文章编号：1003-0077(2020)03-0056-08

融合 BERT 语境词向量的译文质量估计方法研究

李培芸，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）

摘 要：蕴含语义、句法和上下文信息的语境词向量作为一种动态的预训练词向量，在自然语言处理的下游任务中有着广泛应用。然而，在机器译文质量估计中，没有相关研究工作涉及语境词向量。该文提出利用堆叠双向长短时记忆网络将 BERT 语境词向量引入神经译文质量估计中，并通过网络并联的方式与传统的译文质量向量相融合。在 CWMT18 译文质量估计评测任务数据集上的实验结果表明，融合中上层的 BERT 语境词向量均显著提高了译文质量估计与人工评价的相关性，并且当对 BERT 语境词向量的最后 4 层表示平均池化后引入译文质量估计中对系统性能的提高幅度最大。实验分析进一步揭示了融合语境词向量的方法能利用译文的流利度特征来提高翻译质量估计的效果。

关键词：神经译文质量估计；语境词向量；循环神经网络；编码器—解码器网络；质量向量

中图分类号：TP391 **文献标识码**：A

Integrating BERT Word Embedding into Quality Estimation of Machine Translation

LI Peiyun, LI Maoxi, QIU Bailian, WANG Mingwen

(School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China)

Abstract: The word embedding of BERT contains semantic, syntactic and context information, pre-trained for a various downstream tasks of natural language processing. We propose to introduce BERT into neural quality estimation of MT outputs by employing stacked BiLSTM (bidirectional long short-term memory), concatenated with the existing the quality estimation network at the output layer. The experiments on the CWMT18 datasets show that the quality estimation can be significantly improved by integrating upper and middle layers of the BERT, with the top-improvement brought by average pooling of the last four layers of the BERT. Further analysis reveals that the fluency in translation is better exploited by BERT in the MT quality estimation task.

Keywords: neural quality estimation of machine translation; contextual word embedding; recurrent neural network; encoder-decoder network; quality vector

0 引言

机器翻译是利用计算机把一种语言自动转换成另一种语言的过程[1-3]。在相关研究中，机器输出译文质量的自动估计是推动机器翻译快速发展与应用的关键环节。译文质量估计方法（quality estimation of machine translation，QE）是在没有人工参考译文对照的情况下对机器译文质量自动进行打分[4]。与使用且依赖人工参考译文的译文自动评价方法（automatic evaluation of machine translation）相比，其使用限制少，更方便、快捷。

然而，仅仅给出源语言句子和其待评价的机器输出译文，在缺乏人工参考译文作对照的情况下，如何定量地估计翻译质量呢？学者们提出多种利用外部语言资源抽取描述翻译质量的特征，然后使用机器学习算法估计翻译质量的方法。

Specia 等结合大规模的单语语料和训练翻译系

收稿日期：2019-08-22 **定稿日期**：2019-10-17

基金项目：国家自然科学基金(61662031,61462044,61876074)

统的双语平行语料等语言资源，利用语义、句法等分析工具从待翻译的源语言句子及其机器译文中提取包括流利度指标、忠实度指标和翻译复杂度指标等特征，然后利用支持向量机回归算法估计翻译质量。该方法简称为 QuEst[4]，QuEst 能有效利用外部语言资源抽取译文的句法和语义信息，但是其抽取过程严重依赖语言学分析，且与目标语言种类相关，缺乏通用性。为了解决这个问题，Shah 等[5]、陈志明等[6-7]提出使用大规模单语语料训练词向量[8]，利用句子的平均池化词向量来提取反映翻译质量的特征。尽管上述方法不需要语言学分析，但是源语言端和目标语言端词向量独立训练，没有考虑到源语言词和目标语言词之间的对应关系。

近年来，许多学者提出了基于深度神经网络的译文质量估计方法[9-14]，即神经译文质量估计，其从双语平行语料中利用编码器—解码器模型抽取描述翻译对应关系的分布式表示，称为质量向量[9]，并将质量向量输入循环神经网络或带自注意力机制的 Transformer 结构[15]中计算译文质量得分。神经译文质量估计方法在近三年的 WMT QE 评测中取得了最好的成绩，受到了研究者的广泛关注。但是，其只使用描述翻译对应关系的质量向量，即反映翻译忠实度的特征，进行译文质量估计，没有结合反映翻译流利度的特征，即机器译文的上下文信息，进行译文质量估计。

语境词向量方法作为一种有效的语言建模途径，其利用双向长短时记忆网络（bidirectional long short-term memory，BiLSTM）或 Transformer 解码器网络从大规模单语语料中学习蕴含丰富的句法、语义和上下文信息的词语分布式表示，其中 BERT 语境词向量根据屏蔽语言建模任务（masked language modeling，MLM）和下句预测任务（next sentence prediction，NSP）进行训练，使用其在多个自然语言处理任务上刷新了最好的性能记录，包括推理、问答和命名实体识别等任务[16]。为了克服神经译文质量估计方法忽略翻译流利度特征的不足，我们尝试将语境词向量作为引入机器译文上下文信息的一个切入点，利用堆叠 BiLSTM 将其引入神经译文质量估计网络结构中。在 CWMT18 译文质量估计评测任务上的实验结果表明，引入 BERT 语境词向量的神经译文质量估计方法显著优于对比的 UNQE 系统[10]和参与评测的最优译文质量估计系统。

