利用语义关联增强的跨语言预训练模型的译文质量评估

叶恒，贡正仙*
（苏州大学自然语言处理实验室，江苏苏州 215006）
20204227003@stu.suda.edu.cn，zhxgong@suda.edu.cn

摘要

机器翻译质量评估(QE)虽然不需要参考译文就能进行自动评估，但它需要人工标注的评估数据进行训练。基于神经网络框架的QE为了克服人工评估数据的稀缺问题，通常包括两个阶段，首先借助大规模的平行语料学习双语对齐，然后在小规模评估数据集上进行评估建模。跨语言预训练模型可以用来代替该任务第一阶段的学习过程，因此本文首先建议一个基于XLM−R的为源/目标语言统一编码的QE模型。其次，由于大多数预训练模型是在多语言的单语数据集上构建的，因此两两语言对的语义关联能力相对较弱。为了能使跨语言预训练模型更好地适应QE任务，本文提出用三种预训练策略来增强预训练模型的跨语言语义关联能力。本文的方法在WMT2017和WMT2019英德评估数据集上都达到了最高性能。

关键词： 译文质量评估；语义关联；跨语言预训练模型；预训练策略

A Cross-language Pre-trained Model with Enhanced Semantic Connection for MT Quality Estimation

Ye Heng, Gong Zhengxian
(Natural Language Processing Laboratory, Soochow University, Suzhou, Jiangsu 215006)
20204227003@stu.suda.edu.cn，zhxgong@suda.edu.cn

Abstract

Quality Estimation(QE) of Machine Translation(MT) can automatically estimate the quality of MT outputs without references, but its model needs to be trained on manually annotated evaluation data. Due to the lack of manual data, the QE Systems with neural network architecture usually work in two stages, involving firstly building bilingual alignment on large scale parallel corpora and secondly creating estimation model on small scale evaluation dataset. Currently the first stage can be taken over by employing cross−language pre−trained models. This paper firstly proposes a unified encoder based on XLM−R for simultaneously encoding source and target text. Second, cross−language pre−trained models,usually trained on monolingual data in multiple languages, have weak semantic connection between pairwise languages. In order to make pre−trained models adapted to QE task, this paper proposes three pre−training strategies to enhance cross−language semantic connection for pre−trained models. The proposed method in this paper has reached the new SOTA performance on both the WMT2017 and WMT2019 quality estimation data sets.

*通讯作者corresponding author

第二十届中国计算语言学大会论文集，第23页-第34页，呼和浩特，中国，2021年8月13日至15日。

Keywords: Quality Estimation , Semantic Connection , Cross−Language Pre−trained Model , Pre−training Strategy

1 引言

近年来神经机器翻译迅猛发展(Ilya Sutskeveret al., 2014; Bahdanauet al., 2014)，机器翻译系统给出的译文质量也越来越高，但是机器翻译系统仍会产生许多不可靠的译文。例如英文待译句：“John met his wife in the hot spring of 1988. ”，谷歌翻译给出的德语译文却是：“John lernte seine Frau in der heißen Quelle von 1988 kennen ”，翻译系统将“hot spring”（春天）错翻成了“Quelle”（热泉），这种一词多义导致的错误译文就需要翻译专家再进行一次修正。为了在翻译开销和翻译质量之间达到平衡，通常做法是先机翻一遍，再进行一次人工修正。为了减少人工修正的难度，设计一个能自动评估机器翻译质量的系统也成了迫切的需求。传统机器翻译的评估方法，如BLEU(Dzmitry Bahdanauet al., 2014)，Meteor(Denkowski and Lavie, 2014)和ROUGE(Chin-Yew Lin, 2014)，需要高质量的参考译文。译文质量评估(Quality Estimation,QE)能在没有参考译文的条件下自动评估机器译文的质量(Blatzet al., 2014; Speciaet al., 2013)，它通常分为三个粒度：词级别，句子级别和文档级别。在词级别，QE系统会给出目标端每个词的翻译正误情况，翻译正确给“OK”标签，错误给“BAD”标签。在句子级，QE系统会给出每句译文的可靠性，通常使用HTER指标(Snoveret al., 2006)。在文档级，QE系统会给出译文的翻译可靠性。

src	to select one frame , click the frame .
mt	klicken sie auf den rahmen , um einen rahmen auszuwählen .
pe	klicken sie auf ein bild , um es auszuwählen .
tags	OK OK OK BAD BAD OK OK BAD BAD OK OK
hter	$\frac{the\ count\ of\ BAD}{the\ token\ num\ of\ pe}=\frac{4}{10}=0.4000$

Table 1: WMT19质量评估数据集中的一个样例。其中src表示源语言待译句，mt表示src的翻译结果，pe为人工后编辑的结果，hter为HTER值。

随着预训练模型的快速发展，QE任务可以将m−Bert(Devlinet al., 2018)，XLM−R(Conneauet al., 2019)，XLM(Conneau and Lample, 2019)等预训练模型当成双语编码器，但这些预训练模型大多数在多语言的单语语料上构建，因此缺少双语对应信息。而译文质量评估任务更需要编码器的双语语义关联能力使其能根据源端推断当前目标端译文的质量水平。

本文首先构建一个基于XLM−R的译文质量评估模型，并通过增强XLM−R的双语语义关联能力来提高译文质量评估能力，为此分两步进行了研究：首先，为了减少编码过程中双语信息的丢失，本文抛弃独立编码器方法，仅在统一编码器的基础上进行译文质量评估任务，并通过实验证明统一编码器的优势；其次，本文充分利用WMT19译文质量评估比赛所提供的平行语料，根据译文质量评估任务的特点设计三种掩码预训练策略。在最终的实验中，本文所提出的语义关联增强模型在WMT2017和WMT2019质量评估英-德数据集上达到了新的单模型最高性能。

