语义关联增强的 XLM-R 的译文质量评估

叶恒，陈世男，冯勤，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 CCMT2021 中英译文质量评估任务（QE），本文探索了预训练阶段和微调阶段的语义关联增强方法。首先本文在跨语言预训练模型 XLM-R 的基础上构建了 QE 基线系统，在实验过程中，本文发现原始的 XLM-R 并不具备良好的双语语义关联能力。因此本文首先使用双语掩码预训练方法继续预训练 XLM-R，其次在预训练阶段使用对比学习增强模型的语义表征能力，为了进一步提升模型性能本文使用 nist08 中英 1357 条数据制作假译文质量评估数据。本文在 CCMT2021 开发集 en-zh 上达到了 0.598 的皮尔森相关性，在 zh-en 上达到了 0.5575 的皮尔森相关性。

关键词：译文质量评估；跨语言预训练模型；对比学习；数据增强

中图分类号：TP 391.2　　**文献标志码：** A

译文质量评估任务（QE）[1-3]自提出以来就广受关注，它是在没有参考译文的条件下自动评估机器翻译译文的质量，QE 有着许多有趣且有用的应用场景，例如在外语教学中不依靠教师即可对不同的译文进行打分，通知用户商用机器翻译系统所给出的译文质量，清洗机器翻译任务所需要的平行语料等等。

在粒度划分上，QE 包含句子级别，词级别，以及文档级别。由于文档级别的 QE 粒度较大，因此其实现难度也较大，相比较句子级和词级的 QE，其短期的应用场景也较窄。句子级别的 QE 旨在只给定源语言待译文的条件下，自动给出译文的可靠性（HTER[4]，DA[5]）评分。词级别的 QE 在同样的条件下，自动给出译文中每个词的翻译正误情况，正确和错误分别给予 OK/BAD 标签。本文主要针对句子级别给出 CCMT2021QE 任务的解决方案。

通常句子级的 QE 任务被视为有监督的回归问题，而词级别的 QE 则被视为有监督的序列标注问题。在 QE 的发展历史中，首先为手工特征阶段。在这个时间节点上，由于计算设备和神经网络理论的限制，研究者主要集中于研究如何利用人工设计的语言特征[6]，基线特征[7]以及伪参考特征[8-9]，去联合机器学习算法来构建 QE 模型。在随后的发展中，联合多特征的神经网络方法逐渐成了主流，其描述语言的特征也不再仅局限于手工特征，一些额外的信息例如 Word2vec 的词向量表示，神经机器翻译模型自身对译文的困惑度可以更好的补充手工特征[10-11]。由于神经网络理论上可以对任何函数建模，因此神经 QE 模型可以更好的学习多特征与译文质量之间的联系，从而达到更好的性能表现。最近一段时间的 QE 模型转变为完全神经网络化的“端到端”模型，首先对语言的表示不再完全依靠于手工特征和额外的信息补充，而是利用自身编码器学习得到。此外，编码器和输出译文质量的评估器之间通过梯度连接，在训练过程中，能够一起优化。出乎意料的是，这种完全神经网络化的模型在许多任务中都大幅度超越了之前基于特征的模型，这其中当然也包括 QE 模型。近年来具有代表性的神经网络模型有“预测器-评估器”模型[12]，由于人工标注的 QE 语料稀缺且昂贵，而“预测器”的巧妙之处在于，通过词预测的方式在平行语料中学习通用的双语信息，他们通过实验表明预测器的词预测质量是影响整个模型质量评估能力的关键。这对后来的工作起了很大的启发作用，在其基础上衍生出一大批基于“预测器-评估器”框架的改进模型。Fan 等人提出可以将机器翻译系统看作一个黑匣子，质量评估模型只需要对这个黑匣子建模即可，由此他

基金项目：国家自然科学基金(61976148)

***通讯作者：**zhxgong@suda.edu.cn

们提出“双语专家”模型[13]，通过双向的 transformer 可以学习条件目标语言模型，在下游任务中，通过强制解码的方式可以得到译文和条件目标语言模型所认为的真实译文分布之间的差异。近年来，出现了一批跨语言预训练模型，例如 mBert[14]，XLM[15]和 XLM-Roberta[16]等等，利用多语言的单语语料或平行语料，跨语言预训练模型可以学习到多语言之间通用的语言知识。在语言表示上，跨语言预训练模型可以得到不同语言下的收缩的特征分布。在 QE 任务中，使用跨语言预训练模型可以直接编码源语言和目标语言[17]。但是由于跨语言预训练模型并不具备良好的双语关联能力，因此在 QE 任务中需要对模型进行进一步的预训练和微调[18]。