1　背景知识

1.1　基于双向 RNN 编码器—解码器的质量向量提取

带有注意力机制的双向循环神经网络（recurrent neural network，RNN）编码器—解码器模型在神经机器翻译中得到了广泛应用[2-3]。该模型将长度为 n 的源语言句子 $x_{1:n}$ 映射为长度为 m 的目标语言句子 $y_{1:m}$。其使用编码器将源语言句子 $x_{1:n}$ 抽象表示为上下文向量 $c_t=\mathrm{RNN}^{\mathrm{enc}}(x_{1:n})(t=1,\cdots,n)$，随后在解码器端使用 RNN 语言模型和注意力机制，根据当前已预测出的词以及编码的上下文向量 c_t 来生成目标序列 $y_{1:m}$，如式(1)所示。

$$p(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t},\boldsymbol{x}) = \partial(\boldsymbol{y}_{t-1},\boldsymbol{h}_{t-1},\boldsymbol{c}_t) = \frac{\exp(\boldsymbol{y}_t^T\boldsymbol{W}\boldsymbol{r}_t)}{\sum_{k=1}^{K_y}\exp(\boldsymbol{y}_k^T\boldsymbol{W}\boldsymbol{r}_t)} \tag{1}$$

其中，∂ 是一个非线性映射函数；h_{t-1} 表示 RNN 的上一隐含层状态；c_t 表示源语言 x_t 的上下文向量；$\boldsymbol{y}_t\in R^{K_y\times 1}$ 为目标词 $\boldsymbol{y}_t$ 的 one-hot 编码表示；K_y 表示目标语言词汇量的大小；$\boldsymbol{W}$ 是一个权重矩阵；$\boldsymbol{r}_t\in R^{d\times 1}$ 是一个中间表示；d 表示目标语言的词向量维度。

$\boldsymbol{r}_t$ 可由式(2)推导得出：

$$\boldsymbol{r}_t = \tanh(\boldsymbol{W}_0\boldsymbol{h}_{t-1}+\boldsymbol{V}_0\boldsymbol{E}y_{t-1}+\boldsymbol{C}\boldsymbol{c}_t) \tag{2}$$

其中，$\boldsymbol{W}_0$、$\boldsymbol{V}_0$、$\boldsymbol{C}$ 均表示模型参数；$E\in R^{d\times K_y}$ 表示目标语言的词向量矩阵。

已训练的编码器—解码器模型中，目标词的条件概率 $p(\boldsymbol{y}_t|y_{<t},x)$ 中包含了译文中词 y_t 的翻译质量信息，即目标词 y_t 是否被正确翻译。Li 等提出使用式(3)计算描述翻译质量的质量向量[10]。

$$\boldsymbol{q}_{y_t} = [(\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{W}_0)\odot\boldsymbol{r}_t^{\mathrm{T}}]^{\mathrm{T}},t=1,\cdots,m \tag{3}$$

其中，符号⊙表示逐元素相乘操作。

1.2　BERT 语境词向量

在语言建模任务中，Devlin 等提出了 BERT 语境词向量预训练方法[16]，该方法使用了多层双向 Transformer 编码器，并提出两个新的语言建模任务——遮挡语言模型和下句预测任务进行训练。WordPiece 子词切分方法被用来对训练语料中的句

子进行切分。在网络模型的输入中，模型使用了位置向量和句子切分向量。

在自然语言处理的下游任务中，BERT 预训练语言模型有两种应用策略：一是在微调阶段，根据不同任务对整个 BERT 模型的所有参数进行微调；二是使用 BERT 语言模型在不同自然语言处理任务的数据集上提取长度固定的语境词向量，将其当作特征应用到具体任务模型中。和传统的词向量相比，BERT 语境词向量蕴含了丰富的句法、语义和上下文信息，并且其是一个动态的词向量，同一个词在不同的上下文环境中，其语境词向量可能不同。

给定 BERT 语言模型 bert(*)，对输入序列 $T=(t_1,t_2,\cdots,t_c)$ 提取 BERT 语境词向量方法如式(4)所示。

$$v_{\text{bert}_i}^{L}=\text{bert}(f_{\text{token}}(T),L),L=-1,\cdots,-12;\ i=1,\cdots,c \tag{4}$$

其中，$f_{\text{token}}(*)$ 是符号化函数，用于对输入句子进行符号化处理(对中文而言该函数是一个分词器)。L 表示 Transformer 编码器网络的隐含层层号，指定由哪一隐含层生成 BERT 语境词向量(−1 表示 Transformer 网络的最后一个隐含层，−2，…，−12 依次类推)。本文使用第二种策略将 BERT 语境词向量应用到译文质量估计方法中。