2 相关工作

译文质量评估方法主要分为无监督方法和有监督方法。无监督方法依照机器翻译系统自身对翻译结果的置信度或基于预训练的词向量(Fanet al., 2020)。在没有标注数据的情况下，基于机器翻译系统模型的方法能在低资源语种和中资源语种上能取得和有监督方法相媲美的性能，但在高资源语种上无监督方法却远落后于有监督方法。基于预训练词向量的方法使用起来简单，但由于预训练词向量和译文质量评估数据之间的差异性以及缺少对词向量的微调

基金项目：国家自然科学基金(61976148)，江苏高校优势学科建设工程资助项目

过程，这类方法的性能并不高。有监督方法将译文质量评估任务当成回归/分类问题，其依赖如表1所示的标注数据。传统的线性神经网络模型十分依赖手工标注的语言特征(Speciaet al., 2013; Martinset al., 2017)，然而这种特征表示方法领域性强且有不鲁棒的缺点，这些缺陷导致线性神经网络模型可迁移性弱，无法在复杂的译文质量评估场景中被继续使用。近年来有监督方法演变为“预测器-评估器”双阶段方法(Kimet al., 2017)，第一阶段在平行语料上进行词预测预训练，预测器因此具备一定的双语理解能力(Iveet al., 2018; Iveet al., 2018; Miaoet al., 2019; Blainet al., 2020)，源语言和目标语言之间使仅用注意力机制(Bahdanauet al., 2014)关联以交互信息，且词预测器需从头训练，不仅耗时极大且受限于平行语料规模。除此之外，预测器的预训练任务为预测目标语言译文被随机选中的子词，这种预训练任务较简单，导致模型预训练阶段的双语学习效率较低。“双语专家”模型(Fanet al., 2019)基于transformer探索了模型预测和译文词之间的差异对质量预测的影响，但其编码器模块基于目标语言，源端和目标端之间的交互较小。Cui(2021) 等人为了缓解质量评估数据集稀缺的问题，提出“生成器–检测器”双模块模型。生成器模型和检测器模型都基于transformer。生成器模块为掩码语言模型，训练目标为还原目标端被掩码的部分。检测器为序列标注模型，训练目标为检测译文中错误的部分。“生成器–检测器”模型充分利用了平行语料，利用生成器将平行语料转换成假译文质量评估数据，在一定程度上缓解了质量评估数据集稀缺的问题。

3 质量评估任务的形式化定义

译文质量评估任务是在没有参考译文的条件下预测译文的翻译质量，本文专注于词级别和句子级别，假设给定源语言待翻译句$S = \{s_1, s_2, ..., s_n\}$和翻译结果$T = \{t_1, t_2, ..., t_m\}$，译文质量评估系统需要输出两种质量标签：

- 词级别标签：$O = \{o_1, o_2, ..., o_m\}$，其中$o_j$表示翻译结果中第j个词的质量标签，通常为“OK”或“BAD”。
- 句子级标签：$HTER$(Snoveret al., 2006)，表示整个翻译句子的人工后修正率。

译文质量评估数据集的样本特征为$E = \{S, T, O, HTER\}$。

4 基于XLM−R的QE基线系统

本文的QE基线系统如图1所示，包括编码器和评估器两个部分。编码器负责将待评估数据的文本进行向量化表示，评估器接受编码器的隐层向量构建神经网络，通过评估数据进行参数学习，最终获得自动评估的能力。

