融合XLM词语表示的神经机器译文自动评价方法

胡纬　　李茂西*　　裘白莲　　王明文
江西师范大学 计算机信息工程学院 江西 南昌 330022
Email:{huwei, mosesli, mwwang}@jxnu.edu.cn, qiubl@ecjtu.edu.cn

摘要

机器译文自动评价对机器翻译的发展和应用起着重要的促进作用，它一般通过计算机器译文和人工参考译文的相似度来度量机器译文的质量。该文通过跨语种预训练语言模型XLM将源语言句子、机器译文和人工参考译文映射到相同的语义空间，结合分层注意力和内部注意力提取源语言句子与机器译文、机器译文与人工参考译文以及源语言句子与人工参考译文之间差异特征，并将其融入到基于Bi-LSTM神经译文自动评价方法中。在WMT'19译文自动评价数据集上的实验结果表明，融合XLM词语表示的神经机器译文自动评价方法显著提高了其与人工评价的相关性。

关键词： 机器翻译 ； 译文自动评价 ； 跨语种预训练语言模型 ； 差异特征

Neural Automatic Evaluation of Machine Translation Method Combined with XLM Word Representation

Wei Hu　　Maoxi Li*　　Bailian Qiu　　Mingwen Wang
School of Computer Information Engineering, Jiangxi Normal University
Nanchang, 330022, China
Email:{huwei, mosesli, mwwang}@jxnu.edu.cn, qiubl@ecjtu.edu.cn

Abstract

The automatic evaluation of machine translation plays an important role in promoting the development and application of machine translation. It generally measures the quality of machine translation through calculating the similarity between machine translation and its reference. This paper uses the cross-lingual language model XLM to map source sentences, machine translations and reference to the same semantic space, and combines layer-wise attention and intra attention to extract the difference features from source sentences and machine translations, machine translations and its references, source sentences and its references, then integrates them into the automatic evaluation method based on neural network Bi-LSTM. The experimental results on the dataset of WMT'19 Metrics task show that the neural automatic evaluation method of machine translation combined with XLM word representation significantly improves its correlation with human judgments.

Keywords: machine translation , automatic evaluation of machine translation , cross-lingual language model , difference features

基金项目： 国家自然科学基金(61662031)
通信作者： 李茂西(mosesli@jxnu.edu.cn)

1 引言

机器译文自动评价方法是机器翻译研究的直接推动力。它极大地促进了机器翻译的研究和系统开发。一方面，译文自动评价结果方便用户选择更好的翻译系统；另一方面，译文自动评价结果能够使系统开发者及时地了解翻译性能，以便开发更好的翻译系统(Shiwen, 1993)(李良友et al., 2014)。

机器译文自动评价方法大都是通过对比机器翻译系统的输出译文和人工参考译文来定量计算译文的质量。BLEU(Papineni et al., 2002)、模糊匹配的BLEU(刘洋et al., 2005)、NIST(Doddington, 2002)、METEOR(Banerjee and Lavie, 2005)、METEOR-SD-Makov(张丽林et al., 2017)和TERp(Snover et al., 2008)等基于词语匹配统计信息的方法使用词形、词根和同义词等信息对机器译文和人工参考译文进行对比计算译文质量；基于句法(姚建民et al., 2004; Popović, 2015)、语义结构匹配(Lo, 2017; Zhu et al., 2010)的方法使用词语的词性知识、句子的短语结构树、依存结构树和语义角色标注信息等等对机器译文和人工参考译文进行对比计算译文质量。近年来，随着深度神经网络在自然语言中的成功应用，许多学者将词语的分布式表示应用在译文自动评价中，包括基于静态词向量的方法(Chen and Guo, 2015)和基于动态上下文词向量的方法(Mathur et al., 2019)。

