文章编号：1003－0077(2022)11－0020－18

机器翻译译文质量估计综述

邓涵铖，熊德意

（天津大学 智能与计算学部，天津 300350）

摘 要：机器翻译译文质量估计(Quality Estimation，QE)是指在不需要人工参考译文的条件下，估计机器翻译系统产生的译文的质量，对机器翻译研究和应用具有很重要的价值。机器翻译译文质量估计经过最近几年的发展，取得了丰富的研究成果。该文首先介绍了机器翻译译文质量估计的背景与意义；然后详细介绍了句子级 QE、单词级 QE、文档级 QE 的具体任务目标、评价指标等内容，进一步概括了 QE 方法发展的三个阶段：基于特征工程和机器学习的 QE 方法阶段，基于深度学习的 QE 方法阶段，融入预训练模型的 QE 方法阶段，并介绍了每一阶段中的代表性研究工作；最后分析了目前的研究现状及不足，并对未来 QE 方法的研究及发展方向进行了展望。

关键词：机器翻译；译文质量估计；文献综述

中图分类号：TP391 **文献标识码**：A

A Survey on Machine Translation Quality Estimation

DENG Hancheng, XIONG Deyi

(College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

Abstract: Machine translation quality estimation refers to the estimation of the quality of the outputs by machine translation system without the human reference translations. It is of great value to the research and application of machine translation. In this survey, we firstly introduce the background and significance of machine translation quality estimation. Then we introduce in detail the specific task objectives and evaluation indicators of word-level QE, sentence-level QE, and document-level QE. We further summarize the development of QE methods to three main stage: methods based on feature engineering and machine learning, methods based on deep learning, and methods integrated with pre-training model. Representative research works in each stage are introduced, and the current research status and shortcomings are analyzed. Finally, we outline the outlook for the future research and development of QE.

Keywords: machine translation; translation quality estimation; literature review

0 引言

机器翻译(Machine Translation，MT)技术在全球化中扮演着十分重要的角色。随着全球化推进，机器翻译技术也在不断地进步，尤其是近些年来神经机器翻译技术(Neural Machine Translation，NMT)的出现，将机器译文质量提升到了新的高度。尽管如今机器翻译技术达到了较高水平，但不少机器译文仍存在着错译、漏译、过译等问题，且无法在机器翻译系统中实时反映给使用者。即当前机器翻译系统尚无法满足人类的翻译需求，尤其是在缺少人类后期编辑(post-editing，PE)干预时。目前机器翻译仅能作为计算机辅助翻译(Computer-Assisted Translation，CAT)的手段之一[1]。如何在使用机器翻译系统的过程中，实时地掌握机器译文质量成了人们研究的问题。因此针对不需要参考译文的机器翻译质量估计(Machine Translation Quality Estimation，MTQE，QE)的研究应运而生。

与如 BLEU[2]、METEOR[3]、NIST[4]、TER[5]

收稿日期：2021-03-30 定稿日期：2021-05-10

基金项目：国家重点研发计划(2019QY1802)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

等需要参考译文来计算对应的指标以评价机器译文质量的方法不同，QE 仅需源语言文本及其经过机器翻译系统生成的目标语言文本，即可自动估计出目标语言文本的翻译质量。Gandraburd 等人[6]受语音识别领域中置信度估计(Confidence Estimation)的启发，最早将置信度估计引入到机器翻译中。Quirk 等人[7]将机器译文句子人工标注为 Ideal、Acceptable、Possibly、Unacceptable 四类质量标签作为数据集，从双语中提取有效特征，并通过 SVM 算法对机器译文进行分类。早期的 QE 任务并没有准确定义，针对 QE 的研究[8-10]大多集中在对统计机器翻译系统本身，且并未形成规模。2009 年，Specia 等人[11]提出了一套包括译文句子人工打分标注、双语特征提取、机器学习算法训练译文分数模型在内的 QE 方案。自 2012 年机器翻译研讨会(Workshop on Machine Translation，WMT)针对该方案正式将译文质量估计作为一项任务[12]，QE 任务开始广泛被研究。发展至今，QE 研究可分为三个阶段：基于特征工程和机器学习的 QE 方法阶段、基于深度学习的 QE 方法阶段、融入预训练模型的 QE 方法阶段。

本文组织结构如下：引言部分主要介绍机器翻译质量估计研究的背景及其意义；第 1 节介绍机器翻译质量估计作为 WMT 的经典任务，在 WMT 中的具体任务描述，包括任务目标、任务所使用的数据集、任务评价指标等内容；第 2 节介绍基于传统机器学习与特征工程的机器翻译质量估计的方法，包括常用的机器学习方法及常见特征；第 3 节介绍翻译质量估计方法过渡到神经方法阶段后出现的主流方法及其存在的问题；第 4 节介绍以 BERT 为代表的自然语言处理预训练模型(Pre-trained Models，PTMs)出现后，融入预训练模型的 QE 方法；第 5 节介绍除从 QE 模型方面改进之外，围绕数据增强展开的 QE 方法；第 6 节讨论目前机器翻译质量估计所面临的一些挑战和未来的研究方向；第 7 节为本文小结。

1　三种不同粒度的 QE 任务描述

按照不同的质量估计粒度划分，QE 任务可分为单词级(word-level)、短语级(phrase-level)、句子级(sentence-level)、段落级(paragraph-level)及文档级(document)五种，其中单词级 QE 任务与短语级 QE 任务较为相似，也被称为亚句子级(subsentence-level)QE 任务[13]。

QE 任务是 WMT 上的一项经典任务，最早作为 WMT 的具体任务出现是在 WMT12 中。此后，不少 QE 的工作都围绕 WMT 上的 QE 任务来展开，所以以下主要按照 WMT 上的 QE 任务，来具体介绍不同粒度 QE 任务的具体内容。由于短语级 QE 任务和段落级 QE 任务分别与单词级 QE 任务和文档级 QE 任务较为相似，均只在 WMT 早期某些年份中少次出现，且最近的研究工作较少围绕短语级 QE 任务及段落级 QE 任务展开，故在本文中只介绍单词级、句子级及文档级三种粒度的 QE 任务。

1.1　单词级 QE 任务

单词级 QE 任务即预测给定机器译文中每一个单词及符号的质量，可以帮助机器翻译系统用户直接了解到翻译不好的位置，帮助后编辑工作者直接定位翻译质量差的单词进行修改。

1.1.1　预测目标

单词级 QE 任务的目标为估计译文中每一单词或标点符号的针对源语言文本的翻译质量。单词级 QE 任务可以被认作是一种有监督的分类任务，可分为二分类目标(Binary Classification)、Level 1 分类目标(Level 1 Classification)和多分类目标(Multi-class Classification)。

单词级 QE 任务二分类的目标是预测译文中每个词或符号的好/坏(OK/BAD)标签，以表示每个词或符号翻译的优劣。自 WMT 2018 起，除预测译文中的词或符号质量外，QE 任务还要求参与者预测词或符号间空格的质量，即判断翻译中有无遗漏单词，并用 BAD 来标注空格以表示有漏译情况，OK 表示无漏译情况。

Level 1 分类目标是在二分类的基础上，将错误翻译的单词(即 BAD 标签所对应的单词)按照多维度质量指标[14](Multidimensional Quality Metrics，MQM)中的一级错误分类(包含准确度错误、流利度错误两类)细粒度化，即预测出翻译中的错误属于准确度错误还是流利度错误。

单词级 QE 任务的多分类目标是在 Level 1 分类目标的基础上将错误翻译更细粒度化，将每个错误翻译的单词都用 MQM 中的细粒度错误类别(大小写、标点、术语、错误翻译、遗漏等)来标记。

Level 1 分类目标与多分类目标在早期的 WMT 中均有出现，但其数据集标注相对于二分类

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

目标的数据集标注更加复杂耗时，并且各参赛系统获得在前两项任务上的效果与二分类目标任务相比差距较大，作为二分类任务外的子任务，较少研究团队参加。所以在WMT15及之后年份的WMT中，单词级QE任务仅采用二分类目标任务作为唯一任务。且相对于难度较大的Level 1分类目标与多分类目标，二分类目标相对简单并以其实用性成为人们在单词级QE任务上的主要研究目标。

1.1.2 数据集

总体来说，单词级、句子级、文档级三种粒度的QE任务所使用的训练集和开发集都包含以下几部分：源语言文本(src)、机器译文文本(mt)、译后编辑文本(pe)、数据标签(labels)。其中，历届WMT中的单词级QE和句子级QE任务都使用同样的src、mt及pe，仅因其预测目标的区别而有不同数据标签，测试集不包含pe及labels。

具体而言，WMT中的单词级及句子级的QE任务数据集一般选取特定领域(新闻、信息科技、制药、生命科学等领域)的不同语言对的平行语料，并使用机器翻译系统对平行语料中的一类语言文本(源语言文本)进行翻译得到目标语言文本，再由专业的译员参照平行语料对目标语言文本后编辑得到译后编辑文本(记作pe)。

不同年份WMT的单词级QE任务因有着不同的预测目标，因而其所使用的数据标签也不尽相同。如今主要使用的是OK/BAD二分类标签，可通过TERCOM[①]工具对比机器译文与译后编辑文本自动计算得来。表1以WMT2019中的英语-德语QE任务为例，展示单词级QE数据集主要内容。

表1 单词级QE任务数据集示例

组成部分	内容
源语言文本(src)	this format is used on Wireless Application Protocol (WAP) pages.
译文文本(mt)	dieses Format wird für PPP (WP)-Seiten verwendet.
译后编辑文本(pe)	dieses Format wird für WAP-Seiten (Wireless Application Protocol) verwendet.
分类标签(labels)	OK OK OK OK OK OK OK OK OK BAD OK OK BAD BAD OK BAD OK BAD OK OK OK OK OK