本文在跨语言预训练模型 XLM-R 的基础上，利用平行语料增强模型的语义关联能力，具体方法是利用 “全掩码预训练策略”在平行语料上预训练。此外，为了进一步增强单语的语义表示，我们使用对比学习方法

1 数据

CCMT2021 QE 语料为三元组$< X, Y, H >$形式的数据，其中定义源语言待译文$X = (x_1, x_2, \ldots x_n)$，目标端译文$Y = (y_1, y_2, \ldots y_n)$，译文错误人工后修正率 HTER: $H =$ 译文中错译字词的长度/参考译文长度，其中每条源语言待译句只对应一条译文。表 1 给出了数据集中的一条样例。

表 1 CCMT 数据样例

Tab.1 CCMT data sample

Src	The graduation rate of black men is lower than that of any group .
Tgt	黑 人 男 子 的 毕 业 率 低 于 任 何 一 组 。
Pe	黑 人 男 子 的 毕 业 率 低 于 任 何 群 体 。
HTER	$\frac{\text{编辑次数}}{\text{参考译文的单词数}} = \frac{2}{15} = 0.133333$

2 基线系统

本文基于跨语言预训练模型 XLM-R 构建译文质量评估基线系统，其架构如图所示。XLM-R 作为底层的双语编码器，可以一次编码源语言和目标语言。在编码过程中，目标端融合了源端信息，在最终的编码表征中本文使用”<s>”(类似于 Bert 中的[CLS]符号，它能在一定程度上反应整体语义信息)，”mean-pooling”（源端和目标端的平均池化），”target-meaning”（仅在目标端的平均池化）三种方式获得句子级的质量评估向量。

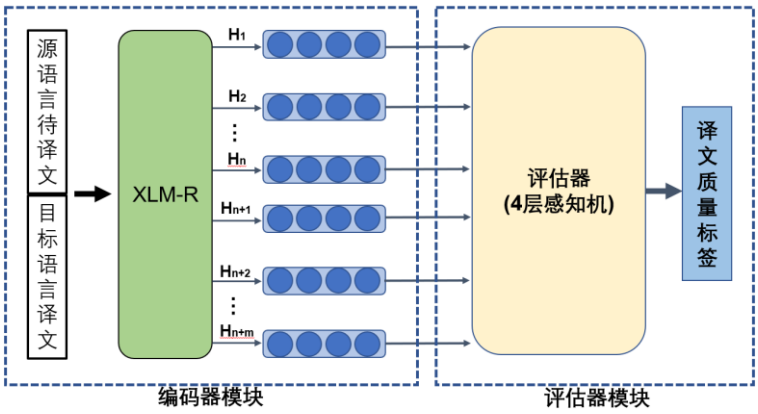

图 1 基线系统

Fig.1 Baseline system

2 增强双语语义关联的训练策略

直接利用 XLM-R 即可快速的搭建一个 QE 系统，但是我们对这样的性能表现并不满足。原因在于（1）XLM-R 仅在多语言的单语语料上预训练，没有任何语言指示符，也未用到多语平行语料。（2）基线系统仅使用“<s>“语义或 pooling 方式代表全句信息，这对于 QE 任务来说处理方式不够细致。为此本文进一步的研究预训练阶段双语语料的引入以及微调阶段语义增强方法对最终 QE 性能的影响。

联合全词掩码策略

本文利用 CMMT2021 所提供的中英平行语料对 XLM-R 进一步的预训练。具体方法是将源端待译句和目标端译文拼在一起，并在同时在源语言和目标语言端随机选取 15%的全词，其中 80%用特殊符号“<mask>”代替，10%不做任何处理，剩下 10%用词表中任意一个词替换。使用全词掩码的好处在于，避免了相对简单的子词掩码，增强了训练信号，因此能够增强模型的语义理解能力。在双语掩码任务中，全词掩码可以鼓励模型从平行语料中寻找与之对应的词，隐式的增强了模型的双语语义关联能力。