然而，当前神经译文自动评价方法均只在目标语言的深度语义空间对比机器译文和人工参考译文，评价时不仅缺乏源语言句子的对照参考，而且没有在同一语义空间对比源语言句子和机器译文的语义差异。针对这个问题，本文尝试使用跨语种预训练语言模型XLM(Lample and Conneau, 2019)将源语言句子、机器译文和人工参考译文映射到同一语义空间以计算差异特征：人工参考译文和机器译文构成的深度语义信息反映了同语种下机器译文语义与真实语义之间的差异；源语言句子和机器译文构成的深度语义信息反映了不同语种下机器译文语义与真实语义之间的差异；源语言句子和人工参考译文构成的深度语义信息作为评价的黄金参考。为使提取的句子表征充分考虑不同网络层、不同词语位置所包含的深度语义信息，我们分别在XLM模型表示的纵向和横向上使用分层注意力(Rei et al., 2020)和内部注意力，将得到的表征向量与黄金参考进行逐元素相减、相乘等操作以增强表示，获取差异特征，并将差异特征融入机器译文自动评价中以指导译文自动评价。在WMT'19译文自动评价数据集上与现有模型进行对比实验，结果表明融合XLM词语表示的神经机器译文自动评价方法在句子级和系统级任务上均显著提高了机器翻译自动评价与人工评价之间的相关性。

2 相关工作

在基于静态词向量的神经机器译文自动评价中，Chen和Guo使用word2vec(Mikolov et al., 2013)静态词向量表征机器译文和人工参考译文中的词语，并通过启发式的方法计算两者在词级别目标语言语义空间中的相似度(Chen and Guo, 2015)；Gupta等人提出利用树结构长短时记忆网络(Tree-LSTM)将机器译文和人工参考译文的词级别Glove静态词向量表征编码为句子级别表征，并以两者句子表征的乘与差逐元素操作的结果作为前馈神经网络的输入计算译文的质量(Gupta et al., 2015)。

近年来，BERT(Devlin et al., 2018)、GPT(Radford et al., 2018)等使用大规模数据进行训练的预训练语言模型被相继提出，使得直接利用句子向量进行机器译文自动评价的方法成为可能。RUSE(Shimanaka et al., 2018)使用预训练的InferSent(Conneau et al., 2017)、Quick-Thought(Logeswaran and Lee, 2018)以及Universal Sentence Encoder(Cer et al., 2018)作为编码器获取句子向量，再通过多层感知机回归器预测机器译文质量。BERT regressor(Shimanaka et al., 2019)则使用更先进的预训练语言模型BERT(Devlin et al., 2018)代替RUSE中的三种句子向量编码器，并与多层感知机回归器一起进行微调。Mathur等人(Mathur et al., 2019)首先使用BERT提取的动态词向量，并将其输入Bi-LSTM模型中进一步学习机器译文和人工参考译文的句子向量，最后将两者间的交互程度用于机器译文质量评价。然而机器翻译是一项开放式任务，对于同一个源语言句子可能存在多个不同的正确翻译。这些方法使用的单一人工参考译文仅能代表一种可能的翻译，不能正确评价所有正确的候选译文(Fomicheva et al., 2020)。Qin(Qin and Specia, 2015)和Fomicheva等人(Fomicheva et al., 2020)通过引入多种参考译文来缓解这个问题，然而获取多种参考译文需要大量的人力。由于源语言句子与参考译文在语义上是等价的，Takahashi等人(Takahashi et al., 2020)提出通过引入源语言句子作为伪参考的方法、罗琪等人(Luo and Li, 2020)使用译文质量估计向量将源端信息引入模型。

与上述方法不同，本文使用跨语种预训练语言模型XLM(Lample and Conneau, 2019)获取源语言句子、机器译文和人工参考译文两两之间的深度语义信息，结合注意力机制提取它们的差异特征，并将得到的差异特征融入机器译文自动评价中，进一步提高机器翻译自动评价方法与人工评价之间的相关性。

3 背景知识

3.1 跨语种预训练语言模型(Cross-lingual Language Model, XLM)

近年来，使用大型语料库进行自监督学习的预训练语言模型，如OpenAI GPT(Radford et al., 2018)和BERT(Devlin et al., 2018)，在一些自然语言理解任务上取得了显著性突破。然而，这些模型仅在单语语料上进行自监督训练。这使得在不同语言任务上不仅需要进行多次训练，而且无法获得跨语言的信息。XLM(Lample and Conneau, 2019) 在BERT上进行改进：使用字节对编码(Sennrich et al., 2016)将子词编码独立于语言；加入语言嵌入层；在多语言平行语料库上使用翻译语言模型(Translation Language Modeling, TLM)进行预训练。在XNLI跨语言分类任务(Conneau et al., 2020)上，XLM取得了比多语言BERT(multilingual BERT, mBERT)(Devlin et al., 2018)更好的性能。

