引入源端信息的机器译文自动评价方法研究

罗琪　　李茂西*
江西师范大学 计算机信息工程学院 / 江西 南昌 330022
Email:{luoqi，mosesli}@jxnu.edu.cn

摘要

机器译文自动评价是机器翻译中的一个重要任务。针对目前译文自动评价中完全忽略源语言句子信息，仅利用人工参考译文度量翻译质量的不足，该文提出了引入源语言句子信息的机器译文自动评价方法：从机器译文与其源语言句子组成的二元组中提取描述翻译质量的质量向量，并将其与基于语境词向量的译文自动评价方法利用深度神经网络进行融合。在WMT'19译文自动评价任务数据集上的实验结果表明，所提出的方法能够有效增强机器译文自动评价与人工评价的相关性。深入的实验分析进一步揭示了源语言句子信息在译文自动评价中发挥着重要的作用。

关键词： 机器翻译；译文自动评价；质量向量；语境词向量；自然语言推断

Research on Incorporating the Source Information to Automatic Evaluation of Machine Translation

Qi Luo　　**Maoxi Li***
School of Computer Information Engineering, Jiangxi Normal University
Nanchang, 330022, China
Email:{luoqi，mosesli}@jxnu.edu.cn

Abstract

Automatic evaluation of machine translation is one of the most critical tasks in machine translation. However, the source sentence information is completely ignored and only the reference is used to measure the translation quality in previous work. For this shortcoming, the paper presents a novel automatic evaluation metric incorporating the source information: extracting the quality embeddings that describes the translation quality from a tuple consist of the machine translations and their corresponding source sentences, and incorporating it into the automatic evaluation method based on contextual embeddings by using a deep neural network. The experimental results on the dataset of WMT'19 Metrics task show that the proposed method can effectively enhance the correlation between the results of the automatic evaluation metrics and that of the human judgments. Deep analysis further reveals that the information of the source sentences plays an important role in automatic evaluation of machine translation.

Keywords: machine translation , automatic evaluation of machine translation , quality embeddings , contextual embeddings , natural language inference

基金项目：国家自然科学基金(61662031, 61462044)

0 引言

机器译文自动评价是机器翻译的重要组成部分。它不仅能一定程度上度量翻译系统的整体性能，还能在翻译系统开发时指导其特征权值的优化。因此，研究机器译文自动评价对机器翻译的发展和应用有着重要的意义。

近年来，许多机器译文自动评价方法被相继提出，它们将机器翻译系统的输出译文与人工参考译文进行对比来定量刻画译文的质量。根据对比时涉及的语言知识层次，它们分为基于词语匹配的方法，如BLEU(Papineni et al., 2002)和NIST(Doddington, 2002)等；基于浅层句法结构匹配的方法，如POSBLEU(Popović and Ney, 2009)和POSF(Popović and Ney, 2009)等；基于深层语义信息匹配的方法，如引入复述的指标Meteor Universal(Banerjee and Lavie, 2005)和TERp(Snover et al., 2008)等、引入语义角色标注的指标MEANT(Lo, 2017)等等。随着深度学习的发展和其在自然语言处理中的广泛应用，一些研究者利用词语深度表示和神经网络结构对比翻译系统输出译文和人工参考译文进行译文自动评价，如基于静态词向量word2vec(Mikolov et al., 2013)的方法(Chen and Guo, 2015)、基于动态词向量BERT(Devlin et al., 2018)的方法(Mathur et al., 2019)、和基于神经网络结构的方法ReVal(Gupta et al., 2015)和RUSE(Shimanaka et al., 2018)等等。

然而，这些方法评价机器译文的主要思路还是遵循BLEU(Papineni et al., 2002)的基本观点："机器译文越接近于人工参考译文，其译文质量越高"。从这个观点出发，译文自动评价即是计算机器译文和人工参考译文的相似度。因此，译文自动评价过程中完全忽略了源语言句子，即在没有对源语言句子充分利用的基础上进行译文的自动评价。所以，找到结合源语言句子进行译文自动评价的切入点，势必能提高译文自动评价与人工评价的相关性。因此，我们尝试引入从源语言句子和其机器译文中提取的质量向量（Quality Embedding, QE），并将其与基于语境词向量的译文自动评价方法(Mathur et al., 2019)进行深度融合来增强译文自动评价，提高译文自动评价与人工评价的相关性。

