原文输入部分。

此外，应用此策略的另一个原因如前文所述，基于“预训练-评估器”模型的 QE 模型所依赖的 XLM-R 采用的训练数据（单语数据）与训练 QE 模型的数据（双语数据，包括源端和目标端）不同，前者是在不同语言但完全正确的数据中训练获得，而后者的目标端包含错误程度不同的数据，为了减少预训练模型和 QE 模型在数据方面的差异性，通过引入质量较高的 Google 翻译结果作为伪 PE 数据进行 QE 模型的辅助训练，可以一定程度保证 QE 模型训练的收敛性和输出的一致性。

伪 QE 数据的具体使用情况如下图 3 所示，系统将原来 Transquest 模型中的 Src 部分进行扩展，不仅仅输入原文，还将谷歌翻译通过特殊标签</s>拼接到 Src 后面，希望模型以此学习一些额外特征，从而提升模型性能，其中原文用 S 表示，谷歌译文用 G 表示。

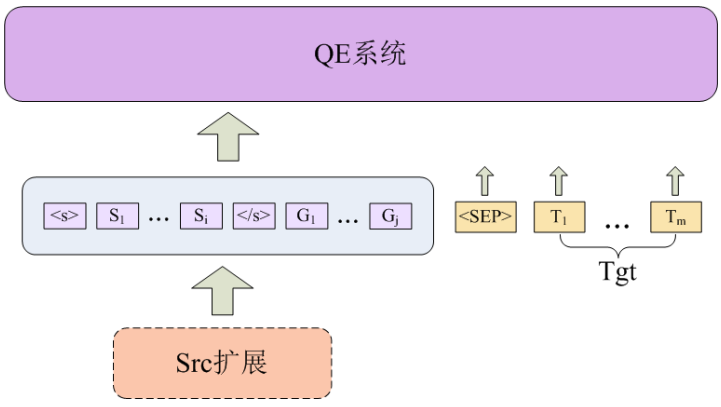

图 3 输入数据扩展策略

Fig.3 Extended source-side part with pseudo PE data

2.3 依存句法树特征表示的融合策略

本文提出了一种基于图注意力网络 （GAT[17]）的句法特征表示方法，通过与预训练模型的输出表示相融合，希望改进后的模型能学习到一些句法特征进行译文的质量评估。

本文利用依存句法分析器为每个句子构造依存句法树，并根据树的节点和边的关系生成依存句法图。使用 300 维的 GloVe[18]来表示图的节点嵌入信息。参考 XLM-R 把信息聚合到[CLS]嵌入的做法，本文使用 GAT 把图的节点信息聚合到与 ROOT 连接的节点上。如图 4 所示，本文在 Ni[19]的基础上，使用 Stanford NLP 和 DDParser 代替 dependency parsing [20]和 NLP 库 spaCy 直接获取句法依存树并转为 GAT 所需的图结构，得到 GAT 的聚合信息后，与 XLM-R 的[CLS]嵌入拼接，经过线性层得到最终结果。