2 融合 BERT 语境词向量和质量向量的译文质量估计方法

通过编码器—解码器模型提取的译文中每个词语的质量向量，描述了其与源语言句子的映射关系，反映了机器译文的翻译忠实度。但是，除了忠实度，机器译文质量的优劣还与其流利度相关。BERT 语言模型在大规模的单语语料上进行训练，给定具体的机器译文，由其推导的语境词向量描述了译文的流利度。鉴于这样的思路，我们研究融合 BERT 语境词向量和质量向量的译文质量估计方法。

为了结合 BERT 语境词向量与质量向量，我们将其分别输入两层堆叠的 BiLSTM 网络中，计算出描述整个译文的忠实度向量和流利度向量，然后将忠实度向量和流利度向量连接，输入单个节点的全连接层，使用 sigmoid 激活函数输出译文质量得分。模型的整体结构如图 1 所示，图中左下部分描述如何通过词语的质量向量计算译文的忠实度向量，图中右下部分描述如何通过词语的 BERT 语境词向量计算译文的流利度向量。需要说明的是，由于 BERT 语言模型方法对机器译文的切分粒度与编码器-解码器模型可能不同，因此我们不在网络的输入层对词语的 BERT 语境词向量与质量向量进行连接。

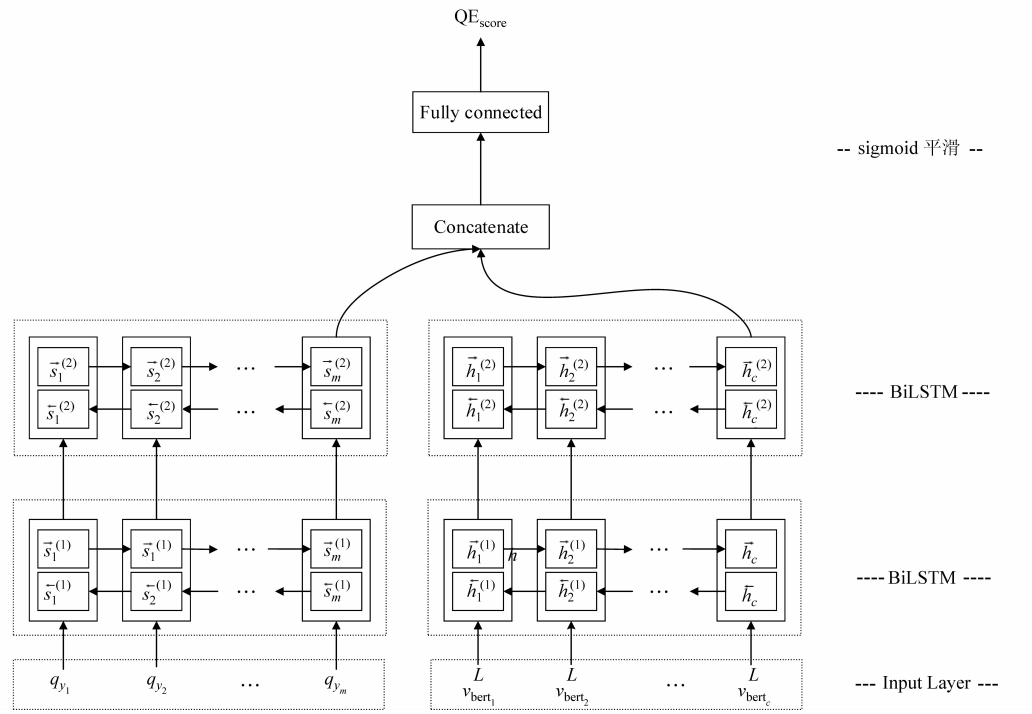

图 1 CUNQE 网络模型结构

我们将图 1 所示的模型称为结合上下文信息的神经译文质量估计模型(contextual UNQE，CUNQE)，其中，$\overrightarrow{s}_{1:m}^{(1)}$、$\overrightarrow{h}_{1:c}^{(1)}$ 和 $\overrightarrow{s}_{1:m}^{(2)}$、$\overrightarrow{h}_{1:c}^{(2)}$ 表示第一层网络及第二层网络中的前向隐含状态，$\overleftarrow{s}_{1:m}^{(1)}$、$\overleftarrow{h}_{1:c}^{(1)}$ 和 $\overleftarrow{s}_{1:m}^{(2)}$、$\overleftarrow{h}_{1:c}^{(2)}$ 表示对应的后向隐含状态。网络中词语的质量向量 $q_{y_i}(i=1,\cdots,m)$ 由式(3)计算，并将其输入到图左下部分两层堆叠的 BiLSTM 网络中，网络中词语的语境词向量 $v_{\text{bert}_i}^{L}(i=1,\cdots,c;L=-1,\cdots,-12)$ 由式(4)计算，并将其输入到图右下部分两层堆叠的 BiLSTM 网络中。由于 BERT 语境词向量包含不同隐含层的输出，为了调查引入不同层 BERT 语境词向量的输出对译文质量估计的影响，12 个隐含层的语境词向量：$v_{\text{bert}_i}^{-1}$、$v_{\text{bert}_i}^{-2}$、$v_{\text{bert}_i}^{-3}$、$v_{\text{bert}_i}^{-4}$、$\cdots$、$v_{\text{bert}_i}^{-12}$ 分别被引入神经译文质量估计方法中，简称为 CUNQE^{-1}，CUNQE^{-2}，CUNQE^{-3}，CUNQE^{-4}，$\cdots$，CUNQE^{-12}；同时，为了充分利用 BERT 语言模型中每个隐含层蕴含的上下文信息，我们对上述提取的语境词向量进行了平均池化运算，并将其结果作为特征引入到译文质量估计中，简称为 CUNQE$^{\text{AVG}}$，如式(5)所示。

$$v_{\text{bert}_i}^{\text{AVG}}=\frac{\sum v_{\text{bert}_i}^{-(1:n)}}{n},n=12 \tag{5}$$