3.2 引入源端信息的机器译文自动评价方法

语境词向量方法将词语映射到一个语义空间中，具有相近含义的词语在这个空间中会获得较高的相似度。Mathur(Mathur et al., 2019)等人从“译文评价是计算机器译文和人工参考译文之间的相似度”的观点出发，将语境词向量空间中机器译文和人工参考译文之间的交互程度用于反映机器译文的质量。罗琪等人(Luo and Li, 2020)通过引入译文质量估计向量的方法将源端信息融入Mathur等人(Mathur et al., 2019)的模型中。

罗琪等人(Luo and Li, 2020)使用基于联合神经网络的模型(Unified Neural Network for Quality Estimation, UNQE)(Li et al., 2018)提取质量估计向量。将长度为l_s的源语言句子s和长度为l_t的机器译文t输入到UNQE模型中，得到词语级别的质量向量$e_{qe} = \{e^1_{qe}, e^2_{qe}, ..., e^{l_t}_{qe}\}$。最后，将词语级别的质量向量经过Bi-LSTM后进行最大、平均池化，再将其拼接得到句子级别的质量向量v_{qe}:

$$h_{qe} = \text{Bi-LSTM}(e_{qe}) \tag{1}$$

$$h^{max}_{qe} = \max_{i=1}^{l_t} h^i_{qe}, \quad h^{avg}_{qe} = \frac{1}{l_t}\sum_{i=1}^{l_t} h^i_{qe} \tag{2}$$

$$v_{qe} = [h^{max}_{qe}; h^{avg}_{qe}] \tag{3}$$

其中，$h_{qe} = \{h^1_{qe}, h^2_{qe}, ..., h^{l_t}_{qe}\}$ 是Bi-LSTM将词语级别的质量向量作为输入后得到的隐藏层向量，h^{max}_{qe}和h^{avg}_{qe}分别是h_{qe}经过最大、平均池化后的结果。

将长度为l_r的人工参考译文r和长度为l_t的机器译文t作为模型的输入，使用Mathur(Mathur et al., 2019)等人的$(\text{Bi-LSTM+attention})_{\text{BERT}}$和$(\text{ESIM})_{\text{BERT}}$分别得到$m_{att}$和$m_{esim}$，作为人工参考译文和机器译文的相互表示。罗琪等人(Luo and Li, 2020)的$(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$和$(\text{ESIM})_{\text{BERT+QE}}$分别将$m_{att}$、$m_{esim}$与$v_{qe}$拼接后得到的向量输入到前馈神经网络中，以计算译文质量的得分。相比没有引入源端信息的$(\text{Bi-LSTM+attention})_{\text{BERT}}$和$(\text{ESIM})_{\text{BERT}}$(Mathur et al., 2019)，$(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$和$(\text{ESIM})_{\text{BERT+QE}}$(Luo and Li, 2020)在WMT'19译文自动评价任务数据集上与人工评分的相关性更高，证明了源端信息在译文自动评价任务中的有效性。

4 融合XLM词语表示的神经机器译文自动评价

4.1 注意力层

预训练语言模型最后一层中首个位置的输出向量通常作为下游任务的输入。然而，之前的研究(Tenney et al., 2019)表明，预训练语言模型编码器的每一层涵盖不同的语言学特征：底层关注词法信息，中间层关注句法信息，顶层关注语义信息。对于机器译文自动评价任务，各种

语言学特征都是评价机器译文质量的重要信息，Zhang等人(Zhang et al., 2019)表明仅使用最后一层通常会导致机器译文自动评价模型性能下降。此外，如果仅使用跨语种预训练语言模型首个位置的输出向量，在一定程度上容易丢失其他位置输出向量所包含的跨语言信息。

为了解决这两个问题，本文在纵向上使用分层注意力机制(Rei et al., 2020)以融合各层次语言学特征，并在横向上使用内部注意力机制将首位置的输出向量与所有位置的平均向量进行加权求和，以获取包含各层次语言学特征的深度语义信息。