1 相关工作

在基于深度神经网络的机器译文自动评价中，Lo(2017)，和Chen(2015)等人提出利用词语的分布式表示，静态预训练的词向量word2vec(Mikolov et al., 2013)，来提高机器译文和人工参考译文对比时同义词、近义词和复述等匹配的准确率。Guzmán(2019)等人提出了一种基于词向量和神经网络的机器译文自动评价方法，其目标是在给定人工参考译文的情况下，从一对机器译文中选择最佳译文，使用神经网络可以方便地融合由词向量捕获的丰富语法和语义表示。Gupta(2015)等人用基于树结构的长短时记忆网络(Tai et al., 2015)（Long Short-Term Memory network, LSTM）对机器译文和人工参考译文进行编码，根据两者之间元素差异和夹角计算机器译文的质量得分。Shimanaka(2018) 等人使用双向LSTM（Bidirectional LSTM, Bi-LSTM）对机器译文和人工参考译文进行编码，并利用多层感知机回归模型计算机器译文的质量得分。Mathur(2019)等人基于BERT(Devlin et al., 2018)语境词向量使用Bi-LSTM网络结构学习机器译文和人工参考译文的句子表示，并将自然语言推理中启发式方法(Mou et al., 2015)和增强序列推理模型(Chen et al., 2016)（Enhanced Sequential Inference Model, ESIM）引入到机器译文自动评价中，该方法在WMT'19译文自动评价任务（Metrics Task）上取得了优异的成绩，因此，本文将在Mathur(2019)等人的工作基础上，将利用源语言句子提取的质量向量融入译文自动评价中，进一步增强译文自动评价的性能。

2 背景知识

2.1 基于语境词向量的译文自动评价

自然语言推断关注假设结论（hypothesis）是否可以从前提语句（premise）中推断获取，它与译文自动评价任务非常类似。译文的质量越好，机器译文被人工参考译文表示（推断）的程度越高，同时人工参考译文被机器译文表示（推断）的程度也越高；反之亦然。在自然语言推断的框架下，Mathur(2019)等人使用语境词向量分别表示机器译文和人工参考译文，并根据两个表示的交互程度来度量机器译文的质量。使用自然语言推断中启发式方法(Mou et al., 2015)以及ESIM方法(Chen et al., 2016)，Mathur等人分别提出了$(\text{Bi-LSTM+attention})_{\text{BERT}}$译文自动评价方法和$(\text{ESIM})_{\text{BERT}}$译文自动评价方法。

第十九届中国计算语言学大会论文集，第414页-第423页，海南，中国，2020年10月30日至11月1日。

2.1.1 (Bi-LSTM+attention)$_{\text{BERT}}$译文自动评价方法

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分别利用BERT语境词向量进行表示，使用Bi-LSTM网络对其进行编码得到人工参考译文和机器译文包含上下文含义的新的向量表示$h_{r_{1:x}}(x=1\dots l_r)$、$h_{t_{1:y}}(y=1\dots l_t)$，通过向量点积求得人工参考译文和机器译文的相似度矩阵A，A中元素$a_{i,j}=h_{r_i}^T h_{t_j}$，利用相似度矩阵$A$，结合$h_r$和$h_t$，计算人工参考译文和机器译文的相互表示：

$$\tilde{h}_t=\sum_{i=1}^{l_r}\frac{\exp(a_{i,j})}{\sum_j\exp(a_{i,j})}\cdot h_r \tag{1}$$

$$\tilde{h}_r=\sum_{j=1}^{l_t}\frac{\exp(a_{i,j})}{\sum_i\exp(a_{i,j})}\cdot h_t \tag{2}$$

其中符号$\tilde{h}_t$表示h_r中每个词与h_t的相关程度，$\tilde{h}_r$表示h_t中每个词与h_r的相关程度。

为了避免向量$\tilde{h}_t$和$\tilde{h}_r$简单求和容易导致结果对序列长度敏感的问题(Chen et al., 2016)，对向量$\tilde{h}_t$和$\tilde{h}_r$分别进行最大池化和平均池化，将池化结果分别拼接得到向量v_t和v_r，并且启发式方法(Mou et al., 2015)被用作对局部推理进行增强得到增强后的表示向量m:

$$m=[v_t\oplus v_r\oplus(v_t\odot v_r)\oplus(v_t-v_r)] \tag{3}$$

其中符号“$\oplus$”表示向量拼接操作；符号“$\odot$”表示两个向量逐元素相乘操作。最后向量m被作为前馈神经网络的输入用于预测机器译文被人工参考译文表示的程度，即译文质量的得分。

