句子级别机器译文质量估计研究综述

罗兰， 何贤敏， 李茂西*

（江西师范大学计算机信息工程学院，江西省 南昌市 330022）

摘要：近年来，随着估计结果与人工评价的相关性逐步增强，译文质量估计引起了机器翻译研究者们的广泛关注和高度重视。该文对句子级别译文质量估计方法进行了综述，将它们分为基于传统机器学习的方法、基于神经翻译模型的方法和基于预训练语言模型的方法，梳理和对比了这三种译文质量估计方法的代表性工作、交叉工作以及不同方法的发展路线，并介绍了推动译文质量估计研究的相关评测活动和性能评价指标，最后展望句子级别译文质量估计今后的研究方向和发展趋势，并对全文进行总结。

关键词：机器翻译，译文质量估计，深度神经网络，预训练语言模型，评价指标

中图分类号：TP391 **文献标识码：**A

0 引言

机器译文质量估计(Quality Estimation, QE)是指不依赖人工参考译文的情况下实时地估计翻译的质量，它是机器翻译的最新研究方向之一[1]。它的主要作用包括[2-3]：（1）以机器译文的后编辑工作量或人工评价结果为翻译质量估计基准，提供一个有实际含义的译文质量指标，使机器翻译的普通用户（主要是仅懂目标语言的用户）了解机器译文的可靠程度；告诉在机器译文上进行后编辑的专业译员后编辑需要的工作量大小并摒弃质量低劣的机器译文。（2）译文质量估计方法针对句子级别相关性进行模型优化，因此可以克服利用人工参考译文的机器译文自动评价方法(Automatic Evaluation of Machine Translation, MTE)[4]在句子级别与人工评价相关性低的不足。（3）由于机器译文质量估计不需要人工参考译文，因此它能辅助神经翻译模型进行网络权重的自训练，以代替传统的通过开发集（开发集中每个待翻译的句子都有人工参考译文）优化翻译系统网络权重的方法。

机器译文质量估计根据对翻译结果的评价粒度不同而分为单词级别、短语级别、句子级别和文档级别。由于围绕句子级别译文质量估计的研究工作较多且其应用广泛，本文对其进行综述。在缺乏人工参考译文，仅给定源语言句子和它的机器译文的情况下，译文质量估计通常被看作一个有监督的回归/分类任务[2]。形式化描述为给定训练集$\{(s_i,z_i),y_i\}_{i=1}^{N}$，其中符号 s_i，z_i 分别表示训练集中第 i 个源语言句子 s_i 和其机器翻译输出译文 z_i；y_i 表示对应的译文质量人工评价值，一般为人工标注的译文质量等级(介入 1-5 的整数值)[2]、或度量后编辑工作量大小的 HTER 值(0-1 之间的实数值)[3,5]、或直接评价值(Direct Assessment, DA)[6]，N 表示训练集中样本数量，由于人工标注译文质量比较耗时费力，训练集的规模一般比较小，仅包含几万个样本；然后利用该训练集建立模型预测（估计）未知的源语言句子 s_j 的译文 z_j 的翻译质量。因此如何从源语言句子和其机器译文中自动提取表征译文质量的特征，并利用这些特征构建有效的分类/回归模型是译文质量估计的两个重要问题。我们根据特征提取和模型构建方法的不同将句子级别译文质量估计分为基于传统机器学习的方法、基于神经翻译模型的方法和基于预训练语言模型的方法。

基于传统机器学习的句子级别译文质量估计方法是早期的方法，它通过启发式规则人工设计影响译文质量的特征，使用传统机器学习算法预测译文的质量。在特征提取时通常结合外部语言资源（大规模语言模型和双语平行语料等）和词法、句法以及语义分析工具（词形分析、句子依存结构分析和语义角色标注等等）从源语言句子和机器译文二元组中提取表征译文质量的流利度和忠实度等统计信息；在模型构建时采用支持向量机、贝叶斯分类器或随机森林等等算法预测译文质量。该类方法典型的代表是 QuEst 框架[1]。

基于神经翻译模型的句子级别译文质量估计方法假设待估计的译文由神经翻译系统生成，通

基金项目：国家自然科学基金(61662031)

***通信作者**：mosesli@jxnu.edu.cn

过迁移学习利用神经翻译模型部分网络层提取描述译文质量的词级序列特征，使用循环神经网络(Recurrent Neural Network, RNN)将词级特征抽象表示为句子级别特征并预测译文质量。其中神经翻译模型由于参数量大，通常使用双语平行语料进行预训练；而 RNN 预测网络则在译文质量估计训练集上进行训练。由于神经机器翻译分为基于 RNN 带注意力机制的编码器-解码器模型[7]和 Transformer 模型等等[8]，因此，其特征提取方法也分为基于 RNN 编码器-解码器的方法和基于 Transformer 的方法，前者代表性的工作是预测器-估计器模型(Predictor-Estimator)[9]，而后者代表性的工作是双语专家(Bilingual Expert)[10]。

基于预训练语言模型的句子级别译文质量估计方法利用在大规模语料上训练获取的语言模型[11]提取表征译文质量的特征，使用神经网络或支持向量机预测机器译文质量。由于预训练语言模型的种类很多，包括静态预训练语言模型[12]、动态上下文预训练语言模型[13]和跨语种预训练语言模型[14, 15]等等，该类译文质量估计方法也可以据此进行细分，其中基于静态预训练语言模型的译文质量估计方法代表性工作为 SHEF-NN[16]，基于动态预训练语言模型的译文质量估计方法代表性工作为 Multi-BERT QE[11]，基于跨语种预训练语言模型的译文质量估计方法代表性工作为 TransQuest[17]。

不同种类译文质量估计方法提取的特征从不同角度描述了译文的质量。为了提高译文质量估计效果，许多工作[11, 18]将不同种类特征拼接融合，这些方法不能简单的归为某一类，下文将根据其主要使用的特征对其进行归类介绍。

本文 1， 2， 3 小节分别详细介绍这三种方法，第 4 节介绍相关的评测活动 WMT QE 任务、CCMT QE 任务和其评价指标，最后对未来的研究方向和发展趋势进行展望。

1 基于传统机器学习的句子级别机器译文质量估计

基于传统机器学习的句子级别译文质量估计方法采用机器学习中“特征工程+任务建模”的范式进行译文质量估计，由人工指定与译文质量相关的词法、句法和语义统计特征，利用计算机自动从源语言句子和机器译文中通过语言学分析提取这些特征[19]，根据回归模型建立特征与译文质量之间的映射函数。

QuEst 框架把机器译文质量估计问题看作是一个回归问题，它使用基于径向基函数核的支持向量机回归算法估计机器译文的质量，利用网格搜索进行特征权重学习。同时，该框架提供随机拉索(randomized lasso)和高斯过程(Gaussian Process)算法进行特征的选择。