我们将源语言句子src、机器译文mt和人工参考译文ref两两拼接组成三组句子对分别输入到XLM模型中："$src+ref$"和"$src+mt$"表示由源语言到目标语言的句子对，"$ref+mt$"表示由同一个源语言句子产生的两个目标语言句子的组合。以长度为l_{src+mt} 的句子对"$src+mt$"为例，注意力层结构如图1所示。

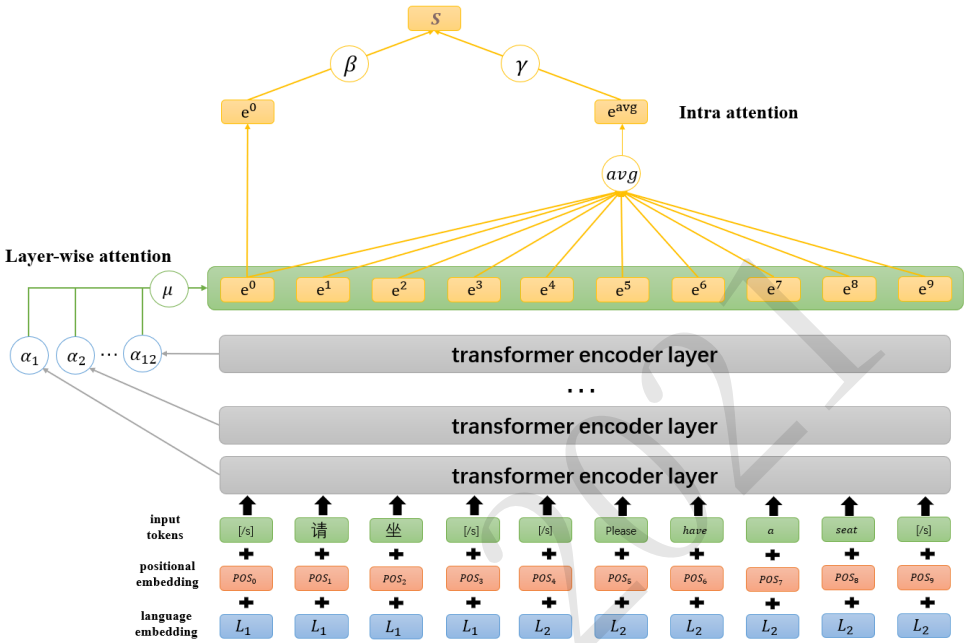

图 1. 注意力层结构

将XLM每一层的隐藏层向量作为XLM的输出，通过分层注意力机制对这些隐藏层向量含有的各种语言学特征进行融合，得到e_{x_j}:

$$e_{x_j} = \mu H_j^{\top} \alpha \tag{4}$$

其中，$j \in \{0,1,...,l-1\}$，l为句子对的长度，μ是可学习的参数，$H_j = [h_j^{(1)}, h_j^{(2)}, ..., h_j^{(12)}]$为XLM输入句子中第$j$个位置对应的所有隐藏层向量(共有12层)，$\alpha = softmax([\alpha_1, \alpha_2, ..., \alpha_{12}])$是可学习的分层注意力权重。

我们将分层注意力机制的输出$e_{src+mt} = [e^0_{src+mt}, e^1_{src+mt}, ..., e^{l_{src+mt}-1}_{src+mt}]$中首个位置的向量$e^0_{src+mt}$和所有位置平均池化后的向量$e^{avg}_{src+mt}$通过内部注意力机制得到句子对"$src+mt$"的表征向量$s_{src+mt}$:

$$e^{avg}_{src+mt} = \frac{1}{l_{src+mt}} \sum_{k=0}^{l_{src+mt}-1} e^k_{src+mt} \tag{5}$$

$$s_{src+mt} = \beta_{src+mt} e^0_{src+mt} + \gamma_{src+mt} e^{avg}_{src+mt} \tag{6}$$

其中，β_{src+mt}和γ_{src+mt}为可学习的权重参数。"$src+ref$"和"$ref+mt$"对应的句子对向量计算过程与此类似，分别得到$s_{src+ref}$和s_{ref+mt}。