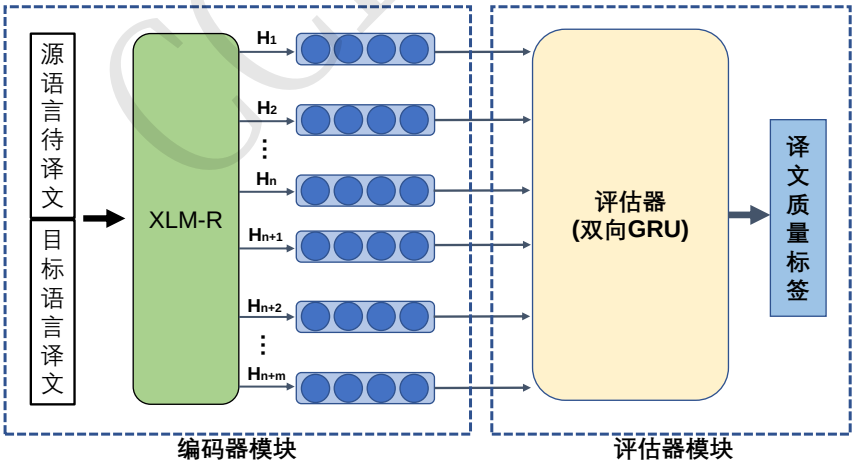


Figure 1: 基于XLM−R的译文质量评估模型图

4.1 统一编码器

本文的编码器基于XLM−R预训练模型，XLM−R在多项跨语言理解任务上都取得了SOTA的效果，利用XLM−R可以直接编码源语言待译句和译文。

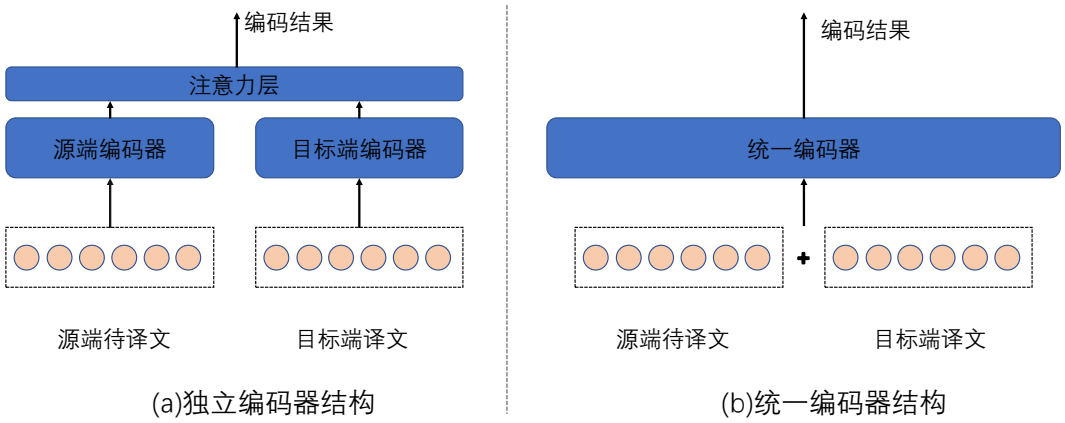


Figure 2: 独立编码结构和统一编码结构

在双语编码过程中，有“独立编码”和“统一编码”这两种编码方法。其中独立编码是源语言和目标语言使用共享参数的编码器分别编码，统一编码是将源语言待译文和目标语言译文拼接起来一次编码(图2展示了两种编码方式的区别)。使用独立编码器编码双语时，借助注意力机制能在一定程度上融合源端和目标端的信息，但是由于编码过程相互独立，因此独立编码器并不能完全捕获源端和目标端之间的交互。当编码器为transformer时，“自注意力机制”(Vaswaniet al., 2017)仅会在各自编码过程中起作用，因此采用独立编码方式会导致编码器端的双语信息缺失问题，而统一编码器则不存在这样的问题。统一编码器通过自注意力机制在源语言词和目标语言词之间进行注意力关联，而这种直接的注意力关联能帮助模型捕获更深层更复杂的翻译质量信息。

自注意力机制会将输入分别投影到“查询”(Q)，“键”(K)和“值”(V)这三个不同的子空间，随后计算注意力权值，其计算公式为：

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

在双语编码过程中，输入序列中的每一词都会和序列中其它的词进行匹配计算，因而更容易捕获双语信息。采用统一编码器结构，源端和目标端间直接的注意力关联相比较独立编码器更能挖掘译文质量信息。除此之外，利用多头注意力能使统一编码器在不同的表示子空间中学习到更多的双语信息。其计算公式为：

$$MultiHAtt(Q,K,V) = Concat(head_1, head_2, ..., head_h)W^O \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

参数矩阵$W_i^Q \in R^{d_m \times d_k}, W_i^K \in R^{d_m \times d_k}, W_i^V \in R^{d_m \times d_k}, W^O \in R^{hd_v \times d_m}$，其中h为多头注意力的数量,$d_m$和$d_k$分别为编码器的编码维度和注意力头的维度。

根据Jawahar(2019)等人的研究，Bert预训练模型每一层都学习到不同的语言知识，其中底层能学习到通用的语言学知识，而顶层则更多的是学习和语义相关的信息。受此启发，本文融合编码器12层的隐状态向量，通过参数学习的方式确定每一层的权重。统一编码器编码后每一层的隐状态为$\{H_1, H_2, ..., H_{12}\}$，$H_i = \{h_1, h_2, ..., h_{m+n}\}, h_i \in R^{d_m}$。其中每一层所对应的权重为$W_l = \{w_1, w_2, ..., w_{12}\}$，混合后得到$H_{mix} = \sum_{i=1}^{12} w_i * H_i, H_{mix} = \{h_{mix-1}, h_{mix-2}, ..., h_{mix-n+m}\}$。其中源语言待译句编码向量为$H_s = \{h_{mix-1}, h_{mix-2}, ..., h_{mix-n}\}$，目标语言编码向量为$H_t = \{h_{mix-n+1}, h_{mix-n+2}, ..., h_{mix-n+m}\}$

4.2 基于GRU的评估器

基线系统的编码器将双语编码为向量，评估器在编码器输出的基础上，联合译文质量评估标注数据进行学习。RNN网络可以在双语编码器的基础上对时序信息建模，因此本文的评估器

采用双向GRU单元构建RNN网络，其编码过程为：

$$\overleftarrow{\mathbf{h}}_{GRU-i} = \overleftarrow{GRU}(h_{mix-i}, \overleftarrow{\mathbf{h}}_{GRU-i+1}) \quad (4)$$

$$\overrightarrow{\mathbf{h}}_{GRU-i} = \overrightarrow{GRU}(h_{mix-i}, \overrightarrow{\mathbf{h}}_{GRU-i-1}) \quad (5)$$

融合双向编码得到$H_{GRU-i} = Concat(\overleftarrow{\mathbf{h}}_{GRU-i}, \overrightarrow{\mathbf{h}}_{GRU-i})$，其中Concat表示拼接操作。