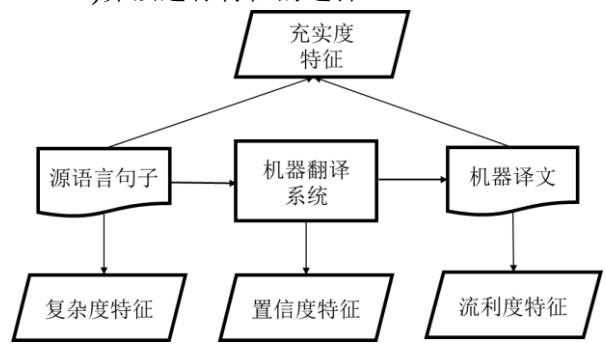

图 1 QuEst 特征提取模块

Figure 1 QuEst feature extraction module

其特征提取模块从源语言句子、机器译文和外部语言资源中提取 3 类总共 17 个与译文质量相关的基本特征。这 3 类特征包括：（1）从待翻译的源语言句子中提取定量反映翻译句子复杂度的 4 个特征；（2）从机器译文中提取描述译文流利程度的 3 个特征；（3）从源语言句子与机器译文的对应关系中提取描述译文充实度的 10 个特征，提取这 10 个特征需要使用[20]训练得到的词对齐关系以及源语言句子和机器译文的词性和句法分析结果。除了能提取这 3 类与翻译系统无关的基本特征，如果能获取机器翻译系统解码的细节，比如翻译系统对机器译文的全局打分、*n*-best 列表等，那么，该平台还能提取描述翻译系统置信度[21]的“黑盒子”特征。提取的这些特征都完全不需要人工参考译文。

QuEst 框架作为早期应用广泛的译文质量估计系统，在 WMT12-18 评测的句子级别译文质量估计子任务中被作为基线系统提供给参加评测的单位使用，其在不同年份的评测中性能如下图所示。

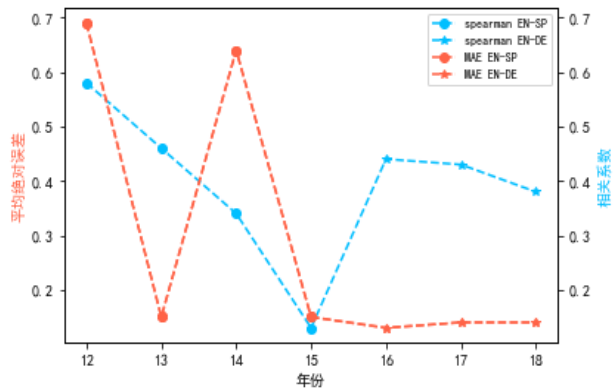

图 2 QuEst 在 WMT12-18 评测中与人工评价的相关性

Figure 2 Correlation of QuEst with manual evaluation in the WMT12-18 QE task

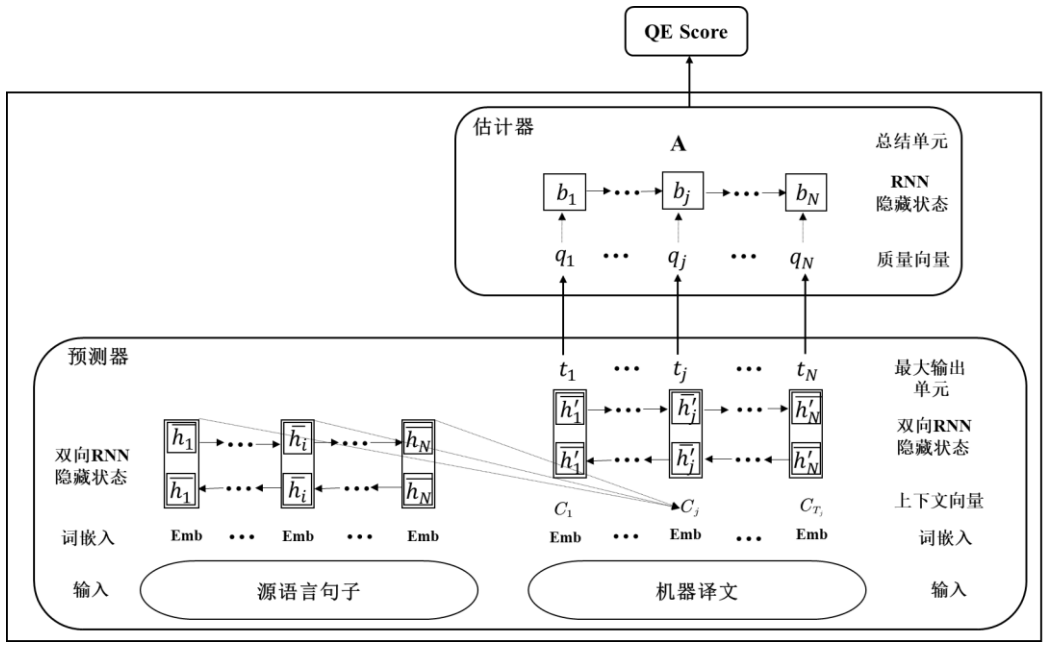

图 3 预测器-估计器总体网络框架

Figure 3 Overall network framework of predictor-estimator

QuEst 框架为机器译文质量估计搭建了一个基准平台，在其基础上许多工作对其进行了扩展，包括：（1）针对特征提取的研究，Hokamp 等人结合停止词（stop words）和词性标注 POS 构建语言模型，引入反向“白盒子”特征来作为目标端的质量特征[22]；Scarton 等人采用随机森林算法先对特征进行排序，然后采用逆向特征选择方法，以获得更优的特征集[23]。（2）针对机器学习算法的研究，Bicici 等人基于参考翻译机器(RTM)和并行特征衰减算法(ParFDA5)提取特征，在提取特征后采用岭回归算法进行机器译文质量估计[24]；Beck 等人提出稀疏高斯过程对机器译文质量进行估计[25]；Esplà-Gomis 等人采用多层感知器算法对机器译文质量进行估计[26]。

由于基于传统机器学习的方法在特征提取时需要对译文进行复杂的语言学分析，这些语言学分析不仅需要额外的资源，且与待估计的机器译文语言种类相关，这导致该类方法不易扩展，且泛化性差[27]，另外人工提取的特征大多数是一些语法和浅层语义特征，很少涉及译文的深层次语义信息[28]。随着深度学习的引入，深度神经网络强大的特征学习和表征能力，为译文质量估计提供了一个更好的选择。

2 基于神经翻译模型的句子级别机器译文质量估计

该类方法通过强制学习（Teacher Forcing）将源语言句子和其机器译文输入已在双语平行语料上训练好的神经翻译模型上提取表征翻译质量的词语级别质量向量，利用 RNN 网络汇总该质量向量获取句子级别质量向量，通过前馈神经网络估计译文质量值。根据提取质量向量使用的神经翻译模型的不同，该类方法可以分为基于 RNN 编码器-解码器的句子级别质量估计模型和基于 Transformer 的句子级别质量估计模型。