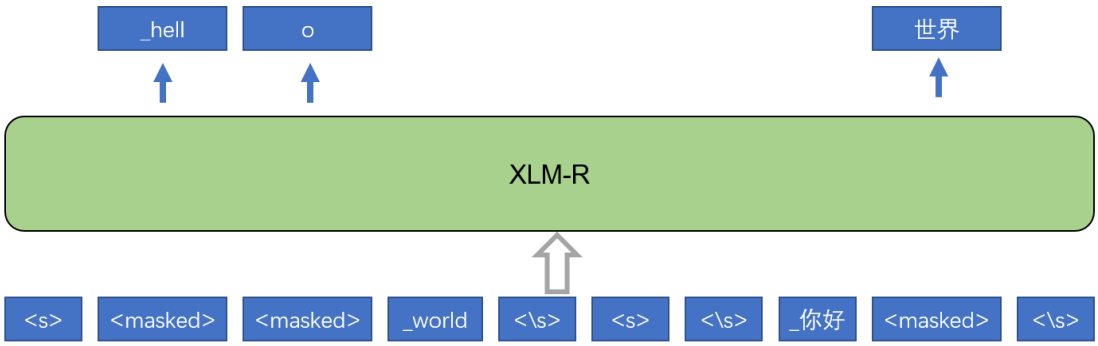

图 2 联合全词掩码策略

Fig.2 Joint full word mask policy

3 对比学习增强语义表示

对比学习简单的说就是将相似的样本距离拉近，将不相似的样本距离拉远。本文参考 SimCSE 的方法[19]，使用 Dropout 掩码直接做数据增强。具体的，同一个样本经过两次模型前向传播得到两个正例对，负例则是同一个 batch 内部其它句子。对于 QE 任务而言，借助这种对比学习方法可以在一定程度上增强模型的句子表征能力。

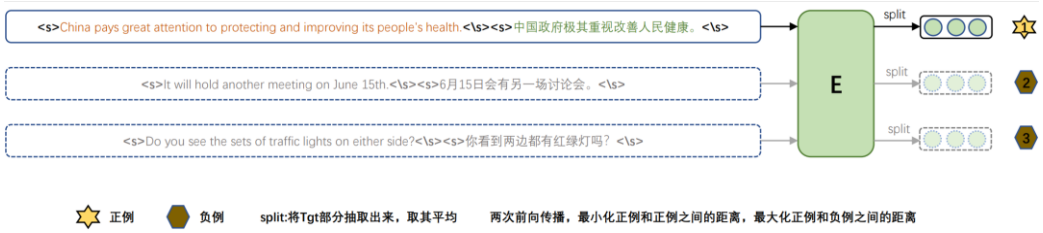

图 3 对比学习策略

Fig.3 Contrastive learning strategy

4 数据增强

在 ZH-EN 和 EN-ZH 两个 QE 任务上，本文都使用了 nist08 的平行语料，但是具体的使用方法却并不一样。本文将分别介绍它们。

在 EN-ZH QE 任务上，句子级质量评估的数据集包括<src, tgt, hter>，其中 hter 是通过计算将译文 tgt 编辑成可供出版的标准译文 pe 所需要的编辑距离，这个过程往往需要人工参与，因此质量评估任务不可避免的成为小语料任务。为此，我们使用 nist08 数据集进行了数据增强。首先通过谷歌翻译 API 将 nist08 中文(src)翻译成英文译文(tgt)，再通过 TER 工具计算英文译文与第一个参考翻译(pe)的编辑距离，即伪 hter，从而构造出质量评估数据，并与原训练语料打乱顺序合并。

在 ZH-EN QE 任务上，nist08 中每条中文包含 4 条参考英文译文，因此本文从 4 种类别中随机均匀采样得到增强数据集。

5 实验设置及结果

5.1 实验设置

为验证上述方法的性能 ，我们在 CCMT2021 句子级译文质量评估训练集进行训练，并在 19 验证集和测试集进行验证和测试，表 1 给出了 EN-ZH 和 ZH-EN 两个方向上译文质量评估语料、用于数据增强的 nist08 数据集以及用于训练 en-zh 预训练了模型的平行语料的的规模。

表 1 实验语料规模统计

Tab.1 Size statistics of experimental corpus

语料	训练集	验证集	测试集
EN-ZH	14789	1381	1445
ZH-EN	10070	1143	1385
Nist08	1357	-	-
EN-ZH 平行语料	913M	9000	

EN-ZH 预训练模型采用 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 12 个头。训练设置为：最大序列长度为 200，批次大小 16，学习率为 1e-5，使用的优化器为 AdamW，β_1=0.9，β_2=0.999，ϵ=1e-8，训练 6 个 epoch，早停数为 10，并进行了 3 折交叉验证。