每个单词都被标记为OK或BAD。此外，在WMT18之后，如果两个单词之间有一个或多个单词需要被插入，那么每个单词之间的间隔都被标记为BAD，否则标记为OK。所以，如果目标句子单词的数量若为N个，则每个目标句子的标签数量是$2N+1$。

1.1.3 评价指标

与其他分类任务相似，单词级QE方法可使用准确率(precision)、召回率(recall)、F_1值(precision和recall的调和平均数)作为评价指标，precision和recall的计算方式如式(1)、式(2)所示。

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{1}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

其中，TP、FP、FN分别表示QE模型预测出的真正类(True Positive)、假正类(False Positive)、假负类(False Negative)的样本数。F_1计算方式如式(3)所示。

$$F_1=\frac{2\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}=\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{3}$$

单词级QE任务的最终评价指标是“OK”和“BAD”类别的F_1值的乘积，记作F_{mult}。

由于数据集中的OK标签远远多于BAD标签，即单词级QE这一分类任务中的类具有非常大的不平衡性。因此在WMT19及之后，马修斯相关系数[15](Matthews correlation coefficient，MCC)也因其在类不平衡时的有效性，作为一项额外的评价指标被引入到单词级QE任务中[16]，其计算如式(4)～式(6)所示。

$$S=\frac{\text{TP}+\text{FN}}{N} \tag{4}$$

$$P=\frac{\text{TP}+\text{FP}}{N} \tag{5}$$

$$\text{MCC}=\frac{\frac{\text{TP}}{N}-\text{SP}}{\sqrt{\text{SP}(1-S)(1-P)}} \tag{6}$$

其中，N表示所有的预测总数，即TP+TN+FP+FN，TN为模型预测的真负类(True Negative)的样本数。

1.2 句子级QE任务

句子级QE任务，旨在对每条翻译的句子进行整体的质量估计，是最早被定义和研究的QE任务[7]，同时因为机器翻译系统最常用于翻译句子上，机器翻译系统通常以句子为输入并处理整条句子，因此句子级机器翻译质量估计具有很高的实用性，

① http://www.cs.umd.edu/~snover/tercom/

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

句子级 QE 任务也是最受各研究团队关注的任务。

1.2.1　预测目标

在 WMT 中，句子级 QE 任务可分为打分和排名两个子任务，其中打分任务是给出一个代表句子质量的绝对分数，而排名任务是对所有机器译文句子按照估计的质量进行排序，打分任务为主要任务。作为最早出现在 WMT 上的 QE 任务发展至今，打分任务本身的预测目标也是一直在变化的，但其始终是围绕将机器译文句子后编辑至可出版水平的所需工作量①出发的。根据 Krings[17] 的观点，后编辑工作量可分时间、认知及技术三个维度。其中时间维度的后编辑工作量是指将机器译文后编辑至可出版水平所需的时间，简称后编辑时间。认知维度指的是从人类（特指人类译员）的认知角度（译员付出的脑力劳动），将机器译文后编辑至可出版水平所需的工作量，具有很强的主观性。技术维度的后编辑工作量指的是将机器译文后编辑至可出版水平所涉及到的技术操作（如插入、删除、替换、移动等）的工作量。

后编辑时间的长短能够直接比较不同机器译文的好坏程度，其作为一种直观的后编辑工作量体现方式，曾作为句子级 QE 子任务的预测目标出现于 WMT13[18] 及 WMT14[19] 中。但是后编辑时间是一项具有很强主观性的指标，在后编辑时间数据标签标注过程中，不同译者可能因翻译经验、熟练水平、打字速度等因素，对于相同机器译文句子，后编辑时间差异较大，后编辑时间同样可能会受因译者个人状况（如分心、劳累）等因素影响。并且，译者在后编辑的过程中还需要时间阅读、修改、校对，这些时间也具有较强的主观性，它们与后编辑时间之间的关系也难以定义。因此，将后编辑时间作为句子级 QE 任务的预测目标缺乏一定的客观性和实用性，WMT 在 2015 年及之后不再将预测译文后编辑时间来作为句子级 QE 的子任务，辅助的数据标签出现于 WMT16-WMT18 句子级 QE 任务中。

认知维度的后编辑工作量主要由人类译者对译文后编辑工作量打分来间接体现，又称感知后编辑工作量（perceived post-editing effort[19]）。例如，在 WMT12[12] 中，句子级 QE 任务的预测目标为基于李克特量表的 1～5 分的质量分数[20]，其中 1 分表示译文无法进行后编辑，需要从头开始翻译，2～4 分分别表示约 50%～70%、25%～50%、10%～25%的译文需要后编辑，5 分表示译文清晰易懂，几乎不需要后编辑。在 WMT14[19] 中，句子级 QE 任务的预测目标为基于李克特量表的 1～3 分的质量分数，与之前不同的是，分数越低表示译文需要的后编辑工作量越少，译文质量越高，1 分表示无需任何后编辑的完美译文，2 分表示译文中包含的错误不超过 3 个及可能带有一些易于修正的简单错误（如大小写、标点符号等），3 分表示译文质量非常低，且无法轻易修正。认知维度的后编辑工作量同样具有很强的主观性，同一译文句子需要多个译者（后编辑者）进行打分标注，而不同的译者对于相同的译文句子打分可能差异较大，因此认知维度的后编辑工作量（感知后编辑工作量）作为数据标签是耗时耗力且不稳定的，不适合作为句子级 QE 任务的预测目标。

技术维度的后编辑工作量中最常用且最具代表性的衡量指标是人工翻译编辑率（Human-targeted Translation Edit Rate[5]，HTER），是翻译编辑率（Translation Edit Rate[5]，TER）的变种。TER 的计算方法为机器译文到参考译文的最小编辑（插入、删除、替换、移动等四类编辑）次数除以其若干条非定向参考译文（Untargeted Reference Translations）的平均长度。HTER 的计算方式同样为最小编辑次数与参考译文的比值，但其参考译文为经过人工后编辑的定向参考译文②，由人类译者参考非定向参考译文以了解语义后，本着最少编辑次数的原则对机器译文进行后编辑得来。HTER 相对于 TER 有更小及更客观的最小编辑次数，能更合理地反映机器译文的质量。HTER 的计算如式(7)所示。

$$\mathrm{HTER}=\frac{I+D+S+\mathrm{Sh}}{R} \tag{7}$$

其中，I，D，S，Sh 分别代表插入（Insert）、删除（Delete）、替换（Substitute）、移动（Shift）操作的次数，R 代表定向参考译文中单词的个数。

HTER 的范围在[0,1]之间，其值越高，表示译文需要修改的次数越多，质量越差。相比于其他指标它更能直观且客观地反映机器译文所需工作量。因此从 WMT13[18] 开始，预测机器译文的 HTER 成为句子级 QE 的一项子任务，后于 WMT15 开始成为句子级 QE 唯一打分任务，并沿用至今。

1.2.2　数据集

如 1.1.2 节中所提到，单词级 QE 和句子级 QE 任务共用数据集中的 src、mt 及 pe。使用 TERCOM

① 简称后编辑工作量，Post-Editing Effort。

② Human-target Reference Translation，即 1.1.2 节中提到的 PE。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

工具即可自动计算出句子级QE任务所需的HTER标签。除此之外，如1.2.1节中提到的，部分数据集中还有部分额外的如后编辑时间、后编辑者键盘点击次数等辅助数据标签。

1.2.3 评价指标

不同的子任务及不同的预测目标有着不同的评价指标。早期的句子级QE任务采用平均绝对误差(MAE)作为主要评价指标，均方根误差(RMSE)作为次要评价指标。同时使用DeltaAvg，Spearman作为排名任务的评价指标。

Graham等人[21]指出，若QE模型的预测结果中方差较高，它将导致较高的平均绝对误差，即使是在预测结果的分布遵循真实标签分布的情况下。该问题在用于句子级别QE的数据集中很常见，因此建议使用皮尔逊相关系数r(Pearson correlation coefficient)作为句子级QE预测HTER任务的评价指标，其计算方法如式(8)所示。

$$r=\frac{\sum_{i=1}^{N}(H(T_i)-\bar{H})(V(T_i)-\bar{V})}{\sqrt{\sum_{i=1}^{N}(H(T_i)-\bar{H})^2\sum_{i=1}^{N}(V(T_i)-\bar{V})^2}} \tag{8}$$

其中，$H(T_i)$和$V(T_i)$分别表示机器译文T_i的质量自动估计得分和对应的真实质量标签；$\bar{H}$和$\bar{V}$分别为其均值；N表示测试集中机器译文句子总数量。皮尔逊相关系数r最大为1，其值越大说明QE模型质量估计结果与机器译文真实质量越吻合。

Souza等人[22]指出，由于皮尔逊相关系数r使用的前提假设之一是两个变量均服从正态分布，而句子级QE任务中的HTER标签往往并非呈正态分布，因此将皮尔逊相关系数r作为句子级QE任务的唯一评价指标是不可靠的，建议将MAE与皮尔逊相关系数r结合考虑，以更好评价句子级QE模型的效果。

1.3 文档级QE任务

文档级(又称篇章级)QE任务是指在没有人工参考译文的情况下对给定的翻译文档进行质量估计，其文档泛指包含多个句子(3个句子及以上)的文本。

1.3.1 预测目标

文档级QE任务自2016年作为一项新任务出现于WMT，发展至今主要分为两类预测目标，一类预测是WMT16[23]中采用的两阶段后编辑方法[24]计算质量得分，另一类是预测WMT18[25]中采用的由多维度质量指标(Multidimensional Quality Metrics，MQM)计算得来的MQM分数及译文文档中句子级的MQM错误标签。