2.1.2 (ESIM)$_{\text{BERT}}$译文自动评价方法

ESIM方法利用式(4)和(5)计算机器译文被人工参考译文表示的增强向量m_t和人工参考译文被机器译文表示的增强向量m_r。为降低模型参数复杂性，利用一个前馈神经网络层将m_t和m_r转换至模型的维度。Bi-LSTM网络被用作对降维后的信息进行编码，以便得到其局部信息的上下文表示向量。将编码后的向量进行平均池化和最大池化，并将池化后的结果$v_{r,avg}$、$v_{r,max}$和$v_{t,avg}$、$v_{t,max}$进行拼接，形成固定长度向量p，即：

$$m_t=[h_t\oplus\tilde{h}_t\oplus(h_t\odot\tilde{h}_t)\oplus(h_t-\tilde{h}_t)] \tag{4}$$

$$m_r=[h_r\oplus\tilde{h}_r\oplus(h_r\odot\tilde{h}_r)\oplus(h_r-\tilde{h}_r)] \tag{5}$$

$$p=[v_{r,avg}\oplus v_{r,max}\oplus v_{t,avg}\oplus v_{t,max}] \tag{6}$$

最后向量p被作为前馈神经网络的输入用于预测机器译文质量的得分。

2.2 译文质量向量提取方法

译文质量向量是译文质量估计中描述翻译质量的向量，它从源语言句子和其相应的译文中抽取，完全不需要借助人工参考译文进行计算。目前主流的质量向量提取方法包括基于循环神经网络（Recurrent Neural Network, RNN）的编码器-解码器模型(Bahdanau et al., 2014)的方法(Kim et al., 2017; Li et al., 2018)和基于Transformer模型(Vaswani et al., 2017)的方法(Fan et al., 2019; Wang et al., 2019)。它们将源语言句子和其机器译文使用强制学习的方式输入已训练好的神经机器翻译模型，截取在使用前馈神经网络进行$softmax$分类前一层网络的输出向量，作为机器译文当前位置词语的质量向量。

给定源语言句子，为了获取机器译文中每个词语的质量向量，基于联合神经网络的模型（Unified Neural Network for Quality Estimation, UNQE）(Li et al., 2018)被用作提取质量向量。联合神经网络模型使用译文质量估计任务数据集联合训练基于RNN的编码器-解码器模型和基于RNN的预测器，可以提取更优的质量向量，并且该模型在WMT18句子级别质量估计任务中取得了优异的成绩(Specia et al., 2018)，证实了其效果。

3 结合质量向量的机器译文自动评价

为了把源语言句子信息引入译文自动评价中，我们以质量向量作为切入点，将给定源语言句子情况下机器译文质量的表示和给定人工参考译文情况下机器译文的增强表示进行融合。模型结构如图1所示，其中符号src、mt和ref分别表示源语言句子、机器译文和人工参考译文。图左边描述通过UNQE方法(Li et al., 2018)从源语言句子和其机器译文中提取出描述翻译质量的词语级质量向量，并将其利用Bi-LSTM网络处理成句子级别的质量向量；图右边描述通过$(\text{Bi-LSTM+attention})_{\text{BERT}}$或$(\text{ESIM})_{\text{BERT}}$方法(Mathur et al., 2019)将机器译文和人工参考译文抽象为交互表示的增强向量，图上表示将质量向量与交互表示的增强向量进行拼接，将拼接后的向量输入前馈神经网络以预测机器译文质量得分。