ZH-EN 预训练模型采用 xlm-roberta-large 模型，该模型包含 24 层编码器，隐层维度为 1024 维，多头注意力机制设置 16 个头。训练设置为：最大序列 120，批次大小 16，学习率为 5e-6，使用的优化器为 AdamW，β_1=0.9，β_2=0.999，ϵ=1e-8，训练 6 个 epoch，早停数为 10，并进行了 3 折交叉验证。

为了评估译文质量评估模型的性能，我们用皮尔森相关系数（Pearson）进行衡量。该系统用于反映预测值与真实值的线性相关程度，系数值越接近 1，表示相关性越高，越接近 0，表示相关性越低，若为负数，则表示负相关。

5.2 实验结果

表 2 不同数据集下的 Pearson 值

Tab.2 Pearson values under different data sets

方法	EN-ZH		ZH-EN	
	DEV	TEST	DEV	TEST
+Pretrained	0.613	-	-	-
+DataAug+SimCSE	-	-	55.75	52.51
Baseline	0.548	0.453	55.03	50.14

如表 2 所示，在 EN-ZH 数据集上，预训练方法在验证集上提升了 7 个点；在 ZH-EN 数据上，通过数据增强和对比学习方法，在验证集上提升了 0.7 个点，在测试机上提升了 2.4 个点。

6 结论

本文介绍了我们在 CCMT2021 中英译文质量评估任务上的解决方案，首先基于 XLM-R 和 4 层感知机构建译文评估系统基线预测 HTER。在此基础上，在预训练和微调阶段使用了不同的训练方法，目的都是为了得到更好的译文表征。

参考文献：

[1] Soricut Radu, Bach Nguyen, Wang Ziyuan. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task[C]//Proceedings of the WMT，2012：145-151.

[2] Christian Hardmeier, Joakim Nivre, Jorg Tiedemann. Tree Kernels for Machine Translation Quality Estimation[C]//Proceedings of the WMT，2012：109-113.

[3] David Langlois, Sylvain Raybaud, Kamel Smaïli. LORIA System for the WMT12 Quality Estimation Shared Task[C]//Proceedings of the WMT，2012：114-119.

[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] Graham Y, Baldwin T, Moffat A, et al. Can machine translation systems be evaluated by the crowd alone[J]. Natural Language Engineering, 2017, 23(1): 3-30.

[6] Felice M, Specia L. Linguistic features for quality estimation[C]//Proceedings of the Seventh Workshop on Statistical Machine Translation. 2012: 96-103.

[7] Specia L, Shah K, De Souza J G C, et al. QuEst-A translation quality estimation framework[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013: 79-84.

[8] Soricut R, Echihabi A. Trustrank: Inducing trust in automatic translations via ranking[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 612-621.

[9] Kozlova A, Shmatova M, Frolov A. Ysda participation in the wmt'16 quality estimation shared task[C]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016: 793-799.

[10] Martins A F T, Junczys-Dowmunt M, Kepler F N, et al. Pushing the limits of translation quality estimation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 205-218.

[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale[J]. arXiv preprint arXiv:1911.02116, 2019.
[16] Lample G, Conneau A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv:1901.07291, 2019.
[17] Pires T, Schlinger E, Garrette D. How multilingual is multilingual bert?[J]. arXiv preprint arXiv:1906.01502, 2019.
[18] Yan Y, Li R, Wang S, et al. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer[J]. arXiv preprint arXiv:2105.11741, 2021.
[19] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.

Evaluation of translation quality for XLM-R with semantic relevance enhancement

YE Heng，CHEN Shinan，FENG Qin，GONG Zhengxian*

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Aiming at the CCMT2021 Chinese/English Translation Quality Assessment (QE) task, this paper explores the semantic relevance enhancement methods in the pre-training and fine-tuning stages.Firstly, this paper builds a QE baseline system on the basis of the cross-language pre-training model XLM-R. In the process of experiment, it is found that the original XLM-R does not have good bilingual semantic correlation ability.Therefore, this paper firstly uses the bilingual mask pre-training method to continue the pre-training of XLM-R, and then uses contrastive learning in the pre-training stage to enhance the semantic representation ability of the model. In order to further improve the performance of the model, this paper uses 1357 Chinese and English data of NIST08 to make the fake translation quality assessment data.In this paper, Pearson correlation is 0.598 on the CCMT2021 development set EN-ZH and 0.5575 on the ZH-EN.

Key Words：Quality Estimation；Cross-Lingual Pretrained Model；Contrastive Learning；Data Augmentation