两阶段后编辑方法是Scarton等人[24]从文档级特性出发提出的一种衡量文档级译文后编辑工作量的方法。在第一阶段，句子顺序被随机打乱，然后由译员进行后编辑，记作PE1；在第二阶段，将PE1中的句子按顺序放回原处，由译员将其作为整个文档考虑其篇章特性，再次进行后编辑，记作PE2。该方法的动机是将文档级QE与句子级QE区分开来，体现句子之间的衔接性和连贯性。然后译文到PE1和PE2的后编辑代价HTER分别记为$PE_1\times MT$、$PE_2\times MT$。但Bojar等人[23]发现$PE_1\times MT$与$PE_2\times MT$差值较小，难以体现文档级的后编辑代价，而$PE_2\times PE_1$值较大，表明当只考虑文档级别的错误时，文档的变化更大，显然忽略了单词及句子级别的问题，影响整个文档的质量。因此，Bojar等人[23]提出了设置权重将$PE_1\times MT$和$PE_2\times PE_1$线性组合衡量文档级后编辑代价，计算方法如式(9)所示。

$$f=w_1\cdot PE_1\times MT+w_2\cdot PE_2\times PE_1 \tag{9}$$

其中，w_1及w_2为权重，由经验得来，WMT16中的设置为$w_1=1$，$w_2=13$。

与以上这种基于后编辑代价来估计文档级译文质量不同。基于MQM模型的预测目标的出发点是估计译文文档中的翻译错误程度，参与者被要求预测基于MQM错误类型及错误严重程度计算得来的MQM分数。MQM模型将译文中的错误分为轻微错误(minor)、重大错误(major)、严重错误(critical)三种严重程度，由专业译员参考译文中单词级的错误并按照Sanchez-Torron等人[26]提出的方法进行标注分类。MQM分数计算如式(10)所示。

$$MQM=1-\frac{n_{minor}-5n_{major}-10n_{critical}}{n} \tag{10}$$

其中，n_{minor}、n_{major}、$n_{critical}$分别代表文档中轻微错误、重大错误、严重错误的个数，n表示整个文档的单词数。MQM越大，表示译文质量越高，上限为1(即译文中无任何错误)，若错误很严重，MQM分数有可能为负数。

此外，自WMT19开始，参与者还被要求预测机器译文文档中单词级的错误类型(与1.2.1节类似)。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.3.2　数据集

不同的预测目标对应着带有不同数据标签的数据集。基于两阶段后编辑方法质量分数的数据标签以及 MQM 分数标签都由专业译员标注得来。与单词级和句子级 QE 任务数据集共用 src、mt、pe 不同，文档级 QE 任务的基本单位为文档（至少包含 3 个句子）。一般而言，带有 MQM 分数标签的文档级 QE 任务数据还带有细粒度错误（单词级错误）类型标注及错误严重程度标注。

近期的文档级 QE 任务（WMT18-20）使用的数据集均基于亚马逊产品评论数据集[①]（Amazon Product Reviews dataset），源语言文本来自亚马逊网上最受欢迎的英文的体育和户外产品名称和描述，将其经过最先进的机器翻译系统翻译得到法语机器译文，并由 Unbabel 团队人员标注获得 MQM 分数标签。

1.3.3　评价指标

预测两阶段后编辑质量得分及预测 MQM 得分都与预测 HTER 的句子级 QE 任务同样采用皮尔逊相关系数作为主要评价指标，评价预测值与真实值的相关性，同时采用 MAE、RMSE 作为辅助评价指标。而预测机器译文文档中单词级的错误类型的评价指标则与单词级 QE 任务的相同，同样使用 F_1 值作为评价指标。

2　基于特征工程与机器学习的 QE 方法

如引言部分所提及，早期针对 QE 任务的研究未形成规模，QE 任务未被准确定义，学界对机器译文质量也尚未形成统一分类标准。随着 QE 任务被纳入 WMT12，QE 任务逐渐形成基于特征工程与机器学习的研究框架，该框架将 QE 任务定义为有监督的回归/分类预测任务。基于特征工程与机器学习的 QE 方法框架如图 1 所示，其核心部分为特征提取模块与机器学习模块，特征提取模块用以对源语言及机器译文文本进行特征提取及特征选择，该过程又称特征工程；机器学习模块通过提取好的特征在训练阶段可对不同粒度的数据标签进行训练拟合，学习到特征与质量标签的关系，即可在预测阶段实现对机器译文的质量估计。此阶段相关的工作主要围绕特征工程和机器学习算法的选择两方面进行展开。

2.1　基于特征工程的 QE 方法

基于特征工程的 QE 方法主要从两方面出发，

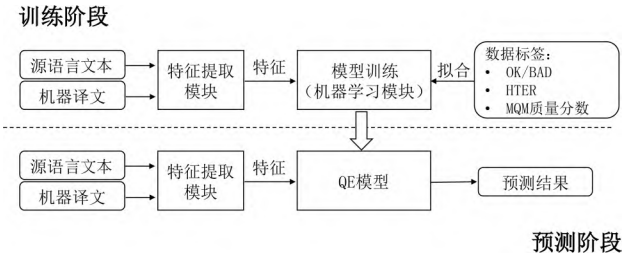

图 1　基于特征工程与机器学习的 QE 方法框架

一是特征抽取，即如何从源语言及机器译文文本中提取与翻译质量相关的特征；二是特征选择与特征过滤，即在众多特征中选取与机器译文质量最为相关的特征。

在特征提取方面，2013 年 Specia 等人[27]提出的 QuEst 模型[②]将 QE 任务使用的特征归纳为了四大类：复杂度特征，流利度特征，忠实度特征，置信度特征（图 2）。其中，复杂度特征由源语言得来，主要反映源语言文本的复杂程度与翻译难度，例如源语言句子长度、源语言句子语言模型概率等；流利度特征由机器译文中得来，包括机器译文句子长度、机器译文句子语言模型概率等；忠实度特征则由源语言文本与机器译文共同得来，以反映翻译是否将源语言文本中的意思完整保留及表达，主要包括源语言句子与译文句子长度比、源语言句子与译文句子中各类词性单词个数比值等特征，以上这三类特征与具体机器翻译系统本身无关，又被称为黑盒特征（Black-box Features）。置信度特征由机器翻译系统得来，依赖于机器系统本身，又被称为白盒特征（Glass-box Features），例如机器翻译系统本身对输出译文的打分、n-best 列表中不同翻译假设（hypotheses）的个数、译文中短语的平均长度等。

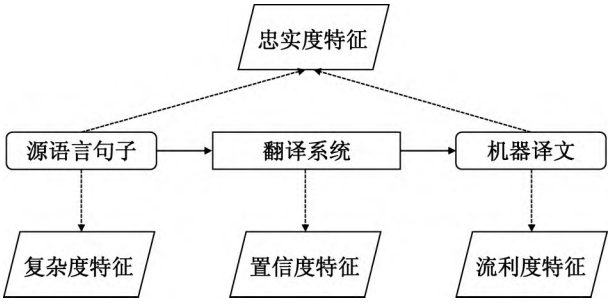

图 2　QuEst 框架特征分类

除 QuEst 模型中提出的的四类特征外，不少团

① http://jmcauley.ucsd.edu/data/amazon/

② 在 2015 年改进为 QuEst++[28]，并作为 WMT13-18 的基线模型。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

队对其他种类的句子级 QE 特征提取展开了研究，其中最为常见的是基于语言学的特征和伪参照译文（pseudo references）、回译[29]（back-translation）特征。Almaghout 等人[30]主要采用了组合范畴语法（Combinatory Categorial Grammar，CCG）特征，他们认为与上下文无关的短语结构语法形式相比，CCG 更适合处理 SMT 文本。他们将翻译分成从 CCG 解析图中提取的最大语法块，使用 CCG 特征来估计翻译的合乎语法性，并对比了 CCG 特征与基线特征在法语-英语和阿拉伯语-英语数据集上的实验效果，结果表明 CCG 特征优于基线特征。Langlois 等人[31]提出的 LORIA 系统首次引入基于潜在语义索引（Latent Semantic Indexing，LSI）的特征来衡量源语言与目标语言的词汇相似性，并加入了基于伪参照译文特征来判断其与机器译文的相似性。Kozlova 等人[32]研究了更为传统的句法特征对源语言与目标语言句法解析树的作用，提取如宽度（来自根节点的依赖数量）、最大深度、内部节点比例、主语数量、关系子句等与句法树、句型、词性标注等相关的句法特征，并将源语言文本输入在线机器翻译系统以获取伪参考译文，又将伪参考译文回译为源语言文本，最后再针对以上提及的文本进行特征设计。Abdelsalam 等人[33]基于词对齐和双语分布式表示，为句子级 QE 任务引入了一组新特征。Sagemo 等人[34]使用不同工具提取词对齐特征、词性（POS）特征、基于短语结构的特征、语言模型特征，并通过量化名词翻译错误、重新排序措施、语法一致性和结构完整性来获得体现 SMT 系统翻译难点的一致性特征。

除了基于语言学的特征和伪参照译文特征外，Biçici 等人[35-36]基于可识别训练语料和测试语料之间翻译行为的参考翻译机器模型[37]（referential translation machines，RTMs），直接估计翻译输出质量，判断文本之间语义相似度的方法，该方法无须依赖 SMT 系统信息及语言学分析，并通过特征衰减算法（feature decay algorithms，FDA5）在大量的候选平行语料中选择与已经给出的训练和测试语料之间翻译行为相同的语料，添加到训练语料中。Shah 等人[38]除了使用 QuEst++中的基线特征外，还使用神经网络提取了连续空间语言模型特征（将在第 3.1 节展开）。

以上特征提取方法主要针对句子级 QE 任务，对句子级 QE 基线方法 QuEst 使用的特征进行扩充。Luong 等人[39]针对单词级 QE 任务使用了基于系统的（图拓扑、语言模型、对齐上下文等）、词法的（词性标签）、语法的（成分标签、到成分树根的距离）和语义的（多义词计数）特征。除了 SMT 系统的现有组件外，还使用了其他外部工具和资源进行特征提取。例如，TreeTagger（用于获取词性标签）、使用 AnCora treebank 训练的 Bekerley parser（用于西班牙语生成组成树）、WordNet 和 BabelNet（用于多义词计数）、谷歌翻译等。该文提出的特征为 WMT15-WMT18 单词级 QE 任务主要特征，并可使用 MARMOT 工具①进行抽取。