2.1 基于 RNN 编码器-解码器的句子级别 QE 模型

预测器-估计器[9]是该类方法的一个典型代表。其模型整体结构如图 3 所示，预测器通过基于 RNN 的编码器-解码器模型依次提取译文中每个词语的质量向量 q_i；估计器将具有时序的质量向量 q_i 通过 RNN 预测译文质量。

形式化描述如下，给定源语言句子 s 和其待估计机器译文 z，将 s 和 z 中每一个词 s_i 和 z_j 表征为词向量，作为编码器和解码器的输入。编码器利用上一时间步的输出 h_{i-1} 和当前时间步的输入 s_i 得到当前时间步的隐藏状态 h_i，解码器将编码器中各个时间步的隐藏状态做加权平均来获得上下文向量 c_j。

$$h_i = f(s_i, h_{i-1}) \tag{1}$$

$$c_j = p(h_1, ..., h_N) \tag{2}$$

函数 $f(\cdot)$表示循环神经网络隐藏层的变换，函数 $p(\cdot)$表示将编码器各个时间步的隐藏状态做加权平均。

解码器利用上一时间步的输入和隐藏状态，结合上下文向量 c_j 对当前词 z_j 进行预测。

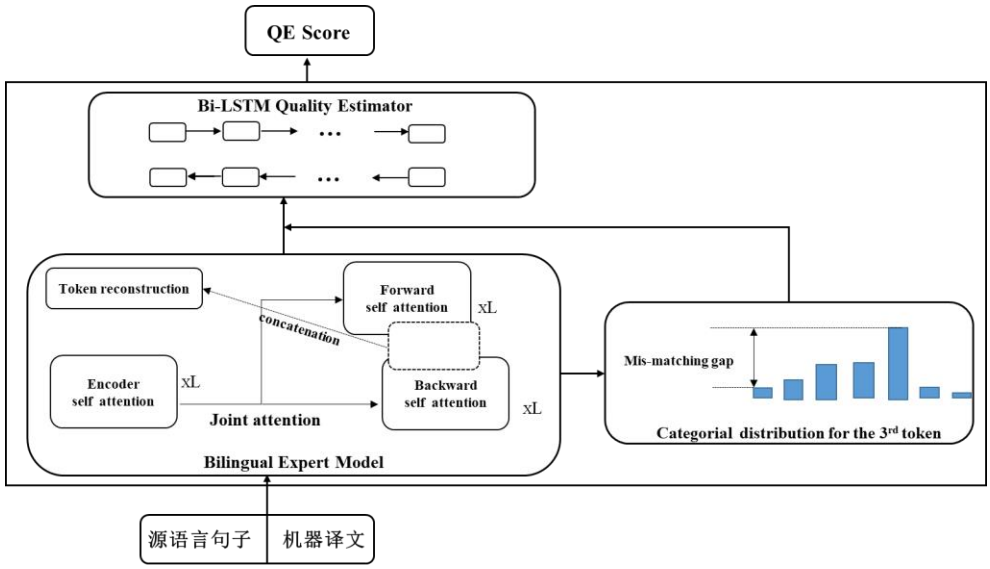

图 4 双语专家模型整体框架

Figure 4 Overall framework of the bilingual expert model

$$\mathrm{p}(z_j \mid z_j \notin Z) = g([z_{j-1}; z_{j+1}], [\overrightarrow{h'_{j-1}}; \overleftarrow{h'_{j+1}}], c_j) = \frac{\exp(z_j^T W_{O_1} W_{O_2} t_j)}{\sum_{K=1}^{K_z} \exp(z_K^T W_{O_1} W_{O_2} t_j)} \quad (3)$$

函数 $g(\cdot)$为预测目标词 z_j 概率的非线性函数，$W_{O_1} \in \mathbb{R}^{Kz \times d}$ 和 $W_{O_2} \in \mathbb{R}^{d \times l}$ 为权重矩阵，Z 表示机器译文当前时间步的所有输出，K_z 是机器译文的词汇表大小，d 是质量向量的维度，l 是最大单元输出的维度，t_j 是最大单元的输出，它包含了目标词 z_j 的质量信息。

$$t_j = [\max\{\tilde{t}_{j,2k-1}, \tilde{t}_{j,2k}\}]_{k=1,\ldots,l}^{T} \quad (4)$$

其中，$\tilde{t}_{j,k}$ 为向量 $\tilde{t}_j$ 的第 k 个元素。

$$\tilde{t}_j = H'_o[\overrightarrow{h'}_{j-1}; \overleftarrow{h'}_{j+1}] + Z'_o[z_{j-1}; z_{j+1}] + C_o c_j \quad (5)$$

其中，$H'_o \in \mathbb{R}^{2l \times 2n}$，$Z'_o \in \mathbb{R}^{2l \times 2m}$，$C_o \in \mathbb{R}^{2l \times 2n}$ 分别是机器译文输出的词嵌入矩阵，m 是词嵌入的维度，n 是双向 RNN 隐藏状态的维度。

通过 t_j 可以计算出各个词的质量向量 q_j:

$$q_j = [row_j(W_{O_1}) \circ [W_{O_2} t_j]^T]^T \quad (6)$$

在估计器中，RNN 最后一个隐藏状态 b_N 包含所有质量特征的信息，将其作为汇总特征 A，使用前馈神经网络得到机器译文质量的预测值。

$$b_j = U(q_j, b_{j-1}) \quad (7)$$

$$QE_{score} = QE_{score}(q_1, \ldots, q_N) = \sigma(W_{QE}^T A) \quad (8)$$

函数 $U(\cdot)$表示 RNN 通过当前时间步的质量向量和上一时间步的隐藏状态得到当前时间步的隐藏状态。$\sigma(\cdot)$表示 sigmoid 函数，$W_{QE} \in \mathbb{R}^r$，r 为汇总特征 A 的维度。

预测器-估计器模型的提出，启发了神经机器译文质量估计的研究，后续很多相关方法以此模型作为基本框架进行扩展，包括 Kim 等人在 WMT17 机器译文质量估计评测任务[30]中对预测器-估计器模型进行改进，提出利用堆栈传播联合学习预测器-估计器两阶段模型，实现从估计器到预测器的反向传播，并且部署了带有堆栈传播的多级任务学习[31]；Patel 等人提出在 RNN 模型中基于双语设置上下文窗口来提取质量特征[32]；Kashif 等人提出的 SHEF-LIUM 模型联合连续空间语言模型(CSLM)和神经网络模型来提取质量特征[33]；André F 等人结合三个神经系统：一个前馈系统,一个卷积神经网络和一个循环系统来进行机器译文的质量估计[34]；而 Li 等人认为将预测器和估计器分别在双语平行语料和机器译文质量估计语料上单独训练，训练后的模型不一定是最优的译文质量估计模型，提出联合神经网络模型进行译文质量估计[35]。