将通过注意力层得到的三个句子对表征向量s_{src+mt}、$s_{src+ref}$ 和s_{ref+mt}进行拼接，以获取跨语言特征空间中同语种和不同语种下的表征信息，并考虑同语义时不同语种间的差异。最

后，对s_{src+mt}与$s_{src+ref}$之间逐元素相减、相乘以突出s_{src+mt}与黄金参考$s_{src+ref}$之间线性与非线性的差异：

$$e_{dv} = [s_{src+mt}; s_{src+ref}; s_{ref+mt}; s_{src+mt} \odot s_{src+ref}; |s_{src+mt} - s_{src+ref}|] \tag{7}$$

我们将e_{dv}称为差异向量。

4.2 模型总体架构

为了提高自动评价方法的效果，我们把提取的差异向量融入前人提出的(Bi-LSTM+attention)$_{\text{BERT+QE}}$和(ESIM)$_{\text{BERT+QE}}$模型中，模型整体结构如图2所示。图左边描述由UNQE模型(Li et al., 2018)和Bi-LSTM网络提取出源语言句子和机器译文的词语级别质量向量，再通过池化层将其处理为句子级别的质量向量。图右边描述通过(Bi-LSTM+attention)$_{\text{BERT}}$或(ESIM)$_{\text{BERT}}$模型(Mathur et al., 2019)提取交互表示的增强向量。图中间部分使用跨语种预训练语言模型XLM(Lample and Conneau, 2019)作为特征提取器，将“$src + ref$”、“$src + mt$”和“$ref + mt$”分别映射到跨语言特征空间中，通过分层注意力和内部注意力获取跨语言信息并进行增强表示。最后将三个部分得到的向量进行融合并通过前馈神经网络得到机器译文的质量分数。

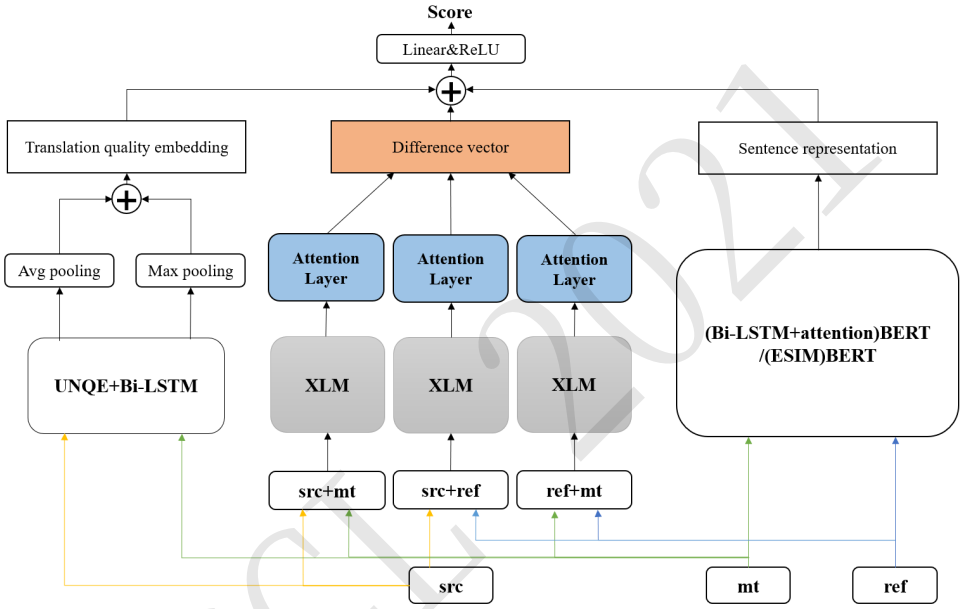

图 2. 融合XLM词语表示的神经机器译文自动评价方法模型总体结构

由式1-3可知模型的左边部分输出为句子级别的译文质量向量v_{qe}(Li et al., 2018; Luo and Li, 2020)。模型右边的输出为m_{att}或m_{esim}，具体细节见(Mathur et al., 2019)。最后将v_{qe}、e_{dv}和m_{att}或m_{esim}拼接得到的向量输入到前馈神经网络中，以预测译文质量分数：