在特征选择方面，González-Rubio 等人[40]指出基于特征工程的 QE 方法存在着特征集高度冗余的问题，特征之间有高度的多重共线性，有些特征可能与预测质量分数无关；且由于特征的数量和种类很多，而且训练集通常相对较小，因此需要对特征进行降维操作，提出了偏最小二乘回归的特征降维方法，并在文献[41]中提出主成分分析的特征降维方法，通过在不同 QE 模型上的实验发现特征降维方法能显著提升模型性能。Shah 等人[42]使用高斯过程（Gaussian Process，GPs）在 82 个特征中选取了前 20 个特征，且取得较好实验结果。除此之外，特征过滤同样能在一定程度上解决特征冗余的问题，Langlois 等人[31]采用反向算法[43]（backward algorithm）过滤无效的特征。

2.2 基于传统机器学习的 QE 方法

对于句子级 QE 任务预测 HTER 这种表示为连续分数的标签，回归算法是自然的选择。一系列的如逻辑回归[30]、M5P 算法[44]、局部最小二乘法[11]、高斯过程回归[45-46]、极端随机树[47]、单层和多层感知机[8,48-49]、岭回归[35]、支持向量机[50]、基于多项式核的支持向量回归算法[51]均被探索应用于句子级 QE 任务中。Tezcan 等人[52]通过实验对比了在相同实验设置及相同特征工程下基于支持向量机、线性回归模型、随机森林（Random Forest，RF）这三类机器学习算法，实验结果显示支持向量机的效果最好。

单词级 QE 任务被定义为有监督的分类模型，由于单词级的 QE 任务总在译文句子内进行，因此针对单词级传统 QE 方法的研究可分为非序列类和序列类[13]两类模型。非序列类模型将句子中每一单词独立看待，不考虑单词间的相互依赖性，序列类

① https://github.com/qe-team/marmot

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

模型则是在进行 QE 任务时关注到单词所在的句子序列信息，即上下文信息。许多标准的机器学习模型都可用于非序列模型训练，Singh 等人[53]及 Esplà-Gomis 等人[54]使用随机森林分类器学习训练数据的决策树集成，Rubino 等人[55]使用了支持向量，Esplà-Gomis 等人[56]及 Tezcan 等人[52]使用多层感知机作为单词级 QE 的分类器，非序列模型早期展现了比较好的实验效果，但因忽略上下文信息，逐渐被持续发展的序列模型超越。例如，Esplà-Gomis 等人[56]在 WMT15 中单词级 QE 任务排名第一，但其在文献[54]中的改进版本在 WMT16 中单词级 QE 任务排名下降到第七。

序列模型中最常用的是条件随机场(Conditional Random Fields，CRF)模型[57]，它类似于生成隐马尔可夫模型，其中任何变量的值都以其邻居的值为条件，能够较好地对单词级 QE 任务进行上下文建模。Luong 等人[39]将 CRF 首次应用于单词级 QE 任务，并在 WMT13-14 上取得较好成绩。但近些年来，用以构建序列模型的 CRF 逐渐被循环神经网络(Recurrent Neural Network，RNN)所替代(基于神经网络的 QE 方法将在第 3 节中展开)。

2.3　问题与挑战

基于特征工程与传统机器学习的 QE 方法的核心在于特征工程，但特征提取和特征选择严重依赖于人们对语言对的语言学分析，并需要进行人工特征设计，若没有强大的语言学分析及人力资源，难以对其开展研究。此外，不同语言对及不同粒度的译文有着截然不同的特征，即特征抽取耗时耗力且难以复用，缺乏在不同语言中的通用性。同时，特征的选择及抽取本身就存在较大误差，大量带有误差的特征导致误差在模型中累积，从而导致模型在 QE 任务上表现较差，且难以突破该框架本身对 QE 任务建模的能力。

3　基于深度学习的 QE 方法

3.1　利用神经网络进行特征提取的 QE 方法

随着神经网络和深度学习技术在自然语言处理领域的初步应用，词嵌入[58-59](又称词向量)技术以及神经网络机器翻译(Neural Machine Translation，NMT)模型的出现，一些研究团队开始将神经网络用于 QE 任务中的特征提取。Shah 等人[38,42]除了使用 QuEst 中的传统手工特征外，还使用基于词袋模型(Continuous Bag-of-Words，CBOW)的 Word2Vec[58]工具提取词嵌入，以及计算源语言和目标语言单词在彼此词嵌入空间映射的相似度，作为单词级 QE 任务的额外特征；在句子级 QE 任务上，将训练连续空间语言模型[60](Continuous Space Language Model，CSLM)所产生的语言模型概率作为特征，并与传统特征相结合，文献[38]的实验结果显示加入了 CSLM 特征的模型的句子级 QE 效果比未加入 CSLM 特征的模型更好。Shah 等人[61]在此基础上将 NMT 系统产生的基于对数似然估计的条件语言模型概率特征与 CSLM 提取的句子向量和交叉熵特征、由 QuEst 提取的传统手工特征相结合，较文献[38]中的方法取得了句子级 QE 任务上更好的实验结果。

Chen 等人[62-63]在 Shah 等人工作的基础上，使用多种方法提取词嵌入特征，并使用算术平均、TF-IDF 加权平均、最小值、乘法等 4 种方法将词嵌入合成为句子向量特征，并且使用循环神经网络的语言模型提取语言模型概率特征，将句子向量特征与语言模型概率特征结合，进一步提升了神经网络在 QE 任务特征提取上的表现。此外，Abdelsalam 等人[33]和 Scarton 等人[64]也将词向量特征结合传统特征分别应用在了句子级和文档级 QE 任务中。

另一方面，一些研究者尝试完全使用神经网络进行特征提取并进行质量估计，Kreutzer 等人[65]提出的 QUETCH 方法利用基于多层感知机的深度前馈神经网络在平行语料上无监督地训练一个将目标词分类为 OK/BAD 的二分类模型，然后将其用于单词级 QE 任务，且以固定大小的滑动窗口形式输入若干个目标词将向量拼接，以将上下文双语表示信息传入下游的前馈神经网络中，该方法属于不依赖传统手工特征而完全使用神经网络提取特征的 QE 方法，且取得了较好的实验效果，但其实验效果并不如融入了在此基础上的传统特征的 QUETCH＋方法。Martins 等人[66]在 QUETCH 的基础上，加入双向门控循环单元(Bidirectional Gated Recurrent Units，BiGRU)网络并堆叠前馈神经网络对 QUETCH 中的神经网络架构进行改进，并加入了源与目标语言输入的词性(Part of Speech，POS)特征，取得了优于基于传统特征 QE 方法的实验结果。Patel 等人[67]也在 QUETCH 的基础上提出一种基于 RNN 的 QE 方法，同样使用了预训练词向量的方法，并基于滑动窗口输入双语单词序列，分别使用

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

LSTM、GRU两种RNN变体提取双语序列的表示；并针对单词级QE训练集中OK/BAD标签的不平衡问题，借用了Shang等人[68]提出的细粒度化标签方法，根据单词在句子中的位置将OK标签分为更细粒度的三类OK标签，以达到均衡标签分布的目的，其实验结果证明了其改进的有效性。除此之外，Paetzold等人[69]、Patel等人[67]也都分别在单词级和句子级QE任务中使用了RNN提取特征。

3.2 完全基于神经网络模型的QE方法

Kreutzer等人[65]提出的QUETCH方法虽然使用了神经网络来进行特征提取，但他们在模型输入部分均采用的是基于滑动窗口以保留双语上下文信息的方法，需要源语言文本和译文之间每一个单词及符号的对齐信息，然而QE数据集中语言对之间的对齐信息本身就是由基于统计方法的工具提取的，具有较大误差，对QE效果造成巨大影响。

随着深度学习技术和计算设备的进一步发展，端到端的神经机器翻译方法[70]被提出，并取得了极大的进展，且其效果也超越了统计机器翻译模型。因此，在QE领域，人们也开始思考完全基于神经网络的模型的QE方法，即无须提取手工特征的方法。

Kim等人[71-72]提出将基于双向RNN并引入注意力机制[70](Attention Mechanisms)的机器翻译模型应用到句子级QE任务上，是第一个“纯神经网络”QE方法。2017年，Kim等人[73]将其命名为预测器-估计器(Predictor-Estimator，PredEst)模型，PredEst模型首先对引入注意力机制的RNNSearch[70] NMT模型进行了改进，并将RNNSearch解码器部分改为双向RNN。如图3所示，PredEst模型分为两个模块，并分两个阶段分别对两个模块进行训练：(1)第一阶段，使用大规模平行语料训练单词预测器(Word Predictor)模块(如图3第Ⅰ部分所示)，词预测器的任务是根据输入的源语言及目标语言信息，来预测目标语言中心词，近似一个NMT模型。与NMT模型不同的是，在解码阶段PredEst模型通过双向RNN不仅能接收到从左到右的目标语言信息，同时也能获取从右到左的目标语言信息，可充分利用目标语言上下文信息，更加符合QE任务的实际。在预测中心词的过程中，我们可以充分获取源端和目标端的双向上下文信息，以提取每一特定中心词的质量向量，该质量向量包含了当前位置应该被正确预测为中心词单词的信息；(2)第二阶段，使用带有质量标签的QE数据训练质量估计器(Quality Estimator)模块(如图3第Ⅱ部分所示)，首先将QE数据中的双语文本输入到预测器中，以提取机器译文句子每一单词的质量向量，再将译文单词质量向量逐一输入到估计器模块，经过估计器中的RNN模型输入QE数据中的质量标签，由估计器对机器译文质量向量及质量标签进行拟合。在预测阶段，输入源语言与目标语言，经过预测器提取译文质量，经过估计器即可进行译文质量标签预测。为了有效训练神经网络，Kim等人[74]利用一种堆栈传播(Stack Propagation)算法，针对单词级QE任务、短语级QE任务和句子级QE任务对神经网络进行联合训练。