根据谷歌发布的提取 BERT 语境词向量方法，我们也对－4 层、－3 层、－2 层、－1 层的语境词向量进行平均池化运算(由式(5)计算，$n=4$)，并将其结果作为特征引入到模型中，简称为 CUNQE$^{\text{AVG_4}}$。

CUNQE 模型前向推导过程如下：

(1) 左侧第一层 BiLSTM 的第 i 个位置的输出如式(6)所示。

$$\begin{aligned}\text{BiLSTM}_{q_{y_i}}^{(1)}(q_{y_{1:m}},i)=&[\text{LSTM}_{q_{y_i}}^{f_{(1)}}(q_{y_{1:i}})\oplus\\&\text{LSTM}_{q_{y_i}}^{b_{(1)}}(q_{y_{m:i}})]\boldsymbol{W}_q^{(1)}\\=&\,y_{q_{y_i}}^{(1)},\quad i=1,\cdots,m\end{aligned} \tag{6}$$

$\text{BiLSTM}_{q_{y_i}}^{(1)}(*)$ 表示左侧第一层 BiLSTM 网络，$\text{LSTM}_{q_{y_i}}^{f_{(1)}}(*)$、$\text{LSTM}_{q_{y_i}}^{b_{(1)}}(*)$ 分别为其对应的前向函数和后向函数，$\boldsymbol{W}_q^{(1)}$ 为权重矩阵。

(2) 以上一层输出为输入，左侧第二层 BiLSTM 的第 i 个位置的输出如式(7)、式(8)所示。

$$\begin{aligned}\text{BiLSTM}_{q_{y_i}}^{(2)}(y_{q_{y_{1:m}}}^{(1)},i)=&[\text{LSTM}_{q_{y_i}}^{f_{(2)}}(y_{q_{y_{1:i}}}^{(1)})\oplus\\&\text{LSTM}_{q_{y_i}}^{b_{(2)}}(y_{q_{y_{m:i}}}^{(1)})]\boldsymbol{W}_q^{(2)}\\=&\,y_{q_{y_i}}^{(2)},\quad i=1,\cdots,m\end{aligned} \tag{7}$$

$$y_{q_y}^{*}=y_{q_{y_m}}^{(2)} \tag{8}$$

$\text{BiLSTM}_{q_{y_i}}^{(2)}(*)$ 表示左侧第二层 BiLSTM 网络，$\text{LSTM}_{q_{y_i}}^{f_{(2)}}(*)$、$\text{LSTM}_{q_{y_i}}^{b_{(2)}}(*)$ 分别为其对应的前向函数和后向函数，$\boldsymbol{W}_q^{(2)}$ 为权重矩阵。

(3) 右侧第一层 BiLSTM 的第 j 个位置的输出如式(9)所示。

$$\begin{aligned}\text{BiLSTM}_{v_{\text{bert}_j}^{L}}^{(1)}(v_{\text{bert}_{1:c}}^{L},j)=&[\text{LSTM}_{v_{\text{bert}_j}^{L}}^{f_{(1)}}(v_{\text{bert}_{1:j}}^{L})\oplus\\&\text{LSTM}_{v_{\text{bert}_j}^{L}}^{b_{(1)}}(v_{\text{bert}_{c:j}}^{L})]\boldsymbol{W}_{\text{bert}}^{(1)}\\=&\,y_{v_{\text{bert}_j}^{L}}^{(1)},\quad j=1\cdots c\end{aligned} \tag{9}$$