2.2 基于 Transformer 结构的句子级别 QE 模型

双语专家模型[10]是该类方法的一个典型代表。模型整体结构如图 4 所示，该模型基于 Transformer 构造预测器，基于双向长短期记忆网络(Bi-LSTM)[36]构造估计器。预测器模块主要包括三个部分：基于自注意力机制的编码器，输入机器译文的前向/后向编码器和机器译文的重构器。其中，两个编码器存在一定的差异，第一个编码器包括自注意力层和全连接网络层；第二个编码器包括遮蔽自注意力层，编码器-解码器自注意力层和全连接网络层，因为第二个编码器的作

用不同于翻译模型中的解码器，其作用更像编码器，所以将其称为前向/后向编码器。给定平行语料 (s,z)，输入预测器对模型进行预训练时，条件概率 $p(z|y)p(y|s)$ 是未知的，但是潜在变量 y 包含源语言句子和机器翻译输出之间丰富的深层语义信息，根据贝叶斯公式，可以得出潜在变量 y 的后验分布：

$$p(y|z,s)=\frac{p(z|y)p(y|s)}{p(z|s)} \tag{9}$$

由于 $p(z|s)=\int p(z|y)p(y|s)dz$ 难以直接计算，可使用变分分布 $q(y|z,s)$ 最小化 KL 散度来逼近真实的后验值：

$$\min D_{KL}(q(y|z,s)\|p(y|z,s) \tag{10}$$

优化上面的目标函数，相当于最大化：

$$\max E_{q(y|z,s)}[p(z|y)]-D_{KL}(q(y|z,s)\|\mathrm{p}(y|s)) \tag{11}$$

新的目标函数可以直接计算条件概率 $p(y|s)$；公式中最大似然期望是一个变分自动编码器，近似表示为：

$$E_{q(y|z,s)}[p(z|y)]\approx \mathrm{p}(z|\tilde{y}),\tilde{y}\sim q(y|z,s) \tag{12}$$

基于单向 Transformer 构造的编码器和基于双向 Transformer 构造的前向/后向编码器表示 $q(z|t,s)$，重构器则对应 $p(t|z)$。为了使预测更加高效，将 $p(t|z)$和 $q(z|t,s)$进行因式分解，明确假设条件独立性：

$$q(y|z,s)=\prod_j q(\vec{y}_j|z,s_{<j})q(\overleftarrow{y}_j|z,s_{>j}) \tag{13}$$

基于双向 Transformer 构造的前向/后向编码器，每一次在预测机器译文的当前词时，Transformer 需要使用前向与后向两部分信息。例如，当前要预测机器译文的第 j 个词，对于正向序列而言，模型需要使用目标端第 j-1 个词的前向深层语义特征向量和第 j-1 个词的词向量。而对于后向序列而言，模型需要使用目标端第 j+1 个词的后向深层语义特征向量与第 j+1 个词的词向量。提取的特征有：正向深层语义特征向量 $\vec{y}_j$；反向深层语义特征向量 $\overleftarrow{y}_j$；前一个词的词向量 $e_{z_{j-1}}$；后一个词的词向量 $e_{z_{j+1}}$。

Fan 等人认为模型的翻译结果与预训练模型给出的正确翻译结果会存在一个差值，这个差值在提取的特征中起到关键作用，他们通过实验发现只利用这一特征（即 4 维“不匹配”特征）做下一步预测，也会得到较好的结果。这一部分提取的特征有：目标端强制解码为当前词的概率信息 l_{j,k_j}，概率最高词语的概率信息 $l_{j,i^j_{\max}}$，强制解码为当前词与解码为概率最高词的概率信息差异 $l_{j,k_j}-l_{j,i^j_{\max}}$，当前词与预测词是否一致 $\amalg_{k_j\neq i^j_{\max}}$，即得到 4 维“不匹配”特征：

$$f_j^{kk}=(l_{j,k_j},l_{j,i^j_{\max}},l_{j,k_j}-l_{j,i^j_{\max}},\amalg_{\mathrm{k}_j\neq i^j_{\max}}) \tag{14}$$

l_j 表示使用 softmax 函数之前的对数向量，k_j 表示翻译输出中第 j 个词在字典中的编号，$i^j_{\max}=\arg\max_i l_j$ 表示双语专家预测的编号，$\amalg$ 表示指示函数。

Fan 等人将从预测器中提取的深层语义特征和 4 维“不匹配”特征一起输入到基于 Bi-LSTM 的估计器中，将 Bi-LSTM 最后一个时间步的前向和后向隐藏状态作为机器译文质量估计值[37]。

$$\overrightarrow{h_{1:K}},\overleftarrow{h_{1:K}}=Bi-LSTM(\{f_j\}_{j=1}^K) \tag{15}$$

K 表示翻译输出的总词数，$\{f_j\}_{j=1}^K$ 表示所有序列特征沿深度方向连接得到的单个向量。

最后最小化真实的 HTER 值和预测的句子级质量估计分数之间的差值：

$$\arg\min\|h-sigmoid(w^T[\overrightarrow{h_K},\overleftarrow{h_K}])\|_2^2 \tag{16}$$

其中 w 为线性层的权值向量。

Fan 等人在训练预测器时使用的是大规模双语平行语料，训练估计器时使用的是译文质量估计训练数据，由于 WMT 评测提供的译文质量估计训练数据规模较小，因此他们在英德和德英两个语言方向上分别构造了 30 万左右的译文质量估计伪训练数据。先用构造的伪数据和真实数据来训练估计器，然后再次使用真实数据对估计器进行微调。该模型在 WMT18 译文质量估计评测任务中获得 6 项任务的第一名[38]。

双语专家模型的提出进一步推动了神经译文质量估计的研究，目前大多数译文质量估计模型都是基于 Transformer 结构，包括 Hou 等人提出的 BiQE 模型，其从两个不同的翻译方向运用两种语言之间的翻译知识来提取特征[39]；Wang 等人提出利用层融合机制的 Transformer-DLCL 模型进行译文质量估计特征提取[40]；Chen 等人提出仅使用 Transformer 瓶颈层提取词语特征，结合 Bi-LSTM 网络进行机器译文质量估计[41]。

虽然以预测器-估计器作为基本框架的译文质量估计模型得到了广泛地应用，但是 Cui 等人认为该框架存在两个问题：1）预测器是在大量平行语料上训练，而估计器是在译文质量估计数据上进行训练，会造成数据不一致的问题；2）预测器的目的是进行词预测，而估计器的目的是进行译文质量估计任务，这也造成了任务不一致的问题。这两个问题会对质量估计的结果造成负面影响，因此他们提出 DirectQE 模型，直接训练机器译文质量估计模型[42]。该模型包括两个阶段：第

一个阶段是生成器，在大量平行语料上进行训练，这一阶段关键是生成大量的译文质量估计伪数据；第二阶段是检测器，它在生成器生成的译文质量估计伪数据上进行预训练，在真实译文质量估计数据上进行微调。该模型的生成器和检测器都是基于 Transformer 结构。