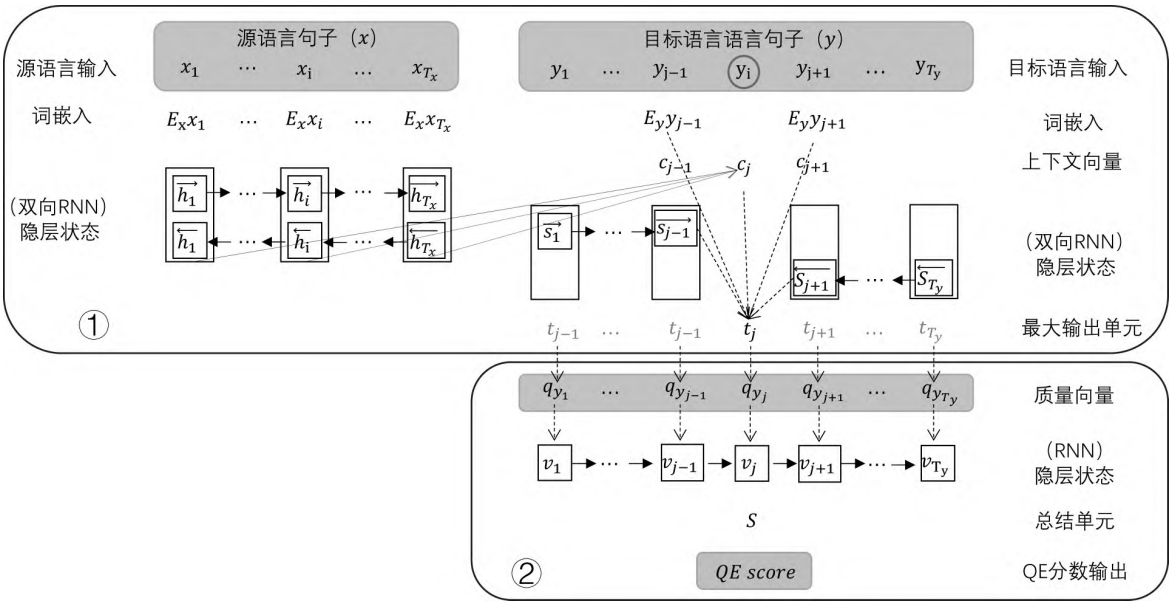

图3 预测器-估计器模型框架

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Li 等人[75-76]针对预测器-估计器模型中两个网络需独立训练的问题，将预测器-估计器框架重构为联合神经网络框架，提出了端到端的 QE 方法，并称之为 UNQE 联合神经模型。

Martins 等人[77-78]在 WMT17 中针对单词级 QE 任务提出了由一个含有大量手工特征的序列线性模型 LINEARQE 和一个神经网络模型 NEURALQE 堆叠而成的 STACKEDQE 模型。LINEARQE 线性模型集成了一元特征（依赖单个输出标签）、二元特征（依赖连续输出标签）、句法特征（POS 标注等）等三大类特征，并使用 MIRA 算法[79]来对计算特征权重。NEURALQE 纯神经网络模型在输入层除输入源文及译文句子外，还增加了词对齐、词嵌入及 POS 特征，并通过多次堆叠前馈神经网络及双向 GRU 获取上下文向量。作者将上述两个模型堆叠成为 STACKEDQE 模型，取得了较两个单独模型更好的实验结果。并在此基础上堆叠自动后编辑（Automatic Post-editing，APE）系统将其拓展为 FULLSTACKEDQE 以进行句子级 QE 任务。该方法取得 WMT 17 的单词级 QE 任务上取得第二名的好成绩，但在句子级 QE 任务上较于同年提出的预测器-估计器模型有较明显差距。Hu 等人[80]针对单词级 QE 任务，在 NEURALQE[77]的基础上提出了对目标词的局部和全局上下文信息进行有效编码的方法，并将之命名为上下文编码 QE 模型（Context Encoding Quality Estimation，CEQE）。该模型由三部分神经网络组成，第一部分为词嵌入层，用于对目标中心词进行表征，除与 NEURALQE 模型词嵌入层一样使用了 POS 特征、词对齐信息外，还加入了目标词相邻词及对应源文相邻词词向量，丰富了目标中心词局部上下文信息；第二部分为一维卷积层，用于为每个目标中心词集成局部上下文信息；第三部分由前馈神经网络和循环神经网络堆叠而成，用于对句子中全局上下文信息进行编码。该方法取得 WMT 18 单词级 QE 任务中 6 个语言方向中 3 个方向第一名的优异成绩，但与同期的“双语专家”模型（将于 3.3 节介绍）在另 3 个语言方向上的单词级 QE 任务效果差距明显。

3.3　双语专家（Bilingual Expert）模型

随着带有自注意力（self-attention）机制的 Transformer 模型[81]在机器翻译领域的广泛应用，Wang 等人[82]在预测器-估计器模型框架的基础上引入 Transformer 模型，加强目标语言与源语言关系的建模，并设计了判断机器译文正确与否的人工特征，在 WMT18 单词级及句子级 QE 任务的所有参赛语言方向中取得最好成绩。

该模型基于预测器-估计器模型架构，包含词预测和质量估计两个模块，同样需要分开在两个阶段训练，Fan 等人[83]将基于大规模平行语料训练的词预测器类比为精通双语的专家，并将该模型命名为双语专家模型（Bilingual Expert），如图 4 所示。在词预测模块中，双语专家模型使用 Transformer 编码器代替预计器中的 RNN 编码器，使用双向 Transformer 解码器代替预测器中解码器的双向 RNN 解码器。该模块类似机器翻译系统，Transformer 结构的使用加强了模型对目标语言与源语言关系的建模，并可避免因输入序列过长而导致 RNN 产生的计算问题。但与基于 Transformer 结构的机器翻译模型不同，双语专家模型使用的双向 Transformer 解码结构增加了后向自注意力机制，使模型在预测中心词任务时，分别使用前向自注意力机制和后向自注意力机制，融入前文及后文的目标语言信息，该设置更贴近 QE 任务实际应用场景。词预测模块经过训练后可提取出上下文隐层状态 z 和上下文词向量 e 两种特征。除此之外，作者针对中心词的预测设计了一个用以衡量机器译文中心词与双语专家模型预测中心词间差距的特征，名为错误匹配特征（mis-matching Feature），该特征通过比较译文中心词和模型预测中心词概率分布得来。质量估计模块使用了广泛用于序列标注和序列分类任务的双向 LSTM 模型，将由词预测模块的所有特征拼接为一个向量输入到双向 LSTM 中，即可进行句子级 HTER 回归任务及单词级序列标注任务。由于双语专家模型预测的中心词可被视作参考译文，作者指出可扩展模型以支持结合 QE 和 APE 的多任务学习。

此外，作者还尝试在双向 LSTM 层后添加额外的 CRF 层，但其对原始模型实验结果并无显著改善；还尝试使用自注意力机制模块代替代双向 LSTM，但实验结果反而变差。作者认为是第一阶段用以训练词预测模块的平行语料与第二阶段用以训练质量估计模块的 QE 数据间巨大数据量差异导致了这种结果。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

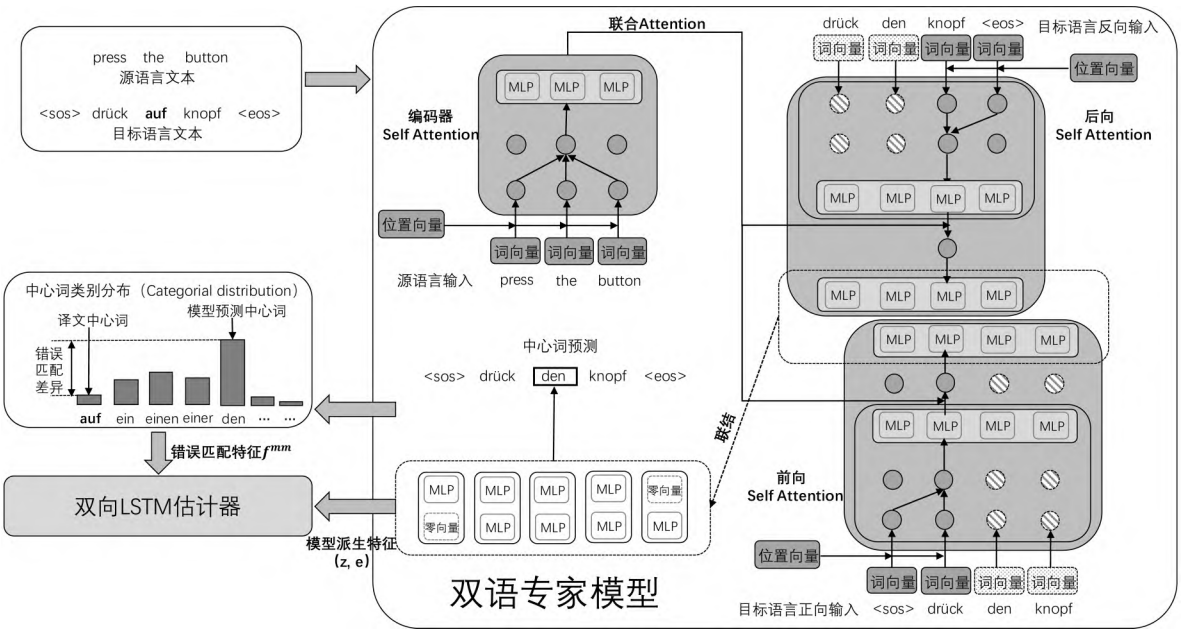

图4 双语专家模型框架

3.4 问题与挑战

PredEst和双语专家模型都是基于预测器-估计器模型架构的QE方法，并在不同时期取得了QE研究领域内最好的实验结果。该框架展现了强大的双语关系建模及特征提取能力，因此成为完全基于神经网络模型QE方法的主流框架。但两阶段训练数据存在巨大数据量差异，从预测器中提取的特征由大量平行语料训练得来，而由数据量小很多且翻译质量参差不齐的QE数据训练而得来的估计器难以利用好这些特征。其次，双语专家模型依赖大规模平行语料进行训练，而不少语种之间缺乏平行语料，因此这一类基于预测器-估计器模型架构的QE方法同样具有难以扩展到其他语种对间的挑战。