$\text{BiLSTM}_{v_{\text{bert}_j}^{L}}^{(1)}(*)$ 表示右侧第一层 BiLSTM 网络，$\text{LSTM}_{v_{\text{bert}_j}^{L}}^{f_{(1)}}(*)$、$\text{LSTM}_{v_{\text{bert}_j}^{L}}^{b_{(1)}}(*)$ 为其对应的前向函数和后向函数，$\boldsymbol{W}_{\text{bert}}^{(1)}$ 为权重矩阵。

(4) 以上一层输出为输入，右侧第二层 BiLSTM 的第 j 个位置的输出如式(10)、式(11)所示。

$$\begin{aligned}\text{BiLSTM}_{v_{\text{bert}_j}^{L}}^{(2)}(y_{v_{\text{bert}_{1:c}}^{L}}^{(1)},j)=&[\text{LSTM}_{v_{\text{bert}_j}^{L}}^{f_{(2)}}(y_{v_{\text{bert}_{1:j}}^{L}}^{(1)})\oplus\\&\text{LSTM}_{v_{\text{bert}_j}^{L}}^{b_{(2)}}(y_{v_{\text{bert}_{1:j}}^{L}}^{(1)})]\boldsymbol{W}_{\text{bert}}^{(2)}\\=&\,y_{v_{\text{bert}_j}^{L}}^{(2)},\quad j=1\cdots c\end{aligned} \tag{10}$$

$$y_{v_{\text{bert}}^{L}}^{*}=y_{v_{\text{bert}_c}^{L}}^{(2)} \tag{11}$$

其中，$\text{BiLSTM}_{v_{\text{bert}_j}^{L}}^{(2)}(*)$ 表示右侧第二层 BiLSTM 网络，$\text{LSTM}_{v_{\text{bert}_j}^{L}}^{f_{(2)}}(*)$、$\text{LSTM}_{v_{\text{bert}_j}^{L}}^{b_{(2)}}(*)$ 为对应的前向函数和后向函数，$\boldsymbol{W}_{\text{bert}}^{(2)}$ 为权重矩阵。

译文的整体质量向量和整体语境词向量分别由式(8)和式(11)计算获取，我们将其相连得到表征译文忠实度和流利度的整体特征向量 $\boldsymbol{y}^{*}$ 如式(12)所示。

$$\boldsymbol{y}^{*}=y_{q_y}^{*}\oplus y_{v_{\text{bert}}^{L}}^{*} \tag{12}$$

其中，符号⊕表示向量连接操作。

为了估计译文质量值，并将其缩放至 0～1 范围内，激活函数为 sigmoid 的单个节点的全连接层用于回归生成译文质量估计得分 QE_{score}，如式(13)所示。

$$\begin{aligned}\text{QE}_{\text{score}}=&\frac{1}{1+\exp(-\boldsymbol{W}_{\text{qe}}y^{*})}\\=&\,\text{sigmoid}(\boldsymbol{W}_{\text{qe}}y^{*})\end{aligned} \tag{13}$$

其中，$\boldsymbol{W}_{\text{qe}}$ 为全连接神经网络层的权重矩阵。

给定 QE 任务的训练集：$\{x^{(i)},y^{(i)},\text{HTER}^{(i)}\}_{i=1}^{M}$，其中 $x^{(i)}$ 为源语言句子，$y^{(i)}$ 为待评价机器译文，$\text{HTER}^{(i)}$ 为人工评价结果，M 为训练集包含的样本数量，模型的优化目标为在训练集上最小化质量估计得分 QE_{score} 与人工评价得分 HTER

的平方差，如式(14)所示。

$$J(\theta)=\frac{1}{M}\sum_{i=1}^{M}(\mathrm{QE}_{\mathrm{score}}(x^{(i)},y^{(i)},\theta)-\mathrm{HTER}^{(i)})^2 \tag{14}$$

其中，符号θ表示模型参数空间。

3 实验与分析

3.1 实验设置

为测试CUNQE方法的性能，我们在CWMT18句子级别译文质量估计任务上进行了实验。CWMT18句子级别译文质量估计任务评价中英和英中翻译方向的机器输出译文质量。为了提取质量向量，评测官方发布的用于新闻翻译任务的casia2015、casict2015、datum2017和neu2017双语平行语料被用作训练双向RNN编码器—解码器模型。为了减少平行语料中噪声对提取质量向量的影响，我们仅保留源语言句子长度和目标语言句子长度的比值在(1/3, 3)范围内的句对，并剔除任意一端长度超过70的句对，对英文端句子使用了BPE子词切分和大小写转换等预处理，对中文端句子进行了中文分词。表1给出了预处理后使用的双语平行语料规模和译文质量估计语料的训练集、开发集和测试集语料数量。实验语料规模统计如表1所示。

表1 实验语料规模统计

	名称	句对数量
中英平行语料	训练集	6M
QE中英语料	训练集	8 785
	验证集	1 064
	测试集	1 544
QE英中语料	训练集	12 865
	验证集	1 040
	测试集	1 634