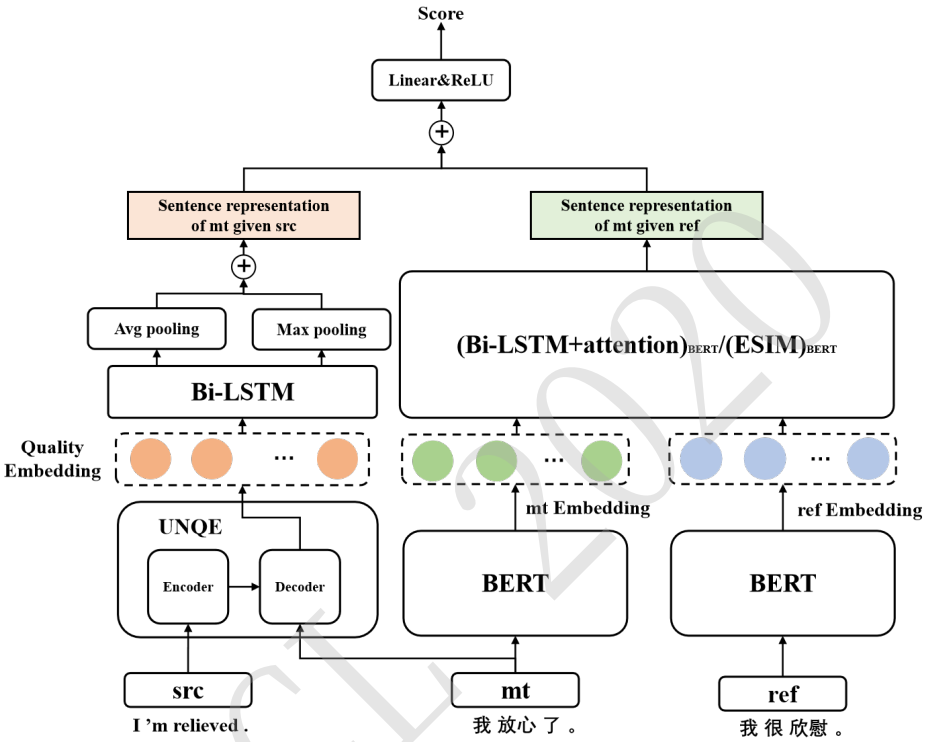

图 1. 引入译文质量向量增强机器译文自动评价的模型架构

3.1 $(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$译文自动评价方法

由于从源语言句子和机器译文中抽取的质量向量是词语级的，即机器译文中每个词（token）使用一个实数向量描述其翻译质量，而机器译文和人工参考译文的交互表示增强向量是句子级，为了在同一层次将二者进行融合，需要将质量向量进一步抽象成句子级别表示。Bi-LSTM网络被用来对词语级质量向量$e_{qe_{1:k}}(k=1\ldots l_t)$进行编码，得到$e_{qe_{1:k}}$的包含上下文信息的向量$h_{qe,k}(k=1\ldots l_t)$，通过对$h_{qe}$进行最大池化和平均池化处理，将池化后的结果拼接即得到了句子的质量向量表示v_{qe}:

$$h_{qe,k} = \text{Bi-LSTM}\left(e_{qe}, k\right), \forall k \in [1, \ldots, l_t] \tag{7}$$

$$v_{qe,max} = \max_{k=1}^{l_t} h_{qe,k}, \quad v_{qe,avg} = \frac{1}{l_t}\sum_{k=1}^{l_t} h_{qe,k} \tag{8}$$

$$v_{qe} = [v_{qe,avg} \oplus v_{qe,max}] \tag{9}$$

其中符号$v_{qe,avg}$表示对h_{qe}进行平均池化后的结果，$v_{qe,max}$表示对h_{qe}进行最大池化后的结果，k表示句子中的词序号。

在机器译文和人工参考译文的交互表示增强向量方面，Bi-LSTM网络被用来对人工参考译文和机器译文的语境词向量编码，利用式(1)-(2)求得人工参考译文和机器译文的相互表示，随后利用式(8)的池化操作和式(9)的拼接操作求得了人工参考译文句子表示v_r和机器译文句子表示v_t。

为了将源端信息有效地引入机器译文自动评价模型中，我们将v_r和v_t进行局部信息增强组合，同时将增强后的信息与式(9)处理后的句子级别质量向量v_{qe}拼接起来形成新的固定长度向量$\tilde{m}$：

$$\tilde{m} = [v_t \oplus v_r \oplus (v_t \odot v_r) \oplus (v_t - v_r) \oplus v_{qe}] \tag{10}$$