4 融入预训练模型的QE方法

融入预训练模型的QE方法，又可称为基于迁移学习的QE方法。随着ELMo[84]、BERT[85]、XLM[86]等大规模预训练语言模型①的出现与应用及发展，尤其是在大规模平行语料上基于掩码(Mask)训练的BERT出现，并在一些下游任务上的表现远超原有方法，一些研究工作开始尝试将预训练模型融入QE模型中，以更好地提取源语言文本和译文文本的质量向量，从而达到提高QE准确度的目的。

Kepler等人[87]分别使用了BERT、XLM等预训练模型代替了预测器-估计器模型框架中的预测器，并对比了基于BERT、XLM、双语专家模型的QE模型的实验效果，实验发现融合了跨语言知识的基于XLM的QE模型性能最好，并获得了当时WMT19中QE任务的最好成绩[16]。Hou等人[88]提出了两种QE模型：双向翻译QE模型和基于BERT的QE模型，双向翻译QE模型利用回译文本从两个不同的翻译方向运用两种语言之间的翻译知识，基于BERT的QE模型则从源端和目标端获取额外的单语知识，该模型取得了WMT19上句子级QE任务的较好成绩。Zhou等人[89]对比使用双语专家模型、ELMo模型、BERT模型在QE上的效果，其中基于ELMo模型的QE方法取得了当时最好效果，他们猜测是因为ELMo减少了目标语言下文的可见信息，使得预测器对中心词预测更加困难，并迫使模型更关注源语言信息，获得更多来自源语言的特征。

Yankovskaya等人[90]使用BERT和LASER[91]；两种预训练模型得到的向量作为回归神经网络模型的特征，并进一步提出了使用机器翻译系统的对数概率作为输入特征，与BERT提取到的向量特征、LASER提取到的向量特征一并输入前馈神经网络进行融合，实验证明了机器翻译系统对数概率特征的有效性。Mathur等人[92]提出了一种基于预训练模型语境向量的无监督机器译文自动评价方法，其

① 以下简称预训练模型

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

实验结果与人类评价相关度较高，说明在不依赖参考译文的情况下对机器译文质量进行判断是可行的，同时也提示我们使用无监督学习方法研究机器译文质量估计的可能性。

Miao 等人[93]基于 BERT 提出了三种融合预训练模型的 QE 方法：①将 BERT 与双语专家模型各自提取的特征融合的混合整合模型(Mixed Integration Model)；②基于 BERT＋LSTM＋MLP 的直接整合模型(Direct Integration Model)；③使用对齐知识约束机制的约束整合模型(Constrained Integration Model)。Miao 等人推测直接整合模型方法可能太依赖于预先训练的语言模型，且有可能学习到一些有偏差的特征，没有充分考虑平行句子对的对齐知识，所以提出了一种约束方法，在预测质量分数时，添加使用一个对齐知识对模型进行约束，实验表明添加了约束条件的模型性能更优。

Wu 等人[94]在 WMT20 上提交的系统集成了两个模型：用平行语料训练的基于 Transformer 的 PredEst 模型和经过微调的基于 XLM 的 PredEst 模型。在这两种模型中，预测器部分均作为特征提取器。基于 XLM 的预测器产生两种上下文表示：掩码表示和非掩码表示，基于 Transformer 的预测器仅产生非掩码表示。估计器采用 Transformer 或 LSTM 训练，将具有不同模型和具有不同参数的同一模型的系统集成在一起，以生成单个句子级的预测。该方法在 WMT20 上英-中句子级 QE 任务上取得最好成绩。

Wang 等人[95]提交在 WMT20 上的系统同样使用 PredEst 架构，使用一个经 WMT 新闻翻译任务的平行语料预训练的不带随机掩码的 Transformer 作为预测器，估计器部分针对特定任务(单词级/句子级)使用特定分类器，采用多任务学习的统一模型对单词和句子级 QE 任务进行联合训练。Wang 等人还指出，由于 QE 数据集与平行语料相比较小，若网络中所有权值均被更新，则容易出现过拟合的现象，因此使用了瓶颈适配器层[96](Bottleneck Adapter Layers)，以保持与训练好的 Transformer 参数固定，以提高迁移学习效率，防止过拟合。

融入了预训练模型的 QE 方法展现了强大的针对 QE 任务的建模能力，但训练预训练模型的庞大数据量和参数量对硬件资源要求较高。不少研究团队无法独立地进行规模庞大的预训练，只能使用其他团队公开发布的预训练模型，为 QE 的研究工作带来了局限性。

5　基于数据增强的 QE 方法

除了在模型上对 QE 方法进行改进之外，由于 QE 数据的稀缺性，一个很自然的想法是使用数据增强的方法来提升 QE 的效果。在一定程度上来说，无论是在模型中使用 Word2Vec、GloVe[59]等外部工具提取的词向量，还是基于 PredEst 结构，使用大规模平行语料训练估计器，还是融入预训练模型，都可算作数据增强的方法。具体说来，当前基于数据增强的 QE 方法可以从以下几个方面出发：①使用额外的平行语料；②伪参照译文及回译方法使用；③伪数据标签构造；④伪后编辑译文的生成。

Kim 等人[73]提出的预测器-估计器模型，在预测器训练阶段，需要大规模的平行语料进行预训练，帮助预测器学习跨语言信息，并在质量估计阶段进行知识迁移，以应对当前 QE 数据集规模较小的问题，但预训练平行语料与 QE 数据集中带噪声的语料之间的巨大数据量差异所导致的 QE 模型无法很好拟合质量标签的问题，也亟待解决。Liu 等人[97]采用平行语料训练额外的机器翻译系统，并对一部分平行语料进行 N-best 解码，最后将机器翻译系统的输出作为 QE 模型的训练数据，在最大边际似然估计的框架下，进行形式化训练，以扩充带噪声的 QE 数据。

Wu 等人[98]提出了一种拟合 QE 数据中错误类型分布的伪数据标签构造方法，不依赖外部机器翻译系统及预训练，使用平行语料进行 QE 数据增强。首先统计 QE 数据集中的插入、删除、替换、移动四类错误的分布，然后选取平行语料中与 QE 数据 TF-IDF 相似度的较高的句子对，并在选取出的句子对的目标语言句子中根据错误分布构造错误，以达到构造 QE 伪数据的目的，因错误分布相同，故伪数据的 HTER 标签也与原 QE 数据相似。该方法较为新颖，且无须训练额外的机器翻译系统或使用大规模语料预训练模型，但由于 QE 伪数据的构造完全由机器自动生成，仅考虑翻译错误的分布而进行构造，无法模拟具体翻译错误的产生及真实的机器翻译场景。

受 Back-Translation 的启发，Junczys-Dowmunt 等人[99]提出一种基于 Round-Trip Translation 的翻译后编辑数据集增强方法，使用 TERCOM 工具对比伪后编辑译文数据集与伪机器译文数据集计算

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

HTER便可得到QE数据集。该方法首先训练一个机器翻译系统,然后使用单语语料进行两次翻译,两次翻译后的原始单语语料可被当作后编辑译文。该方法巧妙地自动获取伪后编辑译文数据及伪HTER数据,但两次翻译会使机器翻译系统中的误差叠加,严重影响数据集质量。受Martins等人[77]APE-QE启发,Kepler等人将APE系统的输出作为伪后编辑文本,并使用TERCOM工具自动生成单词级及句子级质量标签。

Wang等人[95]假设机器译文到参照译文的"距离"约等于机器译文到后编辑译文的"距离"加上后编辑译文到参考译文的"距离",利用APE系统或其他在线翻译系统生成不同的伪参照,并将质量稍差的伪参照作为伪后编辑译文,便可计算伪数据标签以进行QE任务,该方法被称为伪后编辑译文辅助QE方法(Pseudo-PE assisted QE,PEAQE)。实验表明,加入伪后编辑译文数据进行QE任务,与只使用SRC和MT文本相比显著提高了模型性能。

6 未来发展及挑战

总体说来,经过近20年的发展,QE领域的研究取得了长足的进步,尤其是近年来深度学习和神经机器翻译技术的发展带动了QE研究的快速发展,与此同时也伴随着新的挑战,主要有以下问题亟待解决。

(1)本文概述的QE模型主要为句子级别的QE任务模型,单词级与文档级的研究工作相对于句子级QE任务少很多。基于深度学习和迁移学习虽然使单一模型可以进行多任务学习,但较少工作的出发点围绕单词级和文档级QE展开,尤其是单词级QE面向辅助机器翻译或以后编辑较句子级QE更有实用性的情况下,单词级QE和文档级QE理应更受到关注。

(2)如Tu等人[100]所指出,尽管NMT在翻译质量上有了显著的提高,但它往往存在过翻译和欠翻译的问题。在机器译文自动评价(Machine Translation Evaluation)领域中,Yang等人[101]针对NMT中过翻译和欠翻译现象提出的自动评价指标OTEM和UTEM弥补了BLEU等指标只能对译文质量进行机械式评价,而无法针对特定语言现象进行评价的缺陷。目前QE领域暂未出现针对特定语言现象研究的译文质量的工作,该方法为我们提供了从机器译文的具体语言现象着手,并更具有解释地进行质量估计的角度。同时,如何将机器译文估计(QE)与机器翻译自动评估(Evaluation)更好地结合,提高QE与人类评价的相关度,或者利用无监督学习及零资源学习的方式,无须参考译文即可估计译文的质量,也是值得我们思考的问题。