提取的质量向量维数为700，BERT语境词向量维数为768，使用谷歌发布的“BERT-Base, Uncased”语言模型提取英语句子的语境词向量，使用“BERT-Base, Chinese”语言模型提取中文句子的语境词向量[16]。CUNQE网络结构中堆叠BiLSTM的隐含层状态数量设置为100，训练使用“rmsprop”优化器，训练批尺寸为108，并使用dropout机制(rate=0.2)防止过拟合。

为了评价译文质量估计的性能，皮尔森相关系数(Pearson r)被用来测定译文质量估计与人工评价打分的相关性，斯皮尔曼相关系数(Spearman ρ)被用来测定机器译文质量排名与人工评价排名的相关性。皮尔森相关系数或斯皮尔曼相关系数的值越大，表示性能越好。

3.2 实验结果

表2给出了融合不同隐含层BERT语境词向量的CUNQE方法的性能，以及基线系统UNQE[10]、阿里QE Brain[13]和CWMT18参与该评测的最好系统(CWMT18 1st ranked)在CWMT18句子级别译文质量估计任务上的皮尔森相关系数值和斯皮尔曼相关系数值。其中CUNQE^{-12},…, CUNQE^{-2}, 和CUNQE^{-1}分别表示融合了BERT不同隐含层的CUNQE方法，而$\mathrm{CUNQE}^{\mathrm{AVG}}$是指融合平均池化BERT 12个隐含层后的CUNQE方法，$\mathrm{CUNQE}^{\mathrm{AVG_4}}$表示融合平均池化BERT最后4个隐含层(−1层，−2层，−3层，−4层)后的CUNQE方法。

首先，融合中上层BERT语境词向量(第−6层以上)的CUNQE模型与只使用质量向量的基线系统UNQE相比，皮尔森相关系数值和斯皮尔曼相关系数值均有显著提高。皮尔森相关系数值和斯皮尔曼相关系数提升的幅度在中英方向分别为7.3%、2.2%；在英中方向分别为7.2%、6.4%。其中融合最后4层(−1层，−2层，−3层，−4层)平均池化的语境词向量方法$\mathrm{CUNQE}^{\mathrm{AVG_4}}$性能最优，其次是引入−4层、−5层语境词向量的方法CUNQE^{-4}、CUNQE^{-5}，而融合−3层和融合所有层平均池化的语境词向量方法(CUNQE^{-3}、$\mathrm{CUNQE}^{\mathrm{AVG}}$)性能相当。融合12个不同隐含层语境词向量的方法，其性能随着隐含层的变化呈现出规律：以CUNQE^{-4}、CUNQE^{-5}为代表，融合中上层语境词向量的方法表现优异；融合上层语境词向量(−1层，−2层，−3层)的方法表现良好，而融合下层语境词向量(−8层，…，−12层)的方法劣于基线系统UNQE。这也与Liu等[17]的结论相吻合：BERT Transformer语言模型的中间或者靠近中间的隐含层提取的语境词向量蕴含了丰富的语义信息，可迁移性较强。CUNQE方法与UNQE方法对比表明了中上层BERT语境词向量显著提高了译文质量估计与人工评价的相关性。

其次，CUNQE方法和UNQE方法在中英与英中任务上皮尔森相关系数值均超过了参与该评测的

最优系统(CWMT18 1st ranked)，其中 CUNQE 方法的皮尔森相关系数值比参与该评测的最优系统在中英与英中任务上分别最高提升了 12.6% 和 19.1%。

表 2　在 CWMT18 译文质量估计评测任务上不同 QE 系统的性能

模型	平行语料规模	中英		英中	
		Pearson r	Spearman ρ	Pearson r	Spearman ρ
$CUNQE^{-12}$	WMT 6M	0.464	0.393	0.444	0.351
$CUNQE^{-11}$		0.473	0.413	0.463	0.340
$CUNQE^{-10}$		0.466	0.409	0.470	0.341
$CUNQE^{-9}$		0.489	0.403	0.485	0.367
$CUNQE^{-8}$		0.505	0.448	0.515	0.381
$CUNQE^{-7}$		0.516	0.488	0.574	0.434
$CUNQE^{-6}$		0.544	0.467	0.576	0.451
$CUNQE^{-5}$		0.574	0.482	**0.596**	0.433
$CUNQE^{-4}$		0.577	0.491	**0.596**	0.449
$CUNQE^{-3}$		0.554	**0.515**	0.585	**0.455**
$CUNQE^{-2}$		0.552	0.487	0.574	0.434
$CUNQE^{-1}$		0.551	0.485	0.543	0.413
$CUNQE^{AVG}$		0.540	0.461	0.580	0.419
$CUNQE^{AVG\_4}$		**0.591**	0.492	**0.596**	0.451
UNQE	WMT 6M	0.518	0.470	0.524	0.391
CWMT18 1st ranked	CWMT 8M+8M	0.465	\	0.405	\
QE Brain	CWMT 8M	0.564	\	0.588	\
QE Brain	WMT 25M+25M	**0.612**	\	**0.620**	\