从 2019 年开始 WMT 评测将 Unbabel 团队提出的开源框架 OpenKiwi[43]作为译文质量估计评测任务的基线系统，OpenKiwi 基于 Pytorch 深度学习框架实现了四个译文质量估计系统：（1）基于连续的空间深度神经网络的 QUETCH 模型[44]；（2）基于两个连续的前馈层和一个门控循环单元(GRU)层[45]的 NUQE 模型[34]；（3）基于 RNN 的预测器-估计器模型[31]；（4）以及将神经网络模型堆叠成一个含丰富特征的机器译文质量估计系统[46]。

3 基于预训练语言模型的句子级别机器译文质量估计

该类方法在进行机器译文质量估计之前，先训练神经网络语言模型，或者直接采用已经训练好的神经网络语言模型提取句子特征[47]。根据预训练语言模型的不同可以分为基于静态预训练词向量的方法，基于动态预训练词向量的方法，和基于跨语种预训练词向量的方法。

3.1 基于静态预训练词向量的句子级别 QE 方法

SHEF-NN 框架[16]是该类方法的一个典型代表，该框架选用的是连续语言空间模型（CSLM）。模型输入的是预测词的上下文单词，即 $z_j=w_{j-n+1},\cdots,w_{j-2},w_{j-1}$，($z_j$ 为第 j 个待预测词，w 为单词的表示)，采用 one-hot 编码，输出是词汇表中所有单词的后验概率。由于输出词汇表太大，Shah 等采用短列表方法[48]，即在该词汇表中仅选择 32K 最频繁的单词，在标准回退 *n*-gram 语言模型上得到这些单词的后验概率作为输入特征，加上 QuEst 框架提取的 17 个基线特征一起输入到有监督学习算法中，从而获得译文质量估计值。值得注意的是， Shah 等在 CSLM 模型中使用具有四个隐藏层的深度神经网络：第一层用于单词投影（每个上下文单词有 320 个单元），另外三个隐藏层有 1024 个单元用于概率估计。此外，该框架也适用于单词级别的机器译文质量估计，采用连续词袋模型（CBOW），也取到了不错的结果。

基于静态预训练词向量的方法提取特征是对仅使用基线特征作为译文质量估计特征的一大改进，促进了后续的研究。陈等人提出利用上下文单词预测模型和矩阵分解模型训练词向量提取特征，然后采用算术平均方法将词向量转化为句子向量以预测译文质量估计值[18]。

3.2 基于动态预训练词向量的句子级别 QE 方法

该类方法的典型代表有陆等人提出的基于 Multi-BERT[49]和联合编码的预训练语言模型[11]，为了使源语言句子和机器翻译输出能更好的进行语义间的交互，陆等人使用少量平行语料对 Multi-BERT 进行二次训练，并且强制要求遮挡词[MASK]只能出现在机器译文中，使该模型能够捕获所有的源语言句子信息，以更充分地预测译文中带[MASK]标记的单词。

在 Multi-BERT 的基础上，陆等人使用多种不同的网络结构对提取的特征进行了处理,如在 Multi-BERT 后面连接 Bi-GRU 网络，采用“Multi-BERT＋信息交互”的网络结构、使用 Multi-BERT＋Bi-GRU＋LASER[50]＋Baseline 的网络结构融入了多种特征以探究其对于 Multi-BERT 隐状态的应用效率，实验结果表明仅利用不同类型的神经网络进行微调，对模型性能的提升作用不大，而融合外部特征之后，模型性能提升明显，即使用 Multi-BERT＋Bi-GRU＋LASER＋Baseline 的网络结构得到的实验效果最好。

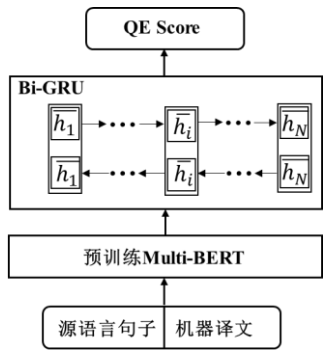

图 5 基于 Multi-BERT 和 Bi-GRU 的 QE 模型
Figure 5 QE model based on Multi-BERT and Bi-GRU

基于动态预训练词向量的方法能更充分地提取语义间潜在信息，基于此 Hou 等人在 WMT19 的机器译文质量估计评测任务中同样采用 Multi-BERT 预训练模型实现句子特征的提取[39]。为了更好的评估译文的质量，他们还使用基于 Transformer 的自注意力机制将源语言句子替换成伪参考译文（即二次翻译），将伪参考译文和其待估计机器译文作为句子对输入 Multi-BERT，提取特征向量，最终通过基于 Bi-LSTM 的评估器得到机器译文质量估计值；李等在联合神经网络模型的基础上增加 BERT 预训练语言模型特征以获得译文的流利度特征和忠实度特征以预测译文质量[51]。

3.3 基于跨语种预训练词向量的句子级别 QE 方法

随着跨语种预训练语言模型的出现，如 XLM[14]、XLM-R[15]等，使不同语言中句子间词汇的交互更加充分。它打破了单语种模型之间的壁垒，单个模型可以应用于多种语言的任务。但对该类模型进行微调以进行译文质量估计很困难，一是因为模型参数量过大，普通的 GPU 设备显存难以完整的加载模型参数；二是译文质量估计训练数据匮乏，在其上不能充分进行微调。所以目前是直接利用预训练的跨语种语言模型以提取表征译文质量的特征向量，该方法的典型代表有 T Ranasinghe 等人提出的 TransQuest 框架[17]。

T Ranasinghe 等人认为之前的译文质量估计模型都仅在单个语言对上起作用，而对于其他语言对，则需要重新训练模型，这一过程不仅泛化性差，而且需要耗费大量的计算资源，于是他们提出一种基于跨语言模型的译文质量估计模型，即基于 XLM-R 模型的 TransQuest 框架，包括 MonoTransQuest 模型和 SiameseTransQuest 模型。

如图 6 所示，MonoTransQuest 模型输入源语言句子和机器译文的拼接，它们间由[SEP]符号分隔。在 XLM-R 模型输出的词向量上使用三种池化策略提取描述译文质量向量的特征:CLS 向量表示（CLS 是序列的第一个词，包含了整个序列的特征）、平均池化和最大值池化。实验结果表明，采用 CLS 策略进行译文质量估计效果最好。