$$\tilde{m} = [v_{qe}; e_{dv}; m_{att}] \tag{8}$$

或

$$\tilde{m} = [v_{qe}; e_{dv}; m_{esim}] \tag{9}$$

$$y_{\text{score}} = w^T ReLU(W^T \tilde{m} + b) + b' \tag{10}$$

其中参数w，W，b，b'均为前馈神经网络中可学习的权重。

我们使用模型最终输出的译文质量得分$\hat{y}$与人工评价得分y的均方误差作为损失函数：

$$loss = \frac{1}{M} \sum_{i=1}^{M} (\hat{y}^{(i)} - y^{(i)})^2 \tag{11}$$

其中M表示训练集包含的样本数量。

5 实验

5.1 实验设置

为了验证融合XLM词语表示的神经机器译文自动评价方法的有效性，我们在WMT'19 Metrics Task(Ma et al., 2019)的德英、中英和英中语言对的语料库上进行实验。表1展示了每种语言对的统计数据。

		de-en	zh-en	en-zh
WMT'19	systems	16	15	12
	sentences	2000	2000	1997
	sum	32000	30000	23964

表 1. WMT'19 Metrics task德英、中英和英中任务的测试集数据统计

对于德英语言对，我们使用WMT'15-17 Metrics task(Bojar et al., 2015; Bojar et al., 2016; Ondrej et al., 2017)德英语言对的句子级别任务数据集，其中训练集和开发集比例为9:1。由于WMT Metrics task 在中英和英中语言对的训练集样本数量过少，我们采用罗琪等人(Luo and Li, 2020)的方法：使用CWMT'18翻译质量评估在中英和英中语言对的语料用于模型训练，并将该语料库中的人工后编辑率($HTER$)处理为译文人工评分($1 - HTER$)。中英和英中任务完全按照CWMT'18翻译质量评估数据集给定的训练集和开发集进行训练，表2展示了每种语言对的训练集和开发集统计数据。

	de-en	zh-en	en-zh
训练集	1458	8785	12865
开发集	162	1064	1040

表 2. 德英、中英和英中训练集、开发集数据统计

我们将BLEU(Papineni et al., 2002)、chrF(Popović, 2015)、BEER(Stanojević and Sima'an, 2014)、仅使用跨语言模型的$hyp + src/hyp + ref$和$hyp + src + ref$方法(Takahashi et al., 2020)等作为基线方法，并将本文提出的方法与Mathur等人的方法(Mathur et al., 2019)以及罗琪等人的方法(Luo and Li, 2020)进行比较。遵循WMT'19 Metrics Task(Ma et al., 2019)中官方做法：使用肯德尔相关系数评价模型在句子级别上与人工评分的相关性，使用皮尔森相关系数评价模型在系统级别上与人工评分的相关性。

本文使用XLM-15(Lample and Conneau, 2019)作为特征提取器，隐藏层向量维度大小为1024。UNQE输出的译文质量向量维度在德英任务中为500，在中英和英中任务上为700。模型中包含的Bi-LSTM隐藏层向量维度大小均为300。$(\text{Bi-LSTM+attention})_{\text{BERT}}$和$(\text{ESIM})_{\text{BERT}}$(Mathur et al., 2019)均使用"bert-base-uncased"提取英文语境词向量、"bert-base-chinese"提取中文语境词向量。使用Adam优化器优化模型参数，初始学习率为0.0004。

5.2 实验结果

表3展示了在WMT'19 Metrics Task的德英、中英和英中任务上各种自动评价方法与人工评价的句子级别相关性。本文提出的融合XLM词语表示的神经机器译文自动评价方法"$(\text{Bi-LSTM+attention})_{\text{BERT+QE+DV}}$"和"$(\text{ESIM})_{\text{BERT+QE+DV}}$"在三个语言对上与人工评分的句子级别相关性均远超过UNQE、sentBLEU等基线模型。仅使用跨语言模型的$hyp + src/hyp + ref$和$hyp+src+ref$方法(Takahashi et al., 2020)也具有一定的竞争性。"$(\text{ESIM})_{\text{BERT+QE+DV}}$"相比罗琪等人(Luo and Li, 2020)未融合XLM词语表示的方法"$(\text{ESIM})_{\text{BERT+QE}}$"在德英、中英以及英中任务上分别提升了38.9%、3.2% 和0.6%；"$(\text{Bi-LSTM+attention})_{\text{BERT+QE+DV}}$"相比"$(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$"在德英、中英以及英中任务上分别提升了26.3%、3.4% 和1.7%。这表明通过融合XLM词语表示的方法可以有效提升机器译文自动评价与人工评价之间的句子级别相关性。