其中符号v_t是$\tilde{h}_t$的平均池化后向量$v_{t,avg}$和最大池化后向量$v_{t,max}$拼接后形成的向量；v_r是$\tilde{h}_r$的平均池化后向量$v_{r,avg}$和最大池化后向量$v_{r,max}$拼接后形成的向量。最后将向量$\tilde{m}$作为前馈神经网络的输入，使用其预测译文的质量得分：

$$y_{\text{score}} = w^T \,\text{ReLU}\left(W^T \tilde{m} + b\right) + b' \tag{11}$$

其中参数w，W，b，b'均为前馈神经网络的权值。

为了训练模型的所有参数，译文自动评价得分y_{score}与人工评价得分h的均方误差被用来对模型进行优化，优化目标正式描述为：

$$loss = \frac{1}{M}\sum_{i=1}^{M}(y_{\text{score}}^{(i)} - h^{(i)})^2 \tag{12}$$

其中$y_{score}^{(i)}$为自动评价方法对待评价机器译文的打分，$h^{(i)}$为人工评价结果，M为训练集包含的样本数量。

3.2 $(\text{ESIM})_{\text{BERT+QE}}$译文自动评价方法

为了控制译文自动评价模型的复杂性，将对式(4)和(5)得到的机器译文和人工参考译文的局部信息表示m_t、m_r使用一个映射F转换至模型的维度后，经过Bi-LSTM进行编码，以得到其局部信息的上下文表示向量$\tilde{m}_t$和$\tilde{m}_r$，如式(13)-(14)所示。为了引入源端信息增强机器译文自动评价，我们将$\tilde{m}_t$和$\tilde{m}_r$平均池化和最大池化后的向量与机器译文质量估计的$v_{qe,avg}$和$v_{qe,max}$向量拼接得到新的信息组合向量$\tilde{p}$。将拼接后的信息表示向量作为前馈神经网络的输入以预测机器译文的质量分数：

$$\tilde{m}_{t,i} = \text{Bi-LSTM}\left(F\left(m_{t,i}\right), i\right), \forall i \in [1, \ldots, l_t] \tag{13}$$

$$\tilde{m}_{r,j} = \text{Bi-LSTM}\left(F\left(m_{r,j}\right), j\right), \forall j \in [1, \ldots, l_r] \tag{14}$$

$$\tilde{p} = [\tilde{v}_{r,avg} \oplus \tilde{v}_{r,max} \oplus \tilde{v}_{t,avg} \oplus \tilde{v}_{t,max} \oplus v_{qe,avg} \oplus v_{qe,max}] \tag{15}$$

$$y_{\text{score}} = w^T \,\text{ReLU}\left(W^T \tilde{p} + b\right) + b' \tag{16}$$

其中符号i、j均表示词序号，F表示激活函数为$ReLU$的单层前馈神经网络层；式(15)中的$\tilde{v}_{t,avg}$和$\tilde{v}_{t,max}$向量分别是$\tilde{m}_t$平均池化和最大池化的向量，$\tilde{v}_{r,avg}$和$\tilde{v}_{r,max}$分别是$\tilde{m}_r$平均池化和最大池化的向量；式(16)中的w，W，b，b'均为该前馈神经网络模型的参数。同样，模型的优化目标也在训练集上最小化译文自动评价得分y_{score}与人工评价得分h的均方差，同式(12)所示。

获取了机器译文句子级别分值后，我们对整个测试集（或文档集）中机器译文的句子级别得分取平均值作为翻译系统的系统级别（或文档级别）得分。

4 实验

4.1 实验设置

为了验证引入源端信息的机器译文自动评价方法的效果，我们在WMT'19 Metrics Task(Ma et al., 2019)的德英任务、中英任务和英中任务上进行实验。为了比较不同译文自动评价方法的性能，我们遵循WMT评测官方的做法利用皮尔森相关系数与肯德尔相关系数分别计算自动评价结果和人工评价结果的系统级别相关性和句子级别相关性，皮尔森相关系数或肯德尔相关系数越大，相关性越好。