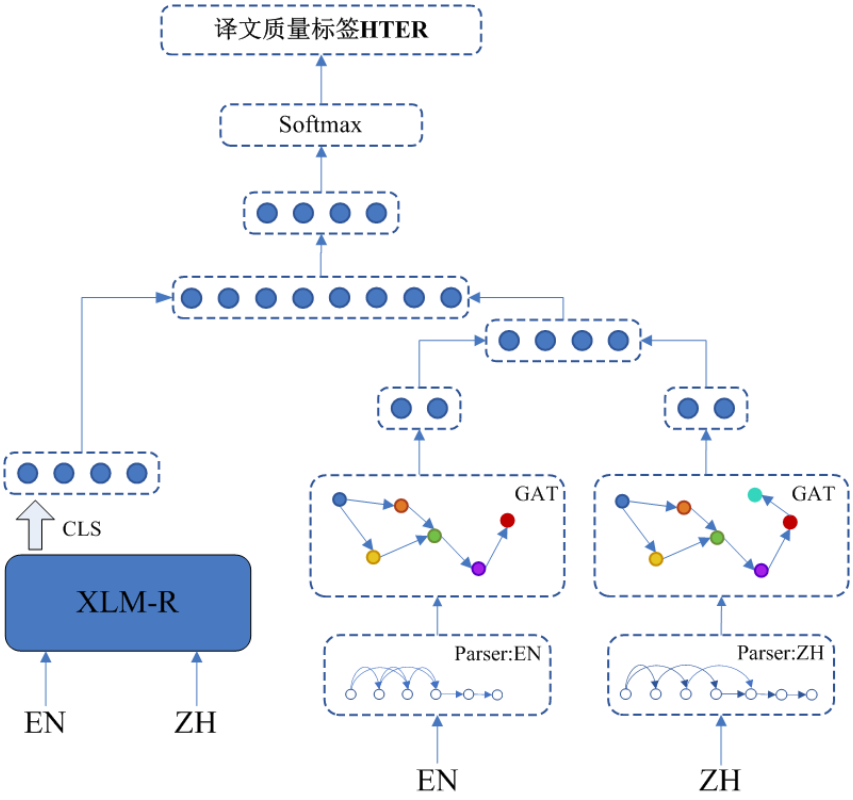

图 4 融合依存句法特征表示的 QE 模型

Fig.4 The improved QE model integrated with dependency parsing representation

图 4 中的 XLM-R 隐藏层有 768 维，GAT 聚合的节点信息英文输出 384 维，中文输出 384 维，把双语信息拼接起来，得到 768 维度。然后将 XLM-R 的[CLS]嵌入和双语信息拼接，经过线性层，得到最后的结果。

3.实验

3.1 数据

机器翻译质量评估的句子级任务的数据由三元组组成，分别是原文 S，译文 T 以及句子级别质量标记 H，记为<*S, T, H*>，另外在 CCMT 2022 的质量评估任务提供的数据集中，训练集和验证集还为每一个句子译文提供了对应的后编辑译文 E，本文给出了一个英-中方向的句子质量评估数据示例，如表 1 所示。

表 1 英-中方向的句子质量评估数据示例

Tab.1 Example of sentence quality assessment data in the English-Chinese direction

原文 S	You put all of the assets onto the TM repository.
译文 T	你把所有的商标上的资产存储库。
句子质量标签 H	0.7500
译后编辑 E	把所有语言资产上传到翻译记忆库.

3.2 实验设置

在本次测评任务中，本文使用了 CCMT2022 提供的数据进行训练，并在 19 年的验证集和测试集进行验证和测试。表 2 给出了 EN-ZH 和 ZH-EN 两个方向上译文质量评估的数据集。

表 2 实验数据统计

Tab.2 Statistical experimental data

语料	训练集	验证集	测试集
EN-ZH	14789	1381	1445
ZH-EN	10070	1143	1385

我们在官方提供的 913M 的英-中平行语料上采用掩码策略进一步训练了 XLM-RoBERTa-Base 模型，该模型包含 12 层编码器，隐层维度为 768 维，多头注意力机制设置 8 个头，使用 AdamW 作为优化器。

本文的 QE 模型使用了上述经过优化的预训练模型，其中 epoch 为 3，并只进行了单折交叉验证，批次大小为 8，优化器是 AdamW，EN-ZH 方向的学习率为 1e-5，ZH-EN 方向的学习率为 2e-5。

3.3 实验结果和分析

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）。实验在开发集和测试集上的表现如表 3 所示。

表 3 实验结果统计

Tab.3 Statistical experimental results

编号#	方法	EN-ZH		ZH-EN	
		DEV	TEST	DEV	TEST
1	Baseline	0.5266	0.4039	0.5355	0.4740
2	Pretrained	0.5825	0.4927	0.5842	0.6096
3	+PseudoPE	0.5556	0.4962	0.5515	0.5531
4	+DPFeature	0.5748	**0.5072**	0.5895	**0.6317**
5	+PseudoPE +DPFeature	0.4868	0.3716	0.5417	0.5356

表 3 中的 Baseline（#1）对应的是采用标准 XLM-R-Base 预训练模型的 QE 系统，Pretrained（#2）对应的是使用了进一步训练过的预训练模型的 QE 系统。从表中可以看出，在 EN-ZH 和 ZH-EN 的测试集上，改进的 Pretrained 系统的预测分值与 HTER 的相关性分别提升了近 9 个百分点和 13 个百分点。Pretrained 的系统性能提升显著，本文后续实验都在此基础上进行，分别进行了采用伪 PE 辅助策略，融合句法树特征策略以及它们的叠加实验。