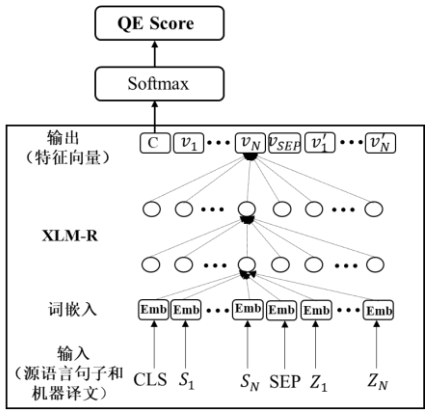

图 6 MonoTransQuest 模型
Figure 6 MonoTransQuest model

如图 7 所示，SiameseTransQuest 模型使用两个独立的 XLM-R 预训练模型，分别输入源语言句子和机器译文，与 MonoTransQuest 模型相同，也使用了三种池化策略，然后计算池化后两个输出向量之间的余弦相似度。实验发现在该模型下平均池化的效果优于其他两种策略。

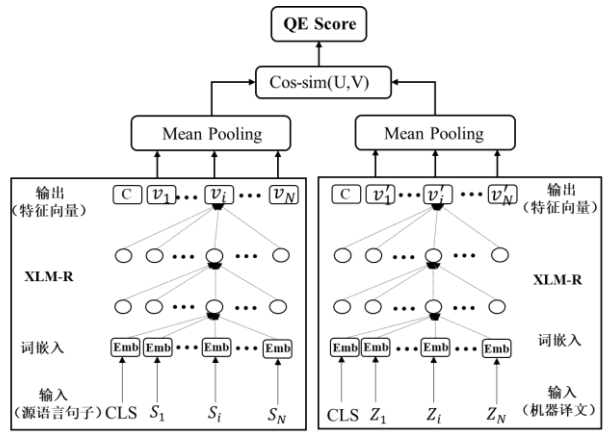

图 7 SiameseTransQuest 模型
Figure 7 SiameseTransQuest model

T Ranasinghe 等人分别在单语言对和多语言对数据集上进行实验，他们发现将高资源语言的知识迁移到低资源语言上有助于提高低资源语言的性能，从而提升模型性能，这也从另一方面验证了迁移学习的有效性。

Amit 等人基于知识蒸馏的方法，以 TransQuest 为教师模型（teacher model),DeepQuest[52]中的双向 RNN（BiRNN）作为学生模型（student model）直接将 TransQuest 的知识转移到 BiRNN 模型中[53]。利用教师-学生框架（teacher-student framework)生成额外的训练数据，即数据增强。具体做法为：使用生成测试数据的机器翻译系统，根据源语言和相关领域的数据生成训练数据，再使用该教师模型生成预测，作为训练学生模型的标签；最后提出了一种基于不确定性量化的机制来过滤掉生成训练数据集中的噪声样本。Li 等人提出基于组置换（group−permutation）的知识蒸馏方法和用于深度神经网络的跳过子层（skipping Sub−layer）正则化方法[54]，实现将深层次模型所学到的知识转移到较小的浅层次模型中。

利用跨语种语言模型提取句子特征进行机器译文质量估计是近两年的研究热点：Hui 等人提出在黑盒设置下执行无监督的译文质量估计[55]，该模型的预测器基于三种不同的预训练模型，即 BERT，XLM 和 XLM-R。编码器的输入为源语言句子和机器翻译输出的拼接。为了在进行句子级质量估计预测时充分利用上下文信息，在预先训练好的模型基础上增加了一层 Bi-RNN 网络。考虑到质量估计直接评估（DA）方向的训练数据十分匮乏，他们使用 MTE 指标来代替人类评估，从而为译文质量估计任务创建大量伪数据，即先使用 MTE 指标来对译文进行评估，估计结果可以替代人工标注的 DA[56]，然后使用伪 DA 分数，

结合源语言句子和机器翻译输出后的句子对，来训练译文质量估计模型；Hu 等人针对迁移学习，多任务学习，模型集成问题在 WMT20 译文质量估计共享任务上提出基于 Multi-BERT，XLM-R 的译文质量估计模型[57]，模型将更深层的基于 Transformer 的机器翻译模型纳入译文质量估计模型中。该模型在 WMT20 译文质量估计评测[58]多项子任务上获得第一名。

4 机器译文质量估计评测与评价

机器译文质量估计评测活动(Quality Estimation Task)发布基准的数据集和评价方法，为不同译文质量估计模型提供了一个公平比较的平台，它极大的促进了译文质量估计的研究。目前机器译文质量估计评测活动主要包括：国际 WMT 译文质量估计评测和国内 CCMT 译文质量估计评测。

WMT 会议最早组织开展译文质量估计评测，从 2012 年至今，刚好 10 年。它主要针对欧洲语言之间互译的质量估计，包括英语-西班牙语、英语-德语和英语-西班牙语等等翻译方向；从 2020 年开始支持英语-汉语翻译方向的译文质量估计。国内很多研究团队参与了该项评测活动，并取得了较好的成绩，包括：（1）在 WMT 18 机器译文质量估计评测任务中阿里团队和江西师范大学团队取得了多个单项上的第一名；（2）在 WMT 19 句子级别机器译文质量估计任务中，南京大学团队在英语-德语的项目上获得第三名；（3）在 WMT 20 句子级别机器译文质量估计任务中，东北大学小牛翻译团队在多个单项上获得第一名。需要说明的是每年评测活动后，评测官方都会公开发布相关数据集供研究者继续使用。

CCMT 会议（2019 年前简称为 CWMT 会议）从 2018 年开始开展译文质量估计评测活动，它主要针对英语-汉语和汉语-英语翻译方向的译文质量估计任务。在 2020 年汉语-英语翻译方向质量估计任务评测中北京交通大学、南京大学和腾讯等参评团队分别取得前三名的好成绩，而英语-汉语翻译方向的前三名分别是北京交通大学、腾讯和南京大学等参评团队。由于对译文质量估计研究的逐步升温，越来越多的研究团队参与了该项评测活动。

在评测活动中，为了比较不同参评模型的优劣，一般使用皮尔森相关系数和斯皮尔曼相关系数定量计算参评系统对译文质量打分和人工对译文质量打分之间的相关性。相关性越高，对应模型越可靠。

皮尔森相关系数 r 计算方法如下：

$$r=\frac{\sum_{i=1}^{N}(y_i-\bar{y})(y_i'-\bar{y'})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2\sum_{i=1}^{N}(y_i'-\bar{y'})^2}} \quad (17)$$

其中，y_i 和 y_i'分别为人工对译文质量打分结果和模型预测译文质量得分；$\bar{y}$ 和 $\bar{y'}$为相应的均值；N 是待估计机器译文的总数。

斯皮尔曼相关系数 ρ 计算方法如下：

$$\rho=1-\frac{6\times\sum_{i=1}^{N}(R(y_i)-R(y_i'))^2}{N\times(N^2-1)} \quad (18)$$

其中，$R(y_i)$和 $R(y_i')$分别是机器译文人工评分排名和预测得分排名。

其它参考的评价指标包括平均绝对值误差(Mean Absolute Error, MAE)和均方根误差(Root Mean Squared Error, RMSE)等等。