(3)辅助译后编辑作为机器翻译质量估计技术的主要应用,体现了APE任务与QE任务是可以互相促进的。理论上,QE任务的预测结果可直接输入到APE系统判断译文是否需要后编辑及完成自动后编辑,而另一方面,APE任务的输出结果也可被QE系统所用,以生成QE任务的质量标签,同时使质量标签更具有解释性。因此,能否使用强化学习等方法使QE模型向APE模型拓展,如何将APE任务与QE任务有效结合,需要我们进一步探究。

(4)融入了预训练模型的QE方法展现了强大的针对QE任务的建模能力,但训练预训练模型的庞大数据量和参数量对硬件资源要求较高。不少研究团队无法独立地进行规模庞大的预训练,只能使用其他团队公开发布的预训练模型,为QE的研究工作带来了局限性。但它还是没有突破PredEst模型的局限性,如何解决预训练数据与QE数据之间巨大的数据鸿沟及如何让更多的源语言信息参与到第二阶段估计器的训练中来,仍然是目前QE研究领域悬而未决的问题。

(5)在数据增强方面,现阶段采用的方法均为使用平行语料库或预训练语言模型等外部资源提前扩增QE数据来达到扩增数据的目的,能否在QE系统中利用现成的QE数据资源自动地生成额外的QE数据,以达到实时的数据增强效果,是值得尝试的方向。

7 小结

机器翻译的质量估计作为一种不需要参考译文就能实时评估机器译文质量的应用,有着很强的实用性,并且能够促进机器翻译本身的发展。本文对机器翻译的质量估计进行了全面的分析和介绍。根据历年WMT中QE任务中的变化,介绍了从句子级、单词级、文档级三个粒度QE任务的具体概念和细节,并将QE方法发展过程归纳为基于传统机器学习、基于深度学习、融入预训练语言模型方法的三个阶段,详细介绍了每一阶段相关研究工作的进展,对各类方法的优点和局限性进行了归纳,并从方法

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

和数据两个方面，对 QE 方法的发展进行了详细介绍和总结，最后针对当前 QE 任务研究工作所存在的问题及挑战提出了未来潜在的研究方向。

参考文献

[1] BARRACHINA S, BENDER O, CASACUBERTA F, et al. Statistical approaches to computer-assisted translation[J]. Computational Linguistics, 2009, 35 (1): 3-28.

[2] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[3] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[4] DODDINGTON G. Automatic evaluation of machine translation quality using *N*-gram co-occurrence statistics[C]//Proceedings of the 2nd International Conference on Human Language Technology Research, 2002: 138-145.

[5] SNOVER M, DORR B, SCHWARTZ R, et al. A study of translation edit rate with targeted human annotation [C]//Proceedings of Association for Machine Translation in the Americas. 2006: 223-231.

[6] GANDRABURD S, FOSTER G. Confidence estimation for translation prediction[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, 2003: 95-102.

[7] QUIRK C. Training a sentence-level machine translation confidence measure [C]//Proceedings of the LREC, 2004: 825-828.

[8] BLATZ J, FITZGERALD E, FOSTER G, et al. Confidence estimation for machine translation[C]//Proceedings of the 20th International Conference on Computational Linguistics, 2004: 315-321.

[9] UEFFING N, NEY H. Word-level confidence estimation for machine translation[J]. Computational Linguistics, 2007, 33 (1): 9-40.

[10] UEFFING N, NEY H. Word-level confidence estimation for machine translation using phrase-based translation models[C]//Proceedings of HLT/EMNLP, 2005: 763-770.

[11] SPECIA L, TURCHI M, CANCEDDA N, et al. Estimating the sentence-level quality of machine translation systems[C]//Proceedings of the 13th Conference of the European Association for Machine Translation, 2009: 28-37.

[12] CALLISON-BURCH C, KOEHN P, MONZ C, et al. Findings of the workshop on statistical machine translation[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, 2012: 10-51.

[13] SPECIA L, SCARTON C, PAETZOLD G H. Quality estimation for machine translation[J]. Synthesis Lectures on Human Language Technologies, 2018, 11 (1): 1-162.

[14] LOMMEL A, USZKOREIT H, BURCHARDT A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics[J]. Revista Tradumàtica, 2014 (12): 455-463.

[15] MATTHEWS B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme [J]. Biochimica at Biophysica Acta -protein Structure, 1975, 405 (2): 442-451.

[16] FONSECA E, YANKOVSKAYA L, MARTINS A F, et al. Findings of the WMT shared tasks on quality estimation[C]//Proceedings of the 4th Conference on Machine Translation, 2019: 1-10.

[17] KRINGS H P. Repairing texts: Empirical investigations of machine translation post-editing processes [M]. 5. Kent State University Press, 2001.

[18] BOJAR O, BUCK C, CALLISONBURCH C, et al. Findings of the Workshop on Statistical Machine Translation[C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 1-44.

[19] BOJAR O, BUCK C, FEDERMANN C, et al. Findings of the Workshop on Statistical Machine Translation[C]//Proceedings of the 9th Workshop on Statistical Machine Translation, 2014: 12-58.

[20] SPECIA L. Exploiting objective annotations for minimising translation post-editing effort[C]//Proceedings of the 15th Annual Conference of the European Association for Machine Translation, 2011: 73-80.

[21] GRAHAM Y. Improving evaluation of machine translation quality estimation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1804-1813.

[22] de Souza J G. Adaptive Quality Estimation for Machine Translation and Automatic Speech Recognition [D]. PHD Thesis, University of Trento, 2016.

[23] BOJAR O, CHATTERJEE R, FEDERMANN C, et al. Findings of the conference on machine translation [C]//Proceedings of the 1st Conference on Machine Translation, 2016: 131-198.

[24] SCARTON C, ZAMPIERI M, VELA M, et al. Searching for context: A study on document-level labels for translation quality estimation[C]//Proceed-

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

ings of the 18th Annual Conference of the European Association for Machine Translation, 2015: 121-128.

[25] SPECIA L, BLAIN F, LOGACHEVA V, et al. Findings of the WMT shared task on quality estimation [C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 689-709.

[26] SANCHEZ-TORRON M, KOEHN P. Machine translation quality and post-editor productivity[C]// Proceedings of AMTA, 2016: 267-272.

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

[28] SPECIA L, PAETZOLD G, SCARTON C. Multi-level translation quality prediction with quest++[C]// Proceedings of ACL-IJCNLP System Demonstrations, 2015: 115-120.

[29] RAPP R. The backtranslation score: Automatic mt evalution at the sentence level without reference translations[C]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009: 133-136.

[30] ALMAGHOUT H, SPECIA L. A CCG-based quality estimation metric for statistical machine translation [C]//Proceedings of MT Summit XIV of Conference, 2013: 223-230.

[31] LANGLOIS D. LORIA system for the WMT quality estimation shared task[C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 323-329.

[32] KOZLOVA A, SHMATOVA M, FROLOV A. Ysda participation in the wmt'16 quality estimation shared task[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 793-799.

[33] ABDELSALAM A, BOJAR O, ELBELTAGY S R. Bilingual embeddings and word alignments for translation quality estimation[C]//Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers, 2016: 764-771.

[34] SAGEMO O, STYMNE S. The UU submission to the machine translation quality estimation task[C]// Proceedings of the 1st Conference on Machine Translation, 2016: 825-830.

[35] BIÇICI E, WAY A. Referential translation machines for predicting translation quality[C]//Proceedings of the 9th Workshop on Statistical Machine Translation, 2014: 304-308.

[36] BIÇICI E. RTM-DCU: Predicting semantic similarity with referential translation machines[C]//Proceedings of the 9th International Workshop on Semantic Evaluation, 2015: 56-63.

[37] BIçICI E, GROVES D, VAN GENABITH J. Predicting sentence translation quality using extrinsic and language independent features[J]. Machine Translation, 2013, 27 (3-4): 171-192.

[38] SHAH K, NG R W M, BOUGARES F, et al. Investigating continuous space language models for machine translation quality estimation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1073-1078.

[39] LUONG N Q, LECOUTEUX B, BESACIER L. LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT[C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 386-391.

[40] GONZÁLEZ-RUBIO J, NAVARRO CERDÁN J R, CASACUBERTA F. Dimensionality reduction methods for machine translation quality estimation[J]. Machine Translation, 2013, 27 (3-4): 281-301.

[41] GONZÁLEZ-RUBIO J, SANCHIS A, CASACUBERTA F. PRHLT submission to the WMT quality estimation task[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, 2012: 104-108.

[42] SHAH K, LOGACHEVA V, PAETZOLD G, et al. SHEF-NN: Translation quality estimation with neural networks[C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 342-347.

[43] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3 (3): 1157-1182.

[44] SORICUT R, BACH N, WANG Z. The SDL language weaver systems in the WMT quality estimation shared task[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, 2012: 145-151.

[45] BECK D, SHAH K, COHN T, et al. SHEF-Lite: When less is more for translation quality estimation [C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 337-342.

[46] COHN T, SPECIA L. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 32-42.

[47] DE SOUZA J G, BUCK C, TURCHI M, et al. FBK-UEdin participation to the WMT13 quality estimation shared task[C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 352-358.

[48] BUCK C. Black box features for the WMT quality estimation shared task[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, 2012: 91-95.

[49] HILDEBRAND A S, VOGEL S. MT quality estimation: the CMU system for WMT '13[C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 373-379.

[50] FELICE M, SPECIA L. Linguistic features for quality

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

estimation[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, 2012: 96-103.