最后，我们将 CUNQE 方法与阿里的 QE Brain 方法[13]进行比较，在使用同等规模平行语料的情况下，$CUNQE^{AVG\_4}$ 的性能超过了 QE Brain 方法，但是当 QE Brain 方法增加训练语料规模时，CUNQE 方法劣于 QE Brain 方法。

3.3　实验分析

BERT 语言模型的不同隐含层蕴含不同的语言信息。在 BERT 模型层次结构上，最下层(−12 层)隐含层和输入词词向量最接近，因而该层的语境词向量表示更多地保留了输入词的原始释义，较少编码从 QE 任务学习的上下文信息(词的使用方法等)，其邻近隐含层的语境词向量表示有相似的现象。因此，融合下层(−8 层，…，−12 层)语境词向量的 CUNQE 方法没有获得利于 QE 任务推导的语境信息，其性能较差；最上一层(−1 层)隐含层接近 QE 任务层，因而使用该层的语境词向量表示可能使实验产生过拟合现象，但是其语境词向量表示蕴含了较多从 QE 任务学习的上下文信息，利于任务的进一步推导。其邻近隐含层的语境词向量表示有相似的性质。因此融合上层(−1 层，−2 层，−3 层)语境词向量的 CUNQE 方法优于 UNQE 方法；−4 层、−5 层处于 BERT 语言模型结构的中上层，对应的语境词向量表示既蕴含了大量从 QE 任务学习的上下文信息，又适当保留了输入词的原始释义，可迁移性较强，因而对应的 CUNQE 方法性能提升幅度最大。

融合所有层语境词向量表示的 $CUNQE^{AVG}$ 方法，通过平均池化运算有效地利用了每个隐含层蕴含的语言信息，因而优于 UNQE 方法；融合最上 4 层平均池化语境词向量表示的 $CUNQE^{AVG\_4}$ 方法，其语境词向量蕴含的上下文信息最丰富，因此性能最佳。

为了进一步分析融合 BERT 语境词向量的

CUNQE方法的特点，我们在实验开发集上分别抽取了中英和英中翻译质量估计的实例进行分析。表3给出了抽取的两组质量估计实例，其中HTER译文指人工对机器译文进行后编辑的结果，它可以看作是参考译文，HTER得分是指将机器译文转换成HTER译文需要的最少编辑次数与译文长度的比值，它可以看作是描述译文质量的定量标准。$CUNQE^{AVG_4}$得分和UNQE得分指$CUNQE^{AVG_4}$方法和UNQE方法估计该机器译文质量的分值，分值越接近HTER得分，表明其估计的越准确。

表3 不同质量估计方法对机器译文打分实例

源语言句子：一是**树立**健康的消费理念。 机器译文：The first thing we should do is **form** a healthy consumption concept. HTER译文：The first thing we should do is **forming** a healthy consumption concept. HTER得分：0.077 $CUNQE^{AVG_4}$得分：0.072 UNQE得分：0.109
源语言句子：They are trying to adapt, but it can be **confounding**. 机器译文：他们正在尝试适应，但可能会**混淆**。 HTER译文：他们正在尝试适应，但这又**谈何容易**。 HTER得分：0.294 $CUNQE^{AVG_4}$得分：0.339 UNQE得分：0.135

尽管机器译文中的词语与源语言句子中的词语语义基本对应，其中“*树立*”对应“*form*”或“*forming*”、“*confounding*”对应“*混淆*”或“*谈何容易*”，但是译文没有根据词语所在上下文语境翻译生成，因而译文流利度较低。查询词汇化翻译概率表容易得到$p(form|树立)>p(forming|树立)$、$p(混淆|confounding)>p(谈何容易|confounding)$，这表明机器译文都是通过选择最大概率的词汇进行翻译，因此译文的忠实度较好。但是根据其上下文语境信息，以上实例都应该选择较小条件概率的词汇进行翻译。上下文语境信息是刻画译文流利度的一个重要因素，例如，第一个实例中，“*树立*”根据其上下文信息翻译为“*forming*”，第二个实例中，“*confounding*”根据其上下文信息翻译为“*谈何容易*”，这都提高了机器译文的流利度。$CUNQE^{AVG_4}$方法结合语境词向量进行译文质量估计，能更准确地描述译文的流利度特征，因此，相比仅使用忠实度特征的UNQE方法，它的打分更接近于HTER得分，即其打分更准确。这说明结合语境词向量的译文质量估计方法能充分利用上下文信息对译文质量进行估计。