为了将各典型方法的性能进行对比，陈等将这些模型在同一数据集即 CWMT2018 的数据集上进行实验[41]，几个典型的模型分别是参加 CWMT2018 句子级别译文质量估计评测任务最优的系统(CWMT18 1st ranked)、基于 RNN 编码器-解码器的联合神经网络模型(UNQE)、基于跨语种预训练词向量方法的模型 TransQuest、基于双语专家的译文质量估计模型(QE Brain)、融合 BERT 语境词向量的译文质量估计模型($CUNQE_{AVE4}$)和基于 BERT 的联合神经网络模型 $TUNQE_{BERT}$，表 1 按照皮尔森相关系数值由小到大的顺序将以上各个模型进行排序。

表 1 各典型方法性能的对比

Table 1 Comparison of the performance of typical methods

模型	平行语料规模	英中	中英
		Pearson	Pearson
CWMT18 1st ranked	CWMT 8M+8M	0.465	0.405
UNQE	CWMT 6M	0.524	0.528
TransQuest	\	0.516	0.568
QE Brain	CWMT 8M	0.588	0.564
$CUNQE_{AVG4}$	CWMT 6M	0.596	0.591
$TUNQE_{BERT}$	CWMT 6M	**0.627**	**0.613**

各个模型在该数据集上的性能如表 1 所示，$TUNQE_{BERT}$ 的性能最优，$CUNQE_{AVG4}$ 的性能次之，而 CWMT18 1st ranked 虽然当年在该评测任务中取到第一名的好成绩，但是性能还是比不上各典型模型。

5 未来研究趋势

机器译文质量估计可以及时准确预测译文质量的优劣，指导翻译系统的开发和推动机器译文应用，它的未来研究方向包括：

（1）更准确的译文质量估计端到端模型，尽管译文质量估计的性能不断提升，特别是基于神经翻译模型的句子级别机器译文质量估计和基于预训练语言模型的句子级别机器译文质量估计极大的提高了译文质量估计的效果，但是在训练语料缺乏的情况下，更准确的译文质量预测模型仍然是研究者孜孜以求的目标。

（2）其是否可以替代传统译文自动评价方法对译文质量进行自动预测？由于传统的译文自动评价方法依赖人工参考译文，使用不方便，能否利用译文质量估计方法对其进行替代，性能能否获得保障，是研究者关注的一个重要问题，也是 WMT 评测近年来重点调查的一项内容。一些初步实验结果表明，神经机器译文质量估计的预测结果在英语-汉语方向上与人工评价的相关性与当前广泛使用的译文自动评价尺度 BLEU 相当[59]，但是它能否推广到其它翻译方向还有待译文质量估计效果的提升，以及进一步的实验和考察。

（3）指导翻译模型的自训练，从统计翻译系统的最小错误率训练[60]到神经翻译系统的最小风险训练[61]，翻译模型的特征权重优化问题一直伴随着机器翻译的研究。由于译文质量估计能够实时的给出译文质量的度量数值且不需要人工参考译文，当翻译模型完成源语言句子翻译后，它能实时计算出译文质量，从而指导翻译系统的参数调整。这种翻译模型的参数优化方式必然极大减少对开发集的依赖，模型甚至可以在测试集上进行自训练，当然这也很大程度上取决于译文质量估计的可靠性。

6 总结

总之，机器译文质量估计是机器翻译研究中一项新的子任务，它在机器翻译的发展和应用中发挥着重要的作用，目前已有取代传统译文自动评价任务的趋势。本文对句子级别机器译文质量估计进行了全面的综述和分析，介绍了三类句子级别译文质量估计方法：以它们代表性的方法为锚点，逐步扩展描述其它相关方法；并对句子级别机器译文质量估计的评测、评价和未来的发展趋势进行了简要介绍。

参考文献

[1] SPECIA L, SHAH K, SOUZA J G C de, et al. QuEst-A translation quality estimation framework[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Sofia: ACL, 2013: 79-84.

[2] CALLISON-BURCH C, KOEHN P, MONZ C, et al. Findings of the 2012 workshop on statistical machine translation[C]//Proceedings of the seventh workshop on statistical machine translation. Montr:ACL, 2012: 10-51.

[3] BOJAR O, BUCK C, CALLISON-BURCH C, et al. Findings of the 2013 workshop on statistical machine translation[C]//Proceedings of the eighth workshop on statistical machine translation. Sofia: ACL, 2013: 1-44.

[4] 俞士汶, 姜新. 机器翻译译文质量自动评估系统[C]. 中国中文信息学会成立十周年学术报告论文集. 北京:中国中文信息学会, 1991:314-319 页.

[5] 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. Cambridge: ACL, 2006： 223-231.

[6] FONSECA E R, YANKOVSKAYA L, MARTINS A F T, et al. Findings of the WMT 2019 shared tasks on quality estimation[C]//Proceedings of the Fourth Conference on Machine Translation. Florence: ACL, 2019: 1-10.

[7] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. Long Beach:NIPS, 2017: 5998–6008.

[9] KIM H, LEE J-H. A. A recurrent neural networks approach for estimating the quality of machine translation output[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 494-498.

[10] FAN K, WANG J, LI B, et al. “Bilingual Expert” Can Find Translation Errors[C]//Proceedings of the AAAI Conference on Artificial Intelligence.

Honolulu:AAAI, 2019: 6367-6374.

[11] 陆金梁, 张家俊. 基于多语言预训练语言模型的译文质量估计方法[J]. 厦门大学学报(自然科学版), 2020, 59(02): 151-158.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. Lake Tahoe:NIPS, 2013: 3111-3119.

[13] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 4171-4186.

[14] CONNEAU A, LAMPLE G. Cross-lingual Language Model Pretraining.[C]// Advances in Neural Information Processing Systems. Vancouver:NIPS, 2019: 7057–7067.

[15] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised Cross-lingual Representation Learning at Scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 8440-8451.

[16] SHAH K, LOGACHEVA V, PAETZOLD G, et al. SHEF-NN: Translation Quality Estimation with Neural Networks[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon: ACL, 2015: 342-347.

[17] RANASINGHE T, ORASAN C, MITKOV R. TransQuest: Translation Quality Estimation with Cross-lingual Transformers[C]//Proceedings of the 28th International Conference on Computational Linguistics. Online: ICCL, 2020: 5070-5081.

[18] 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(08): 1804-1812.

[19] FELICE M, SPECIA L. Linguistic features for quality estimation[C]//Proceedings of the Seventh Workshop on Statistical Machine Translation. Montr: ACL, 2012: 96-103.

[20] OCH F J, NEY H. A systematic comparison of various statistical alignment models[J]. Computational linguistics, 2003, 29(1): 19-51.

[21] QUIRK C. Training a Sentence-Level Machine Translation Confidence Measure[C]// Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisbon: ELRA, 2004: 825-828.

[22] HOKAMP C, CALIXTO I, WAGNER J, et al. Target-centric features for translation quality estimation[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore: ACL, 2014: 329-334.