句子级评估

句子级涉及译文整体翻译质量的评估，通过平均池化层得到RNN再次编码后的句子表征$H_{sen} = \frac{1}{m+n}\sum_{i=1}^{m+n}\mathbf{h}_{GRU-i}$，随后将其作为句子级评估器的输入得到预测的HTER值：

$$HTER_{pred} = tanh(W_0(tanh((W1 * H_{sen})))) \quad (6)$$

其中$W_0 \in R^{1\times d_m}, W_1 \in R^{d_m \times d_m}$ 。

词级别评估

词级别为序列标注问题，每个词都需要分配一个OK/BAD标签，通过softmax网络确定每个词的类别标签。

$$Prob = softmax(\mathbf{H}_{GRU-i}) \quad (7)$$

如果预测概率值Prob大于0.5，则当前词为“OK”，小于0.5则为“BAD”。

5 基于语义关联增强的跨语言预训练模型的QE系统

5.1 QE系统的改进方案

基线系统的编码器所依赖的XLM−R预训练模型是在大规模多语言的单语数据集上训练获得的，虽然它可以对评估数据的源/目标端的双语数据同时进行编码，但其构造过程决定了它的双语语义空间的对应能力较弱（本文的实验分析部分进一步验证了这个特点），因此也局限了它的编码表示能力。

根据上述分析，再加上受Cui(2021)等人的研究工作的启发，本文建议用一个增强跨语言语义关联的预训练模型来改进QE的性能。这种改进方式非常自由灵活，基线系统的结构不需要发生变化，只要把统一编码器的XLM−R模型替换成改进的跨语言预训练模型即可。

为了建立跨语言语义关联增强的预训练模型，受XLM以及Unicoder的启发，本文利用WMT19译文质量评估比赛所提供的平行语料，采用三种掩码预训练策略来提升XLM−R预训练模型的双语对应能力，包括“全掩码预训练策略”，“目标端全词掩码预训练策略”，以及它们的混合方式。

5.2 语义关联增强的预训练策略

(1)全掩码预训练策略

XLM−R所使用的分词方法为sentence−piece[0]，这种子词切分方法能从语料中学习到更细粒度的子词表示，优点是可以减小词表大小以及减少“UNK”现象(Sennrichet al., 2015)。XLM-R的词表包含250000个子词，其中有很多和训练所用的英德平行语料无关的子词。在不改变词表的条件下，为了让模型更专注于英德语料的子词表示，本文提出“全掩码预训练策略”（下文简称fully mask）。全掩码预训练用特殊符号$< mask >$替代源端和目标端的子词。在训练过程中，模型需要根据源端和目标端的上下文推断出被掩码部分的原文。同时fully mask也鼓励模型平等的对待源端语言和目标端语言，在推断被掩码部分时可以在源端和目标端之间互相参照。例如给定待译句“play the movie in a separate window .”和译文“spielen Sie den Film in einem separaten Fenster ab .”，当待译句中的“play”（播放）和译文中的“Film”（电影）被掩码时。还原“play”可以参照译文中的“spielen”（播放），还原“Film”时可以参照待译句中的“movie”（电影）。

全掩码预训练任务的掩码方法为在目标端和源端间各自随机选择15%的子词，其中80%用特殊符号$< mask >$代替，10%用词表中任意一个词代替，剩下10%则原封不动。形式化定义如8所示：

$$\mathcal{L}_{fully\ mask} = -\sum_{\widehat{x}\in\{T_{s-m(x)}\cap S_{s-m(x)}\}} \log p(\widehat{x}|S_{\backslash s-m(x)}, T_{\backslash s-m(x)}) \quad (8)$$

[0]https://github.com/google/sentencepiece

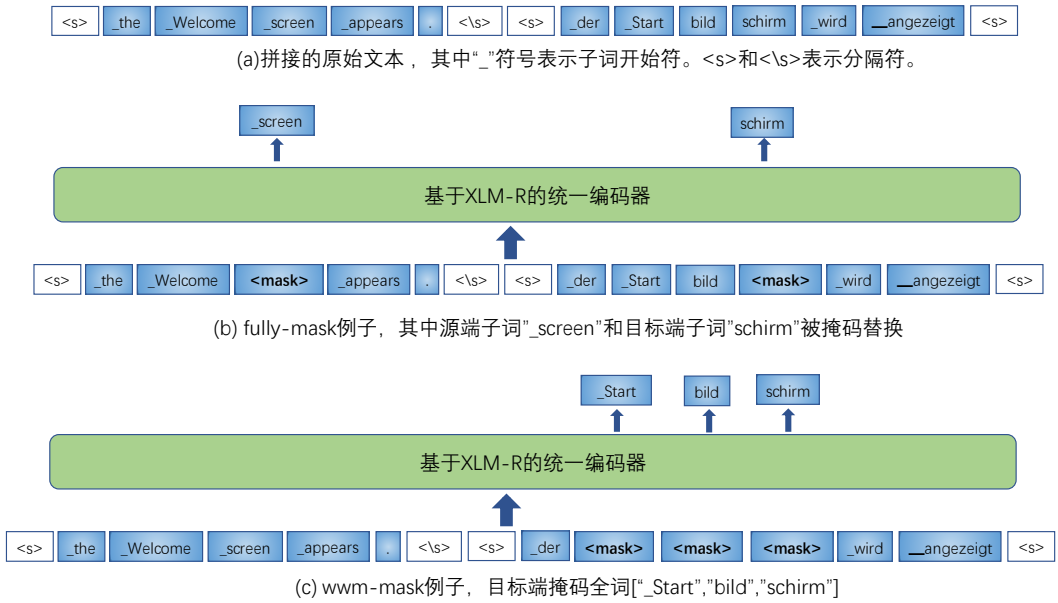


Figure 3: 两种掩码方式对比

其中$T_{s-m(x)}$和$S_{s-m(x)}$表示源端待译句和目标端译文被掩码的子词部分，$S_{\backslash s-m(x)}$和$T_{\backslash s-m(x)}$表示剩下的部分。