UNQE提取的中英、英中任务上的质量向量维度为700，德英任务上质量向量维度为500。模型中Bi-LSTM隐藏层状态维度均固定为300，Dropout设置为0.2，使用Adam优化器优化训练，初始学习率为0.0004，训练批次大小为32，使用"bert-base-uncased"提取英文句子语境词向量，使用"bert-base-chinese"提取中文句子语境词向量。

在实验中，我们不仅将本文提出的方法与BLEU(Papineni et al., 2002)、chrF(Popović, 2015)以及BEER(Stanojević and Sima'an, 2014)等经典的方法进行了比较，而且与Mathur(2019)等人提出的自动评价方法、与不使用人工参考译文的译文质量估计方法UNQE(Li et al., 2018)进行了对比。需要说明的是Mathur等人是混合所有相同目标语言（比如德英和中英）译文自动评价训练集语料进行模型训练，而我们引入了源端信息，考虑实际译文打分需求且避免受不同源语言差异性的负面影响，我们针对每个语言对利用其训练集数据单独训练模型。德英语言对使用的是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。对于中英和英中语言对而言，单独训练可用训练集语料规模太小，因此加入了CWMT'18翻译质量评估在中英和英中语言对上的语料。德英方向按照9：1比例划分训练集和开发集，中英和英中方向完全使用CWMT'18翻译质量评估数据的训练集和开发集，具体数据统计如表1所示。测试集为WMT'19 Metrics Task的数据集，具体数据统计如表2所示。

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

表 1. 德英、中英和英中训练集、开发集数据统计

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

表 2. WMT'19 Metrics task德英、中英和英中任务的测试集数据统计

4.2 实验结果

表3和表4分别给出了在WMT'19 Metrics task上引入源语言句子信息的译文自动评价方法与对比的译文自动评价方法与人工评价的句子级别和系统级别的相关性。

表3的数据表明引入源语言句子信息的方法"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"和"(ESIM)$_{\text{BERT+QE}}$"在德英、中英和英中三个语言对上，与人工评价的句子级别相关性均值分别高于使用语境词向量的方法"(Bi-LSTM+attention)$_{\text{BERT}}$"和"(ESIM)$_{\text{BERT}}$"。"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"相对于"(Bi-LSTM+attention)$_{\text{BERT}}$"在德英、中英、英中三个任务上分别提升了4.6%、3.2%和3.8%，"(ESIM)$_{\text{BERT+QE}}$"相对于"(ESIM)$_{\text{BERT}}$"方法分别提升了7.5%、2.8%和6.3%。其中"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"方法在三个语言对任务中句子级别相关系数均是最高。这说明引入源端信息能增强机器译文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	0.011	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
(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**

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
(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**

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的数据表明本文所提方法"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"和"(ESIM)$_{\text{BERT+QE}}$"在德英、中英和英中三个语言对评测任务上，与人工评价的系统级别相关系数的均值分别高于"(Bi-LSTM+attention)$_{\text{BERT}}$"和"(ESIM)$_{\text{BERT}}$"。"(Bi-LSTM+attention)$_{\text{BERT+QE}}$"相对于"(Bi-LSTM+attention)$_{\text{BERT}}$"方法在德英、中英任务上提升了0.8%和1.7%，在英中任务上保持一致，"(ESIM)$_{\text{BERT+QE}}$"相对于"(ESIM)$_{\text{BERT}}$"方法在中英、英中任务上分别提升了0.7%和0.3%，在德英上保持一致。这说明引入源端信息能增强机器译文自动评价与人工评价的系统级别相关性。

令人惊奇的是仅使用源端信息，完全不使用人工参考译文的UNQE方法也与人工评价结果有较好的相关性。尽管其在平均相关性上劣于所有使用人工参考译文的方法，但是它与sentBLEU方法在平均句子级别相关性和平均系统级别相关性上差距并不大，在英中的句子级别相关性(0.258)上甚至稍高于BEER方法(0.232)，在英中的系统级别相关性(0.916)上高于BLEU(0.901)、BEER(0.803)、chrF(0.880)等方法。这说明了源端信息对译文自动评价非常有帮助，从一个侧面佐证了正确地将质量向量引入译文自动评价必将提高译文自动评价的性能。