表 3 中含 PseudoPE 命名的系统采用了伪 PE 辅助策略。+ PseudoPE （#3）是在改进的 Pretrained 系统上加入伪 PE 数据后的系统，相比 Pretrained，在 EN-ZH 和 ZH-EN 的验证集上的性能有较大幅度的下降，虽然在 EN-ZH 的测试集上有微小的提升，但在 ZH-EN 的测试集上性能却下降了 5.6 个百分点。这一结果验证了华为工作[16]的一个结论，即仅在编码器端简单通过拼接原文与伪 PE 数据无法提升系统性能。受限于系统结构，本文虽然无法像该工作那样通过“预测器”的解码器模块来进一步提取伪 PE 数据的特征表示，从而提升最终的评估性能，但我们也观察到，伪 PE 数据辅助训练的 QE 系统降低了在验证集和测试集上的性能不一致性，具有更好的收敛性。

表 3 中+ DPFeature（#4）表示在改进的 Pretrained 上融合 GAT 聚合依存特征信息的 QE 系统，从表中数据可以看到，该方法相对 Baseline 分别在 EN-ZH 和 ZH-EN 的测试集上获得了近 10 个百分点和 16 个百分点的提升；即使相对 Pretrained，性能提升也非常明显，在两个数据集的测试数据上均达到了最好的性能。如果说 Pretrained 系统是从双语对齐角度给评估器提供了非常有效的评估线索，那么融合双语依存特征携带的结构对应信息对 QE 任务也有着较大的帮助。

表 3 最后一行提供了同时应用三种优化策略的 QE 系统，即采用了进一步训练过的预训练模型，同时又使用伪 PE 数据和依存句法特征的 QE 系统。因为时间关系，本文没有能同

时对伪 PE 数据构建 GAT 句法特征，仅在改进的预训练模型的输入中拼接了伪 PE 数据，因此整体性能相对在 Pretrained 上仅使用 GAT 句法特征的系统性能上有大幅度下降，后续将尝试不在预训练模型的输入加入伪 PE 数据，而是为伪 PE 数据构建 GAT 句法特征表示的方案。

4 总结

本文介绍了我们在 CCMT 2022 中参加第 18 届全国机器翻译大会中参加翻译质量评估评测任务的采用的三个改进策略，分别是改进 QE 系统依赖的 XLM-R 预训练模型的策略、在输入端扩展伪 PE 数据的策略和在 XLM-R 输出端融入依存句法特征的策略。本文通过这些策略来提升 QE 系统提取特征的能力，根据实验结果表明，这三个策略中的第一和第三个策略性能提升显著，并且在使用融入依存句法特征的第三个策略后，QE 系统的性能在 EN-ZH 和 ZH-EN 测试集上均达到了最优。

参考文献：

[1] 科恩, 宗成庆, 张霄军. 统计机器翻译 [M]. [S.l.] : 统计机器翻译, 2012.

[2] COOK R. Machine Translation of Languages[J]. Mathematics Magazine, 1955.

[3] BROWN P F. The Mathematics of Statistical Machine Translation : Parameter Estimation[J]. Computational Linguistics, 1993.

[4] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation[C]//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. 2006: 223-231.

[5] 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.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[J]. Advances in Neural Information Processing Systems 30, 2017.

[7] FAN K, WANG J, LI B, et al. ”Bilingual Expert” Can Find Translation Errors[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[8] 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. 2019.

[9] DEVLIN J, CHANG M-W, 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). 2018.

[10] 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.

[11] Lample G, Conneau A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv:1901.07291, 2019.

[12] Ranasinghe T, Orasan C, Mitkov R. TransQuest at WMT2020: Sentence-level direct assessment[J]. arXiv preprint arXiv:2010.05318, 2020.

[13] Ye H, Gong Z. 利用语义关联增强的跨语言预训练模型的译文质量评估 (A Cross-language Pre-trained Model with Enhanced Semantic Connection for MT Quality Estimation)[C]//Proceedings of the 20th Chinese National Conference on Computational Linguistics. 2021: 23-34.

[14] Lample G, Conneau A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv:1901.07291, 2019.19

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.

[16] Wang M, Yang H, Shang H, et al. Hw-tsc's participation at wmt 2020 quality estimation shared task[C]//Proceedings of the Fifth Conference on Machine Translation. 2020: 1056-1061.

[17] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[18] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[19] Ni B， Lu X， Tong Y . SynXLM-R: Syntax-Enhanced XLM-R in Translation Quality Estimation[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2021.

[20] Huang, B., Carley, K.: Syntax-aware aspect level sentiment classification with graph attention

networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 2019, pp. 5469–5477. Association for Computational Linguistics (2019)