(2)目标端全词掩码预训练策略

为了让模型更好地学习目标语言词对源语言词的依赖关系，本文在预训练阶段引入"目标端全词掩码预训练策略"（下文简称wwm mask）。将源端待译句和目标端译文拼接成一句，但仅掩码目标端译文的全词。全词掩码和子词掩码之间的不同在于是否忽略分词结果。假设给定词"hello"，分词器会根据词表将"hello"分割为"hell"和"o"两个子词，全词会掩码以上两个部分，子词会从所有子词中依概率选取需要掩码的部分。通过全词掩码任务鼓励模型从源端中找到和全词掩码部分对应的源端词。考虑到子词掩码的还原可以更多的借助自身上下文的信息，如"startbildschirm"(屏幕)被分成子词['start', 'bild', 'schirm']后，还原"bild"完全可以借助自身上下文轻松还原。而全词掩码后，对模型语义关联能力要求更高。

目标端全词掩码预训练策略为随机选择目标端15%的全词，其中80%用特殊符号<$mask$>代替，10%用词表中任意一个词代替，剩下10%则原封不动。在此基础上为了避免模型对源端的遗忘，本文在源端添加额外的自编码任务，选择源端10%的子词不做任何处理，鼓励模型尽可能正确的还原出来。形式化定义为:

$$\mathcal{L}_{wwm\ mask} = -\sum_{\widehat{x}\in\{T_{a-m(x)}\cap S_{r-(x)}\}} \log p(\widehat{x}|S, T_{\backslash a-m(x)}) \tag{9}$$

以上$T_{a-m(x)}$表示目标端译文被掩码的全词部分，$S_{r-m(x)}$表示源端被随机选中的子词，S表示源端待译句，$T_{\backslash a-m(x)}$表示目标端剩下的部分。

6 实验

6.1 数据集

本文实验的英德译文质量评估数据和平行语料来源于WMT译文质量评估任务，其中质量评估数据集包含WMT2017和WMT2019两部分，平行语料为WMT2019质量评估任务所提供的IT领域的英德平行语料。表2列出各项语料的规模。

6.2 评估指标

本文沿用WMT2019的的评估指标测试译文质量评估模型的性能。在句子级别，官方评价指标包含Pearson相关性(下文用r代替)，Spearman相关性（下文用p代替），均方误差

任务发布时间	类型	名称	数量
WMT2019	英德平行语料	训练集	3.4M
WMT2019	英德质量评估数据	训练集 验证集 测试集	13442 1000 1023
WMT2017	英德质量评估数据	训练集 验证集 测试集	23000 1000 1000

Table 2: 实验数据规模

（MAE,Mean Absolute Error）和平方根误差（RMSE,Root Mean Squared Error）。在词级别，官方指标为F1−Multi。本文所报告的性能均是在官方评估指标下验证集性能最好的模型在测试集上的性能，对比模型性能时，在句子级别按照Pearson相关性和Spearman相关性排名，在词级别按照F1−Multi排名。

6.3 实验设置

本文使用的优化器为AdamW(Loshchilov and Hutter, 2017)，其中优化器的参数分别为$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$，权值衰减比例为$1e-5$。预训练阶段，训练总步数为100万，学习率按照线性热身方法调整，10000步之后达到最大值$1e-4$，在随后的训练中线性衰减。批大小为16个句子，梯度累计为2倍。在译文质量评估微调阶段，通过在验证集上进行超参数搜索，本文使用在验证集上性能最好的超参数值。其中学习率为$3e-5$，批大小为8。训练所用到的显卡为GTX 1080ti。

6.4 实验结果

(1)统一编码的优势

在上文中提到了独立编码器存在编码时缺失双语信息的问题，因此本文使用统一编码器。表3为两种编码方法的性能比较，从实验结果看出统一编码器在双语编码过程中，充分利用了自注意力机制挖掘译文的质量信息。并且是由于一次编码，因此运算效率相比独立编码更高。需要说明是独立编码和统一编码在编码器端的参数是一致的。

模型	WMT 2017			WMT 2019		
	r↑	F1−Multi↑	训练时间↓	r↑	F1−Multi↑	训练时间↓
独立编码器	0.6496	0.5288	1.319	0.4461	0.3298	1.315
统一编码器	**0.6745**	**0.5418**	1.0	**0.5086**	**0.3703**	1.0

Table 3: 独立编码器和统一编码器器在WMT2017和WMT2019验证集上的性能，其中训练时间以统一编码为标准，越大则用时越长。

(2)预训练方法的有效性

为了验证本文所提出的三种预训练方法的有效性，本文选取了近年来在国际会议上针对译文质量评估任务所提出的模型作为对比系统，首先简要介绍这些对比系统。

- **QE-Bert**：由Kimet.(2019)提出针对译文质量评估任务的预训练方法，使用双语语料，同时在目标语言端插入[GAP]表示间隔，预训练任务为随机掩码平行语料并还原。但其插入的[GAP]特殊符号并不会在真实的译文质量评估数据中出现，这种差异可能影响双语编码。

- **NMT-style**：使用NMT风格的seq2seq模型，预训练任务为利用平行语料预测目标语言的子词。编码器为双向RNN或transfomrer，当编码器为rnn时仅使用简单的注意力机制融合双语信息，当编码器为transformer时，未就QE任务设计语义关联增强模块或任务。

- **NULL-Ins**：由于真实译文中存在漏译问题，因此Cuiet等人(2020)在平行语料中插入空字符，预训练任务为在插入空字符的平行语料上随机掩码并还原（本文简称其为NULL-

Ins）。模型需要判断被掩码的部分为空字符还是为正常掩码的子词，这种掩码方法能强制模型学习源语言子词和目标语言子词之间的对齐关系。但强行插入的空字符具有很强的随机性，无法模拟真实的漏译，因此仅在一定程度适应QE任务。