4.3 实验分析

为了进一步分析融合源端信息的译文自动评价方法的特点，我们在开发集上分别抽取了中英和英中翻译自动评价的实例进行分析。表5给出了对两个译文进行打分的实例，其中HTER是指将机器译文mt转换成人工后编辑的参考译文ref需要的最少编辑次数与译文长度的比值，它可以看作是译文人工打分的结果。自动评价方法对机器译文的打分越接近人工打分(1-HTER)，表明该自动评价方法对译文的评价越准确。

在第一个实例中，源语言句子中"**对城市交通来说**"在机器译文中缺乏对应翻译，存在漏译的情况，但(Bi-LSTM+attention)$_{\text{BERT}}$和(ESIM)$_{\text{BERT}}$却给了很高的分值，而本文的方法打分均更接近人工HTER分值。说明(Bi-LSTM+attention)$_{\text{BERT+QE}}$和(ESIM)$_{\text{BERT+QE}}$方法结合了源语言句子信息对译文进行评价，能更准确地描述译文的完整度特征，因此，相比于仅

结合人工参考译文信息打分的(Bi-LSTM+attention)$_{\text{BERT}}$和(ESIM)$_{\text{BERT}}$方法，引入源端信息的方法其评价更准确。在第二个实例中，机器译文中存在多译、过度翻译的情况，源语言句子中"*Tokyo*, *Japan*"被过度翻译成"**东京**"和"**日本**"两个地方。对于这种情况，本文方法依然比(Bi-LSTM+attention)$_{\text{BERT}}$和(ESIM)$_{\text{BERT}}$更接近人工打分结果HTER。这定性的说明了结合源端信息的机器译文自动评价方法能更充分利用源语言句子的信息对译文质量进行评价。

src：如此规模的城市发展**对城市交通来说**既是挑战，也是机遇。	
mt：This scale of urban development urban traffic is both a challenge and an opportunity.	
ref：This scale of urban development urban traffic is both a challenge and an opportunity to urban transportation.	
人工打分(1-HTER)：0.833	
(Bi-LSTM+attention)$_{\text{BERT}}$得分：0.883	(ESIM)$_{\text{BERT}}$得分：0.862
(Bi-LSTM+attention)$_{\text{BERT+QE}}$得分：0.833	(ESIM)$_{\text{BERT+QE}}$得分：0.845
src: The African Development Conference was dominated by Japan, and the previous five meetings were held in Tokyo, Japan or Yokohama, so this meeting will be the first move to Africa.	
mt: 非洲发展会议由日本主导，前五次会议分别在东京、**日本**或横滨举行，因此这次会议将是第一次到非洲的会议。	
ref：非洲开发会议由日本主导，此前的五次会议均是在日本东京或者横滨举行，因此，本次会议也将是首次移师非洲。	
人工打分(1-HTER)：0.836	
(Bi-LSTM+attention)$_{\text{BERT}}$得分：0.705	(ESIM)$_{\text{BERT}}$得分：0.904
(Bi-LSTM+attention)$_{\text{BERT+QE}}$得分：0.888	(ESIM)$_{\text{BERT+QE}}$得分：0.879

表 5. 不同自动评价方法对机器译文打分实例

5 结论

本文提出引入源端信息的机器译文自动评价方法。与传统的BLEU、BEER、chrF等评价指标相比，引入源端信息的机器译文自动评价方法，融合了源语言句子、人工参考译文、机器译文三者的信息，能更全面更有效地描述译文质量。在未来的工作中，我们将尝试在更大的语料库、更多的语言对上进行实验，以及引入更先进的模型和方法来挖掘源端信息，以提高机器译文自动评价方法的性能。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the ICLR*, pages 1–15.

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.

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

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. In *Proceedings of the ACL*, page 1657–1668.

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.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. "bilingual expert" can find translation errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6367–6374.

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.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2019. Pairwise neural machine translation evaluation. In *Proceedings of the ACL and IJCNLP*, pages 805–814.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–22.

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.

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.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the ACL*, pages 130–136.

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 and Hermann Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32.

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.

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.

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

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

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.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the ACL and IJCNLP*, page 1556–1566.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ziyang Wang, Hui Liu, Hexuan Chen, Kai Feng, Zeyang Wang, Bei Li, Chen Xu, Tong Xiao, and Jingbo Zhu. 2019. Niutrans submission for ccmt19 quality estimation task. In *China Conference on Machine Translation*, pages 82–92.