	de-en	zh-en	en-zh	avg.
UNQE	0.011	0.243	0.258	0.171
sentBLEU	0.056	0.323	0.270	0.216
BEER	0.128	0.371	0.232	0.244
chrF	0.122	0.371	0.301	0.265
hyp+src/hyp+ref	0.127	0.326	0.277	0.243
hyp+src+ref	0.094	0.318	0.256	0.223
(ESIM)$_{\text{BERT}}$	0.134	0.362	0.336	0.277
(Bi-LSTM+attention)$_{\text{BERT}}$	0.153	0.375	0.345	0.291
(ESIM)$_{\text{BERT+QE}}$	0.144	0.372	0.357	0.291
(Bi-LSTM+attention)$_{\text{BERT+QE}}$	0.160	0.387	0.358	0.302
(ESIM)$_{\text{BERT+QE+DV}}$	0.200	0.384	0.359	0.314
(Bi-LSTM+attention)$_{\text{BERT+QE+DV}}$	**0.202**	**0.400**	**0.364**	**0.332**

表 3. WMT'19 Metrics Task的德英、中英和英中任务上自动评价与人工评价的句子级别相关性

	de-en	zh-en	en-zh	avg.
UNQE	0.264	0.688	0.916	0.623
BLEU	0.849	0.899	0.901	0.883
BEER	0.906	0.942	0.803	0.884
chrF	0.917	0.956	0.880	0.918
hyp+src/hyp+ref	0.828	0.934	0.921	0.894
hyp+src+ref	0.855	0.946	0.892	0.898
(ESIM)$_{\text{BERT}}$	0.896	0.951	0.967	0.938
(Bi-LSTM+attention)$_{\text{BERT}}$	0.910	0.956	0.965	0.944
(ESIM)$_{\text{BERT+QE}}$	0.896	0.958	0.970	0.941
(Bi-LSTM+attention)$_{\text{BERT+QE}}$	**0.917**	**0.972**	0.965	**0.951**
(ESIM)$_{\text{BERT+QE+DV}}$	0.911	0.966	**0.973**	0.950
(Bi-LSTM+attention)$_{\text{BERT+QE+DV}}$	0.908	**0.972**	**0.973**	**0.951**

表 4. WMT'19 Metrics Task的德英、英中和中英任务上自动评价与人工评价的系统级别相关性

表4展示了在WMT'19 Metrics Task的德英、中英和英中任务上各种自动评价方法与人工评价的系统级别相关性。本文提出的融合XLM词语表示的神经机器译文自动评价方法"(Bi-LSTM+attention)$_{\text{BERT+QE+DV}}$"和"(ESIM)$_{\text{BERT+QE+DV}}$"在三个语言对上与人工评分的系统级别相关性超过了所有基线模型。同时，"(ESIM)$_{\text{BERT+QE+DV}}$"在所有语言对上均高于对应的罗琪等人(Luo and Li, 2020)未融合XLM词语表示的方法"(ESIM)$_{\text{BERT+QE}}$"，在德英、中英以及英中任务上与人工评价之间的系统级别相关性分别提升了1.7%、0.8%和0.3%；"(Bi-LSTM+attention)$_{\text{BERT+QE+DV}}$"相比"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"，在中英任务上保持一致，在英中任务上提升了0.8%。这表明通过融合XLM词语表示的神经机器译文自动评价方法有助于提升机器译文自动评价与人工评价之间的系统级别相关性。