- **Gen-Det**：由Cui等人(2021)提出的"生成器-检测器"模型（本文简称为Gen-Det），其预训练任务包含生成器的掩码预测和检测器的译文质量检测，其中检测器的预训练数据来源于生成器。"生成器-检测器"是现阶段的最高性能模型，但由于生成器的生成能力较弱，无法生成多样化的译文质量评估数据。

- **Ours**：除了上文所提到的两种掩码策略，本文还综合两种预训练任务的特点，提出将fully mask和wwm mask混合的预训练策略（下文简称mixed mask）。混合方式为进行3次wwm mask策略之后，就进行1次fully mask策略，如此循环。附加这种预训练策略的动机在于避免模型对单一任务的适应性，混合的掩码策略会增加预训练阶段的难度，以更好地挖掘双语语料中的信息。

上述系统在WMT2019和WMT2017句子级的实验结果如表4和表5所示。

预训练策略	方法	句子级 验证集				句子级 测试集			
		r↑	p↑	MAE↓	RMSE↓	r↑	p↑	MAE↓	RMSE↓
QE−Bert	(Kimet al., 2019)⋆	0.545	0.575	-	-	0.526	0.574	-	-
NMT−style	(Hou Qi, 2019)⋆	-	-	-	-	0.5412	0.5665	-	-
NMT−style	(Zhouet al., 2019) ⋆	-	-	-	-	0.5474	0.5947	-	-
NULL−Ins	(Rubinoet al., 2020)	-	-	-	-	0.533	0.609	0.110	**0.162**
Gen−Det	(Cuiet al., 2021)	0.5719	-	**0.101**	**0.1572**	0.5508	-	**0.1125**	0.1633
-	XLM−R基线	0.5086	0.5656	0.1132	0.1837	0.5028	0.5891	0.1216	0.1884
-	our wwm mask	**0.5767**	**0.6045**	0.1134	0.1788	**0.5582**	0.6008	0.1169	0.1734
-	our fully mask	0.5536	0.5914	0.1142	0.1799	0.5341	0.6023	0.1167	0.1807
-	our mixed mask	0.5739	0.6030	0.1072	0.1597	0.5494	**0.6112**	0.1141	0.1758

Table 4: WMT 2019数据集句子集的性能比较，其中⋆表示集成结果。

预训练策略	方法	句子级 验证集				句子级 测试集			
		r↑	p↑	MAE↓	RMSE↓	r↑	p↑	MAE↓	RMSE↓
NMT−style	(Fanet al., 2020)⋆	-	-	-	-	0.7159	0.7402	0.0965	0.1384
NMT−style	(Hou Qi, 2019)⋆	-	-	-	-	0.703	-	0.1007	0.1377
Gen−Det	(Cuiet al., 2021)	0.7278	-	0.0995	0.1497	0.7245	-	**0.0971**	**0.1352**
-	XLM−R基线	0.6745	0.7086	0.1249	0.1722	0.6545	0.6820	0.1385	0.1889
-	our wwm mask	0.7182	0.7394	0.1197	0.1412	0.7148	0.7415	0.1126	0.1554
-	our fully mask	**0.7421**	**0.7670**	**0.0908**	0.1368	**0.7302**	**0.7578**	0.1022	0.1472
-	our mixed mask	0.7351	0.7662	0.1007	**0.1341**	0.7294	0.7549	0.1019	0.1461

Table 5: WMT 2017数据集句子集的性能比较，其中⋆表示集成结果。

由于词级别的粒度更小，相对句子级别而言更难，因此近年来同时做句子级和词级的工作相对较少，词级别实验结果分别如表6和表7所示。

通过对比可以发现本文所提出的预训练策略具有很大的优势，其中fully mask在WMT2017句子级数据集上取得了最高的单模型性能，而wwm mask在WMT2019句子级数据集上取得了最高的单模型性能，混合的mixed mask方法除去WMT2019句子级任务在其他任务上都超过了现有的单模型最高性能。

预训练策略	方法	验证集	测试集
		F1-Multi ↑	F1-Multi↑
QE−Bert	(Kimet al., 2019)⋆	**0.4443**	0.3960
Gen−Det	(Cuiet al., 2021)	0.4063	0.3971
-	XLM−R基线	0.3703	0.3704
-	our wwm mask	0.4047	0.4028
-	our fully mask	0.3884	0.3870
-	our mixed mask	0.4304	**0.4270**

Table 6: WMT 2019数据词级性能比较，其中⋆表示集成结果。

预训练策略	方法	验证集	测试集
		F1-Multi↑	F1-Multi↑
QE−Bert	(Kimet al., 2019)⋆	-	0.5513
Gen−Det	(Cuiet al., 2021)	0.5816	-
-	XLM−R基线	0.5418	0.5217
-	our wwm mask	0.5764	0.5914
-	our fully mask	0.5734	0.5857
-	our mixed mask	**0.5853**	**0.5934**

Table 7: WMT 2017数据词级性能比较，其中⋆表示集成结果。

(3)集成方法

本文集成三种预训练方法所得到的模型输出，通过简单的加权平均方式来集成，结果如表8所示。从结果上看，除了WMT2017验证集上相比fully mask有些许下降，在其它任务上相比单模型都有提升。

数据集	句子级(p)		词级别(F1-Multi)	
	验证集	测试集	验证集	测试集
WMT2017	0.7416	0.7361	0.6176	0.6073
WMT2019	0.5793	0.5609	0.4486	0.4309