[23] SCARTON C, SPECIA L Exploring consensus in machine translation for quality estimation[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore: ACL, 2014: 342-347.

[24] BIÇICI E, WAY A. Referential Translation Machines for Predicting Translation Quality[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore: ACL, 2014: 313-321.

[25] BECK D, SHAH K, SPECIA L. Shef-lite 2.0: Sparse multi-task gaussian processes for translation quality estimation[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore: ACL, 2014: 307-312.

[26] ESPLÀ-GOMIS M, SÁNCHEZ-MARTÍNEZ F, FORCADA M L. UAlacant word-level machine translation quality estimation system at WMT 2015[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon: ACL， 2015: 309-315.

[27] 李培芸, 翟煜锦, 项青宇, 等. 基于子词的句子级别神经机器翻译的译文质量估计方法[J]. 厦门大学学报(自然科学版), 2020, 59(02): 159-166.

[28] 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(02): 271-275.

[29] SCHUSTER M, PALIWAL K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997: 2673-2681.

[30] BOJAR O, CHATTERJEE R, FEDERMANN C, et al. Findings of the 2017 Conference on Machine Translation (WMT17)[C]// Proceedings of the Second Conference on Machine Translation: Volume 2, Shared Task Papers. Copenhagen: ACL, 2017: 169-214.

[31] 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. Copenhagen: ACL, 2017: 562-568.

[32] PATEL R N, SASIKUMAR M. Translation Quality Estimation using Recurrent Neural Network[C]//Proceedings of the First Conference on Machine Translation. San Diego: ACL, 2016: 819-824.

[33] SHAH K, BOUGARES F, BARRAULT L, et al. Shef-lium-nn: Sentence level quality estimation with neural network features[C]//Proceedings of the First Conference on Machine Translation. San Diego: ACL, 2016: 838-842.

[34] MARTINS A F T, ASTUDILLO R F, HOKAMP C, et al. Unbabel's participation in the wmt16 word-level translation quality estimation shared task[C]//Proceedings of the First Conference on Machine Translation San Diego: ACL, 2016: 806-811.

[35] 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, 101(9): 2417-2421.

[36] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]//International conference on artificial neural networks. Warsaw:Springer, 2005: 799-804.

[37] 翟社平, 杨媛媛, 邱程, 等. 基于注意力机制 Bi-LSTM 算法的双语文本情感分析[J]. 计算机应用与软件, 2019: 251-255.

[38] SPECIA L, BLAIN F, LOGACHEVA V, et al. Findings of the WMT 2018 Shared Task on Quality Estimation[C]//Proceedings of the Third Conference on Machine Translation. Belgium:ACL, 2018: 689-709.

[39] HOU Q, HUANG S, NING T, et al. NJU Submissions for the WMT19 Quality Estimation Shared Task[C]// Proceedings of the Fourth Conference on Machine Translation. Minneapolis: ACL, 2019: 95-97.

[40] WANG Z, LIU H, CHEN H, et al. NiuTrans submission for CCMT19 quality estimation task[C]//China Conference on Machine Translation. Singapore: Springer, 2019: 82-92.

[41] 陈聪, 李茂西, 罗琪. 译文质量估计中基于 Transformer 的联合神经网络模型[J]. 中文信息学报, 2021, 35(6): 79-86.

[42] CUI Q, HUANG S, LI J, et al. DirectQE: Direct Pretraining for Machine Translation Quality Estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Online: AAAI, 2021, 35(14): 12719-12727.

[43] KEPLER F, TRÉNOUS J, TREVISO M V, et al. OpenKiwi: An Open Source Framework for Quality Estimation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Minneapolis: ACL, 2019: 117-122.

[44] 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. Lisbon: ACL, 2015: 316-322.

[45] RANA R, EPPS J, JURDAK R, et al. Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech[J], 2016.

[46] 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.

[47] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by

generative pre-training[J], 2018.

[48] SCHWENK H. Continuous space translation models for phrase-based statistical machine translation[C]//Proceedings of 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters. Mumbai:IITB, 2012: 1071-1080.

[49] PIRES T, SCHLINGER E, GARRETTE D. How Multilingual is Multilingual BERT?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 4996-5001.

[50] ARTETXE M, SCHWENK H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610.

[51] 李培芸, 李茂西, 裘白莲, 等. 融合 BERT 语境词向量的译文质量估计方法研究[J]. 中文信息学报, 2020, 34(03): 56-63.

[52] IVE J, BLAIN F, SPECIA L. DeepQuest: a framework for neural-based quality estimation[C]//Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics. Belgium:ACL, 2018: 3146-3157.

[53] GAJBHIYE A, FOMICHEVA M, ALVA-MANCHEGO F, et al. Knowledge Distillation for Quality Estimation[EB/OL]. [2021-07-07]. https://arxiv.org/pdf/2107.00411.pdf.

[54] LI B, WANG Z, LIU H, et al. Learning light-weight translation models from deep transformer [EB/OL]. [2020-10-27]. https://arxiv.org/pdf/2012.13866.pdf

[55] HUANG H, DI H, OUCHI K, et al. Unsupervised Machine Translation Quality Estimation in Black-Box Setting[C]//China Conference on Machine Translation. Singapore: Springer, 2020: 24-36.

[56] 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.

[57] HU C, LIU H, FENG K, et al. The niutrans system for the wmt20 quality estimation shared task[C]//Proceedings of the Fifth Conference on Machine Translation. Online: ACL, 2020: 1018-1023.

[58] SPECIA L, BLAIN F, FOMICHEVA M, et al. Findings of the WMT 2020 shared task on quality estimation[C]. Association for Computational Linguistics. Online: ACL, 2020：743-764.

[59] LUO Q, LI M. Research on Incorporating the Source Information to Automatic Evaluation of Machine Translation[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics. Online: CCL, 2020: 414-423.

[60] OCH F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st annual meeting of the Association for Computational Linguistics. Sapporo: ACL, 2003: 160-167.

[61] SHEN S, CHENG Y, HE Z, et al. Minimum Risk Training for Neural Machine Translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. San Diego: ACL, 2016: 1683-1692.

Review of sentence-level quality estimation of machine translation

LUO Lan, HE Xianmin, LI Maoxi*

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

Abstract : In recent years, with the correlation between estimation results and human judgements gradually increasing, quality estimation of machine translation has gained broad attention and high

recognition from machine translation researchers. The paper reviews the sentence-level quality estimation of machine translation, and classifies them into methods based on traditional machine learning, methods based on neural machine transation models, and methods based on pre-trained language models, compares and contrasts the representative works, intersection works of these three quality estimation methods, as well as the development paths of different methods, and introduces the relevant evaluation campaigns and evaluation metrics that push forward the quality estimation research. Finally, the future research direction and development trend of sentence-level quality estimation are prospected, and the conclusions are drawn.

Key words: machine translation; quality estimation; deep neural networks; pre-trained language models; evaluation metrics