5.3 实验分析

为了定性说明所提出方法的效果，在中英语言对开发集中抽取了一个实例以分析融合XLM词语表示的神经机器译文自动评价方法的特点。

在表5实例中，机器译文将源语言句子中"让权力在阳光下运行"翻译成"let power run in the sunshine"。但通过对比源语言句子和人工参考译文"power is exercised in a transparent manner"，可以发现对于相同语义，不同语种间的表达存在一定的差异，而机器译文并没有表达源语言句子的内在含义。相比未融合XLM词语表示的(Bi-

src：加强党内监督、民主监督、法律监督、舆论监督，让人民监督权力，**让权力在阳光下运行**。	
mt：We should strengthen inner-party supervision, democratic oversight, legal supervision and public opinion supervision, so that the people can supervise power and **let power run in the sunshine**.	
ref：We should strengthen inner-party supervision, democratic oversight, legal supervision and public opinion supervision, so that the people can supervise the exercise of power and that **power is exercised in a transparent manner**.	
人工打分(1-HTER)：0.750	
(Bi-LSTM+attention)$_{\text{BERT+QE}}$ 得分：0.740	(ESIM)$_{\text{BERT+QE}}$得分：0.827
(Bi-LSTM+attention)$_{\text{BERT+QE+DV}}$得分：0.754	(ESIM)$_{\text{BERT+QE+DV}}$得分：0.779

表 5. 不同自动评价方法对机器译文打分实例

LSTM+attention)$_{\text{BERT+QE}}$和(ESIM)$_{\text{BERT+QE}}$ 方法，本文所提方法的打分均更接近于人工评分。通过这个实例表明，融合XLM词语表示的神经机器译文自动评价方法能够充分考虑源语言句子、人工参考译文以及机器译文之间的差异信息，更好地评价机器译文质量。

6 结论

本文提出融合XLM词语表示的神经机器译文自动评价方法。与现有方法相比，融合XLM词语表示的神经机器译文自动评价方法能够充分考虑源语言句子、人工参考译文以及机器译文之间的差异，与人工评价具有更高的相关性。在未来工作中，将尝试在更深层次上挖掘源语言句子、人工参考译文以及机器译文之间的语义差异，进一步提高译文自动评价的性能。

参考文献

姚建民, 周明, 赵铁军, and 李生. 2004. 基于句子相似度的机器翻译评价方法及其有效性分析. 计算机研究与发展, 41(7):1258–1265.

刘洋, 刘群, and 林守勋. 2005. 机器翻译评测中的模糊匹配. 中文信息学报, 19(3):46–54.

李良友, 贡正仙, and 周国栋. 2014. 机器翻译自动评价综述. 中文信息学报, 28(3):81–91.

张丽林, 李茂西, 肖文艳, 万剑怡, and 王明文. 2017. 机器翻译自动评价中领域知识复述抽取研究. 北京大学学报（自然学科版）, 53(2):230–238.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the WMT*, pages 1–46.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the WMT*, pages 131–198.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Boxing Chen and Hongyu Guo. 2015. Representation based translation evaluation metrics. In *Proceedings of the ACL and IJCNLP*, pages 150–155.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R Bowman. 2020. Xnli: Evaluating cross-lingual sentence representations. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the NAACL*, page 4171–4186.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Proceedings of the HLT*, pages 138–145.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Maoxi Li, Qingyu Xiang, Zhiming Chen, and Mingwen Wang. 2018. A unified neural network for quality estimation of machine translation. *IEICE TRANSACTIONS on Information and Systems*, 101(9):2417–2421.

Chi-kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the second conference on machine translation*, pages 589–597.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Qi Luo and Maoxi Li. 2020. Research on incorporating the source information to automatic evaluation of machine translation. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 414–423.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Bojar Ondrej, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the WMT*, pages 169–214.

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

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Ying Qin and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 113–120.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv preprint arXiv:1907.12679*.

Yu Shiwen. 1993. Automatic evaluation of output quality for machine translation systems. *Machine translation*, 8(1):117–126.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2008. Terp system description. In *MetricsMATR workshop at AMTA*, pages 104–108.

Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Junguo Zhu, Muyun Yang, Bo Wang, Sheng Li, and Tiejun Zhao. 2010. All in strings: a powerful string-based automatic mt evaluation metric with multiple granularities. In *Coling 2010: Posters*, pages 1533–1540.