Table 8: 在WMT 2017和WMT 2019句子级和词级数据上的集成结果

7 实验分析

为了分析预训练策略对跨语言预训练模型XLM-R的影响，本文设计了两个实验对比预训练前后的性能变化：

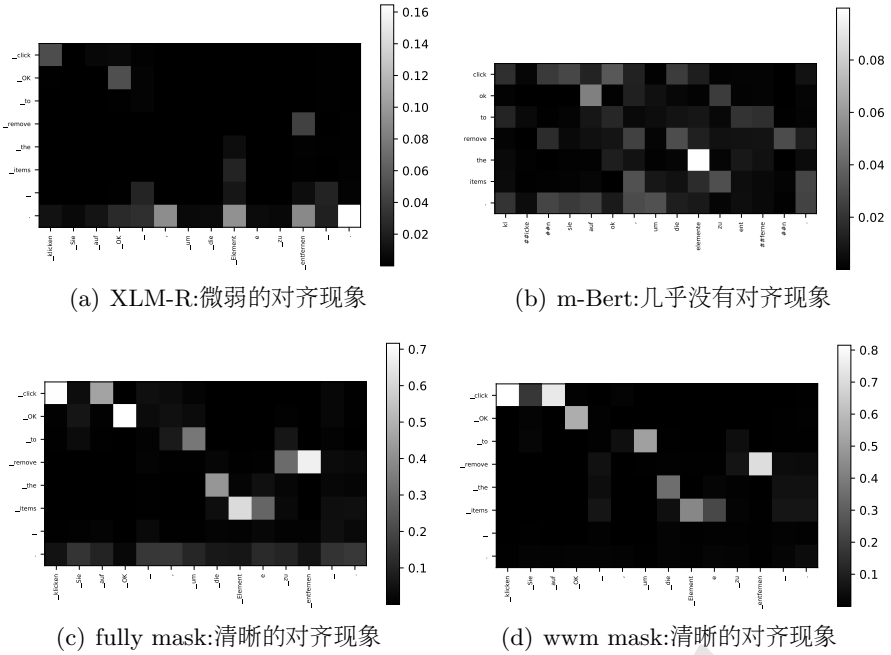
实验一:跨语言预训练模型的词对齐性能

预训练任务的目标为让模型结合源端和目标端上下文还原出被掩码的部分，这项任务非常考验模型的语义关联能力。评估模型语义关联能力的一个非常直观的方法就是观察模型自注意力模块的权重分布，好的注意力权重分布应具有较强的可解释性（如表9所示）。

click OK to remove the items .
klicken Sie auf OK , um die Element e zu entfernen .

Table 9: 一个在英德平行语料上的词对齐例子，其中源端词remove（移除）和目标端词entfernen（移除）的注意力权值较大

图4为本文对注意力的可视化热力图，其中包含了XLM−R和m−Bert，以及本文两种预



(a) XLM-R:微弱的对齐现象　(b) m-Bert:几乎没有对齐现象

(c) fully mask:清晰的对齐现象　(d) wwm mask:清晰的对齐现象

Figure 4: XLM−R，m−Bert以及本文所提出的两种跨语言预训练策略训练后的编码器的注意力头权值分布热力图

训练策略预训练后的XLM−R模型。从热力图可以看出原始的XLM−R只有微弱的双向对齐现象，而m−Bert不具有对齐现象。这种对齐现象是跨语言预训练模型双语理解能力的一个侧面反应，而双语理解能力对下游QE任务非常重要。

实验二:跨语言预训练模型的还原能力

为了量化本文所提出的预训练策略给XLM−R模型带来的提升，本文在译文质量评估数据集上测试模型的翻译理解能力。具体做法是保持源端句子不动，利用本文所提出的"目标端掩码方法"掩码目标端的全词，评价指标为还原出被替换部分的准确率。其结果如表10所示：

模型 \ 掩码概率	15%	20%	30%	40%	50%	60%	70%	80%	90%	100%
XLM−R	60.4%	56.8%	47.7%	37.5%	29.0%	23.0%	19.7%	18.0%	17.5%	17.0%
wwm mask	86.6%	84.7%	80.9%	76.5%	71.5%	64.8%	56.6%	46.8%	37.1%	29.0%
fully mask	86.3%	84.2%	80.5%	76.1%	71.1%	64.4%	56.7%	47.7%	38.4%	29.9%
mixed mask	92.5%	92.1%	90.3%	87.1%	80.2%	69.2%	57.0%	45.2%	35.3%	26.8%

Table 10: 在WMT2019质量评估数据集上模型的掩码还原准确率。

从表10可以看出模型具备的双语知识，当掩码概率增加到100%时，"完形填空"任务就演变成了"机器翻译"任务。三种预训练策略相比基线XLM−R模型都携带更多的"双语翻译信息"。同时，经过预训练之后，对于不同掩码难度的译文，模型的掩码还原能力会有一个线性衰减过程。而XLM−R在掩码概率超过50%之后，还原能力的衰减就不再明显。

8 结论

本文根据译文质量评估任务的特点设计了fully mask，wwm mask和mixed mask三种预训练策略。通过在XLM−R的基础上继续预训练，使预训练模型具有更好的语义关联能力。在WMT2017和WMT2019译文质量评估数据集上的实验都证明了本文所提出的方法的有效性，其中在单模型的条件下，mixed mask 能综合fully mask 和wwm mask两者的优点，在WMT2017和WMT2019大部分任务的测试集上都达到了新的单模型最高性能。本文简单的加权平均集成方式融合三个模型的输出，除了在WMT2017验证集句子级任务上有些许下降，在

其它任务上都有提升。如何更好地简单地集成模型输出是本文未来研究的内容。本文的分析实验也侧面反应增强语义关联能力对译文质量评估任务的重要性。

参考文献

Ilya Sutskever,Oriol Vinyals,Quoc V. Le. 2014. *Sequence to Sequence Learning with Neural Networks.* CoRR,abs/1409.3215.