[51] HARDMEIER C, NIVRE J, TIEDEMANN J. Tree kernels for machine translation quality estimation[C]//Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, Canada, 2012: 109-113.

[52] TEZCAN A, HOSTE V, MACKEN L. UGENTLT3 SCATE submission for WMT shared task on quality estimation[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 843-850.

[53] SINGH A K, WISNIEWSKI G, YVON F. LIMSI submission for the WMT'13 quality estimation task: An experiment with *N*-Gram posteriors[C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 398-404.

[54] ESPLà-GOMIS M, SáNCHEZ MARTíNEZ F, FORCADA M. UAlacant word-level and phrase-level machine translation quality estimation systems at WMT [C]//Proceedings of the 1st Conference on Machine Translation, 2016: 782-786.

[55] RUBINO R, WAGNER J, FOSTER J, et al. DCU-symantec at the WMT quality estimation shared task [C]//Proceedings of the 8th Workshop on Statistical Machine Translation, 2013: 392-397.

[56] ESPLAGOMIS M, SÁNCHEZMARTÍNEZ F, FORCADA M L. UAlacant word-level machine translation quality estimation system at WMT[C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 309-315.

[57] LAFFERTY J, MCCALLUM A, PEREIRA F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning 2001: 282-289.

[58] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 746-751.

[59] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.

[60] SCHWENK H. Continuous space translation models for phrase-based statistical machine translation[C]// Proceedings of COLING: Posters, 2012: 1071-1080.

[61] SHAH K, BOUGARES F, BARRAULT L, et al. SHEFLIUM-NN: Sentence level quality estimation with neural network features[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 838-842.

[62] CHEN Z, TAN Y, ZHANG C, et al. Improving machine translation quality estimation with neural network features[C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 551-555.

[63] **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.

[64] SCARTON C, BECK D, SHAH K, et al. Word embeddings and discourse information for quality estimation[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 831-837.

[65] KREUTZER J, SCHAMONI S, RIEZLER S. QUality estimation from ScraTCH: Deep learning for word-level translation quality estimation[C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 316-322.

[66] MARTINS A F T, ASTUDILLO R, HOKAMP C, et al. Unbabel's participation in the WMT word-level translation quality estimation shared task[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 806-811.

[67] PATEL R N, SASIKUMAR M. Translation quality estimation using recurrent neural network[C]//Proceesings of the 1st Conference on Machine Translation, 2016: 819-824.

[68] SHANG L, CAI D, JI D. Strategy-based technology for estimating MT quality[C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 348-352.

[69] PAETZOLD G, SPECIA L. Simplenets: Quality estimation with resource-light neural networks[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 812-818.

[70] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations, 2015: 1-15.

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

[72] KIM H, LEE J H. Recurrent neural network based translation quality estimation[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 787-792.

[73] 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 Low-resource Language Information Processing, 2017, 17 (1): 1-22.

[74] 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 2nd Conference on Machine Translation, 2017: 562-568.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[75] LI M, XIANG Q, CHEN Z, et al. A unified neural network for quality estimation of machine translation [J]. Ieice Transactions on Information, 2018, 101 (9): 2417-2421.

[76] **李培芸,翟煜锦,项青宇**, et al. **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 (2): 159-166.

[77] MARTINS A F, JUNCZYSDOWMUNT 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.

[78] MARTINS A F, KEPLER F, MONTEIRO J. Unbabel's participation in the WMT translation quality estimation shared task [C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 569-574.

[79] CRAMMER K, DEKEL O, KESHET J, et al. Online passive aggressive algorithms[J] Journal of Machine Learning Research,2006,7: 551-585.

[80] HU J, CHANG W C, WU Y, et al. Contextual encoding for translation quality estimation[C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 788-793.

[81] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.

[82] WANG J, FAN K, LI B, et al. Alibaba submission for WMT18 quality estimation task[C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 809-815.

[83] FAN K, WANG J, LI B, et al. "Bilingual Expert" can find translation errors[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 6367-6374.

[84] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227-2237.

[85] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.

[86] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 7059-7069.

[87] KEPLER F, TRÉNOUS J, TREVISO M, et al. Unbabel's participation in the WMT translation quality estimation shared task[C]//Proceeding of the 4th Conference on Machine Translation, 2019: 78-84.

[88] HOU Q, HUANG S, NING T, et al. NJU submissions for the WMT quality estimation shared task [C]//Proceedings of the 4th Conference on Machine Translation, 2019: 95-100.

[89] ZHOU J, ZHANG Z, HU Z. SOURCE: SOURce-conditional elmo-style model for machine translation quality estimation[C]//Proceedings of the 4th Conference on Machine Translation, 2019: 106-111.

[90] YANKOVSKAYA E, TäTTAR A, FISHEL M. Quality estimation and translation metrics via pre-trained word and sentence embeddings[C]//Proceedings of the 4th Conference on Machine Translation, 2019: 101-105.

[91] 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.

[92] MATHUR N, BALDWIN T, COHN T. Putting evaluation in context: Contextual embeddings improve machine translation evaluation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2799-2808.

[93] MIAO G, DI H, XU J, et al. Improved quality estimation of machine translation with pre-trained language representation[C]//Proceeding of the CCF International Conference on Natural Language Processing and Chinese Computing, 2019: 406-417.

[94] WU H, WANG Z, MA Q, et al. Tencent submission for WMT quality estimation shared task[C]//Proceeding of the 5th Conference on Machine Translation, 2020: 1062-1067.

[95] WANG M, YANG H, SHANG H, et al. HW-TSC's participation at WMT quality estimation shared task [C]//Proceeding of the 5th Conference on Machine Translation, 2020: 1056-1061.

[96] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]// International Conference on Machine Learning, 2019: 2790-2799.

[97] LIU L, FUJITA A, UTIYAMA M, et al. Translation quality estimation using only bilingual corpora[J]. IEEE/ACM Transactions on Audio, Speech, Language Processing,2017,25 (9): 1762-1772.

[98] WU H, YANG M, WANG J, et al. Target oriented data generation for quality estimation of machine translation[C]//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, 2019: 393-405.

[99] JUNCZYSDOWMUNT M, GRUNDKIEWICZ R. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing [C]//Proceedings of the 1st Conference on

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Machine Translation，2016：751-758.

[100] TU Z，LU Z，LIU Y，et al. Modeling coverage for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics，2016：76-85.

[101] YANG J，ZHANG B，QIN Y，et al. Otem&Utem：Over-and under-translation evaluation metric for NMT[C]//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing，2018：291-302.

邓涵铖(1996—)，**硕士研究生，主要研究领域为机器翻译。**

E-mail：hcdeng@tju.edu.cn

熊德意(1979—)，**通信作者，博士，教授，主要研究领域为自然语言处理、机器翻译、多语言信息获取。**

E-mail：dyxiong@tju.edu.cn

第五届“大数据安全与隐私计算”学术会议成功举办

2022 年 11 月 18 日至 20 日，由中国中文信息学会大数据安全与隐私计算专业委员会主办、厦门大学信息学院承办、厦门市美亚柏科信息股份有限公司协办的“第五届大数据安全与隐私计算学术会议”在厦门市线下和线上同步成功举办。会议组织了多场特邀专家报告、竞赛报告、论文报告及企业报告。会议对接互联网＋、数字经济、人工智能、数据安全、个人信息保护等国家发展战略，围绕“万物智慧互联、信息受控共享”的隐私计算与数据安全开展学术交流。会议在线下、腾讯会议、B 站同步进行，共吸引了来自高校、科研机构、企业的 600 余名专家、学者和学生参与。

会议开幕式由厦门大学信息学院副院长袁飞教授致欢迎词。中国中文信息学会理事长方滨兴院士代表学会对专委会 2018 年成立以来持续推动前沿理论与技术研究、坚持举办年度学术会议、创办隐私计算与数据安全挑战赛、推动技术落地应用和人才培养给予了充分肯定，并期望专委会和各位委员锐意创新，为国家数字经济发展战略提供强有力的技术支撑。

专委会主任委员李凤华研究员对专委会发展和历届学术会议进行了回顾，发布了专委会推荐的隐私计算和数据安全学术内涵、学术方向和研究重点，并希望专委会积极发挥学术引领作用，传播正能量，带动产业界沿着正确的技术路线落地应用。

李凤华研究员在开幕式上还宣布了 2022 隐私计算与数据安全挑战赛获奖名单。本次竞赛分为隐私计算赛道和数据安全赛道，各设立一等奖 1 项、二等奖 2 项、三等奖 3 项。由中国科学院信息工程研究所和西安电子科技大学联合参赛的《基于隐私计算的 OFD 文档隐私控制系统》、战略支援部队信息工程大学参赛的《跨域数据安全的医疗影像联合智能诊断系统》分获隐私计算、数据安全赛道一等奖。竞赛奖金赞助方海康威视王滨副总裁、航天信息研究院林文辉院长也出席了开幕式。

开幕式后，李凤华研究员作了题为“隐私计算研究进展与发展趋势”的主旨报告。会议还邀请了华为技术有限公司金意儿教授、厦门大学黄联芬教授、复旦大学杨珉教授、香港科技大学(广州)黄欣沂教授、武汉大学王骞教授、厦门大学肖亮教授、上海交通大学郁昱教授、美亚柏科牛军、蚂蚁集团李滴春做特邀报告。报告内容涵盖异构计算场景下可信执行环境构建、基于隐私保护的人行为检测、开放网络环境中的智能系统安全、后斯诺登时代密码攻击与防御、智能系统数据安全、物联网隐私保护、混淆电路的高效安全设计等方面。

此外，会议还安排了两个竞赛一等奖获奖作品报告、12 个录用论文报告，从不同视角介绍了隐私计算和数据安全的最新理论与技术进展。

专委会确定第六届大数据安全与隐私计算学术会议定于 2023 年 10 月 20 日至 22 日在杭州由浙江理工大学信息科学与工程学院承办。

（中国中文信息学会）

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net