4 相关工作

在神经译文质量估计研究中，Kim等提出了POSTECH质量估计框架[9]，该框架利用双向编码器—解码器神经翻译系统提取质量向量（预测器），利用循环神经网络进行质量评估（估计器），有效提高了机器译文自动估计和人工评价的相关性；Li等人在其基础上将预测器和估计器重构为联合神经网络框架，并提出端到端的译文质量估计方法（UNQE）[10]，该方法在WMT18句子级别机器译文质量估计任务中取得了多个方向上的第一名或并列第一名；Fan等结合基于自注意力机制的Transformer网络提出了一种条件语言模型[12-13]，利用该模型抽取原文和译文的语言特征，然后使用BiLSTM网络进行质量估计。该方法在WMT18句子级别和词语级别译文质量估计任务中取得了优异的成绩。

在语境词向量及其应用上，Peters等最先提出基于BiLSTM网络的语言模型训练上下文相关的词表示，简称ELMo，ELMo可以作为特征加入到自然语言处理任务的有监督模型中[18]；在其基础上，OpenAI团队提出了一种新的模型，该模型使用Transformer网络代替BiLSTM结构训练语言模型来捕获长距离语言依存关系，简称GPT[19]。在进行具体有监督学习任务微调时，GPT可以作为附加训练目标。与ELMo不同，GPT无须根据任务构建新的模型结构。但是ELMo和GPT在预训练时均使用单向的网络模型来学习语言表示，针对这个问题，Devlin等提出了基于Transformer编码器网络结构的词向量训练方法BERT[16]，在推理、问答和命名实体识别等不同的自然语言处理任务上，使用BERT语境词向量比使用ELMo和GPT语境词向量的方法效果有较大幅度的提升。

5 结束语

为了充分利用译文的上下文信息，我们构建并联神经网络，在当前神经译文质量估计方法中引入了BERT语境词向量进行质量估计，实验结果表明

该方法能显著提高译文质量估计的效果，实验分析揭示该方法能有效利用流利度信息进行质量估计。

尽管 BERT 语境词向量在多个自然语言处理任务上优于 ELMo 和 GPT 语境词向量，但是近期学者们提出了一些更先进的语境词向量方法，在未来的研究中，我们将进一步对比不同语境词向量方法对译文质量估计的影响。

参考文献

[1] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2013.

[2] 刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展，2017，54 (06)：1144-1149.

[3] 李亚超，熊德意，张民. 神经机器翻译综述[J]. 计算机学报，2018，41 (12)：2734-2755.

[4] Specia L，Shah K，De Souza J G C，et al. QuEst-A translation quality estimation framework [C]//Proceedings of the ACL，2013：79-84.

[5] Shah K，Logacheva V，Paetzold G，et al. Shef-nn：Translation quality estimation with neural networks [C]//Proceedings of the WMT，2015：342-347.

[6] Chen Z，Tan Y，Zhang C，et al. Improving machine translation quality estimation with neural network features[C]//Proceedings of the WMT，2017：551-555.

[7] 陈志明，李茂西，王明文. 基于神经网络特征的句子级别译文质量估计[J]. 计算机研究与发展，2017，54 (8)：1804-1812.

[8] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the NIPS，2013：3111-3119.

[9] Kim H，Jung H Y，Kwon H，et al. Predictor-estimator：Neural quality estimation based on target word prediction for machine translation [J]. ACM Transactions on Asian and Low-Resource Language Information Processing，2017，17(1)：1-22.

[10] Li M，Xiang Q，Chen Z，et al. A unified neural network for quality estimation of machine translation [J]. IEICE Transactions on Information and Systems，2018，E101-D (9)：2417-2421.

[11] Ive J，Blain F，Specia L. deepQuest：A framework for neural-based quality estimation[C]//Proceedings of COLING，2018：3146-3157.

[12] Wang J，Fan K，Li B，et al. Alibaba submission for WMT18 quality estimation task[C]//Proceedings of the WMT，2018：809-815.

[13] Fan K，Wang J，Li B，et al. "Bilingual Expert" can find translation errors[C]//Proceedings of the AAAI，2019：1-8.

[14] 孙潇，朱聪慧，赵铁军. 融合翻译知识的机器翻译质量估计算法[J]. 智能计算机与应用，2019，9(2)：271-275.

[15] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need [J]. arXiv preprint arXiv：1706.03762，2017.

[16] 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.

[17] Liu N F，Gardner M，Belinkov Y，et al. Linguistic knowledge and transferability of contextual representations[J]. arXiv preprint arXiv：1903.08855，2019.

[18] Peters M，Neumann M，Iyyer M，et al. Deep contextualized word representations[C]//Proceedings of the NAACL，2018：2227-2237.

[19] Radford A，Narasimhan K，Salimans T，et al. Improving language understanding with unsupervised learning[R]. Technical Report，OpenAI，2018.

李培芸(1994—)，硕士研究生，主要研究领域为自然语言处理和机器翻译。

E-mail：lpyjxnu@gmail.com

裘白莲(1981—)，博士研究生，讲师，主要研究领域为计算语言学和机器翻译。

E-mail：57696163@qq.com

李茂西(1977—)，通信作者，博士，副教授，主要研究领域为自然语言处理和机器翻译。

E-mail：mosesli@jxnu.edu.cn