Dzmitry Bahdanau,Kyunghyun Cho ,Yoshua Bengio. 2015. *Neural Machine Translation by Jointly Learning to Align and Translate.* 3rd International Conference on Learning Representations.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation.* In Proceedings of ACL, pages 311–318.

Michael Denkowski and Alon Lavie. 2002. *Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.* In Proceedings of ACL, pages 65–72.

Chin-Yew Lin. 2004. *Rouge: A package for automatic evaluation of summaries.* Text summarization branches out, pages 74-81.

Blatz, John and Fitzgerald, Erin and Foster, George and Gandrabur, Simona and Goutte, Cyril and Kulesza, Alex and Sanchis, Alberto and Ueffing, Nicola. 2004. *Confidence estimation for machine translation.* Coling 2004: Proceedings of the 20th international conference on computational linguistics,pages 315-321

Specia, Lucia and Shah, Kashif and De Souza, José GC and Cohn, Trevor. 2013. *QuEst-A translation quality estimation framework.* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations,pages 79-84

Martins, André FT and Junczys-Dowmunt, Marcin and Kepler, Fabio N and Astudillo, Ramón and Hokamp, Chris and Grundkiewicz, Roman. 2017. *Pushing the limits of translation quality estimation.* Transactions of the Association for Computational Linguistics,pages 205-218

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. *Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation.* In Proceedings of the Second Conference on Machine Translation (WMT).

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; and Makhoul, J. 2006. *A study of translation edit rate with targeted human annotation.* In Proceedings of Association for Machine Translation in the Americas.

Cui, Qu and Huang, Shujian and Li, Jiahuan and Geng, Xiang and Zheng, Zaixiang and Huang, Guoping and Chen, Jiajun. 2021. *DirectQE: Direct Pretraining for Machine Translation Quality Estimation.* In Proceedings of the AAAI Conference on Artificial Intelligence.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need.* In Advances in Neural Information Processing Systems, pages 6000–6010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.,Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual Language Model Pretraining.* arXiv preprint arXiv:1907.11692, 2019.

Ilya Loshchilov and Frank Hutter. 2017. *Fixing weight decay regularization in Adam.* arXiv preprint arXiv:1907.11692, 2019.

Ganesh Jawahar, Benoˆıt Sagot, and Djame Seddah. 2017. *What does BERT learn about the structure of language?* In 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy.

Guillaume Lample and Alexis Conneau. 2019. *What does BERT learn about the structure of language?* arXiv preprint arXiv:1901.07291.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised cross lingual representation learning at scale.* In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. *Neural machine translation of rare words with subword units.* arXiv preprint arXiv:1508.07909, 2015.

Erick Fonseca, Lisa Yankovskaya, Andre F. T. Martins, Mark Fishel, and Christian Federmann. 2015. *Findings of the WMT 2019 shared tasks on quality estimation.* In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 1–10, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and SeungHoon Na. 2019. *Qe bert: Bilingual bert using multi-task learning for neural quality estimation.* In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 85–89.

Hou Qi. 2019. *NJU Submissions for the WMT19 Quality Estimation Shared Task.* In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 95–100.

Junpei Zhou,Zhisong Zhang,Zecong Hu. 2019. *SOURCE: SOURce-Conditional Elmo-style Model for Machine Translation Quality Estimation.* In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 106-111.

Raphael Rubino and Eiichiro Sumita. 2020. *Intermediate Self-supervised Learning for Machine Translation Quality Estimation.* International Committee on Computational Linguistics.

Kai Fan,Jiayi Wang,Bo Li,Fengming Zhou,Boxing Chen and Luo Si. 2019. *"Bilingual Expert" Can Find Translation Errors.* The Thirty-Third AAAI Conference on Artificial Intelligence , pages 6367-6374.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya,Frederic Blain, Francisco Guzman, Mark Fishel,Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. *Unsupervised Quality Estimation for Neural Machine Translation.* arXiv preprint arXiv:2005.10608.

Frédéric Blain,Nikolaos Aletras and Lucia Specia. 2020. *Quality In, Quality Out: Learning from Actual Mistakes.* Proceedings of the 22nd Annual Conference of the European Association for Machine Translation,pages 145-153.

Julia Ive,Frédéric Blain and Lucia Specia. 2020. *deepQuest: A Framework for Neural-based Quality Estimation.* Proceedings of the 27th International Conference on Computational Linguistics,pages 3146-3157.

Maoxi Li,Qingyu Xiang,Zhiming Chen and Mingwen Wang. 2018. *A Unified Neural Network for Quality Estimation of Machine Translation.* IEICE Trans,pages 2417-2421.

Guoyi Miao,Hui Di,Jinan Xu,Zhongcheng Yang,Yufeng Chen and Kazushige Ouchi. 2019. *Improved Quality Estimation of Machine Translation with Pre-trained Language Representation.* Natural Language Processing and Chinese Computing - 8th CCF International Conference,pages 406-417.

Dzmitry Bahdanau,Kyunghyun Cho and Yoshua Bengio. 2014. *Neural Machine Translation by Jointly Learning to Align and Translate.* 3rd International Conference on Learning Representations.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, Luo Si. 2019. *"Bilingual Expert" Can Find Translation Errors* AAAI 2019: 6367-6374.