机器翻译质量评估:方法、应用及展望

王均松[1] 庄淙茜[1] 魏勇鹏[2]

(1. 西北工业大学 外国语学院,陕西 西安 710061;2. 北京语智云帆科技有限公司,北京 100089)

摘 要:随着机器翻译译后编辑模式在语言服务行业的广泛应用,机器翻译质量评估的重要性日益凸显。本文探讨了业界三种主流机器翻译质量评估方法的优劣以及适用性。研究发现,人工评估、有参考自动评估、无参考自动预估三种方法各有其优势与不足,在特定场景下应根据需求不同灵活地选择评估方法。在未来,机器翻译质量评估研究应在提高自动评估准确率、增加质量评估维度、开展垂直细分领域评估等方面深入挖掘。

关键词:机器翻译;翻译质量评估;人工评估;自动评估

中图分类号:H085 **文献标志码**:A **文章编号**:1674-6414(2024)03-0135-10

0 引言

近年来,机器翻译技术取得了突破性进展,尤其是神经机器翻译技术的出现大幅度提升了机器翻译的质量,对翻译行业的生产模式产生了极为深刻的影响。随着机器翻译在行业中的广泛应用,机器翻译译后编辑(MTPE)逐渐取代纯人工翻译,成为翻译和语言服务行业的新业态(王华树 等, 2019;王均松 2023)。在这一背景下,开展面向机器翻译的质量评估研究具有重要的意义:一方面,客观、可信的翻译质量评估结果可以帮助用户更加合理有效地选择和使用机器翻译引擎,“纠正出现不准确的机器译文及其所带来的潜在风险”(张霄军 等, 2021: 5);另一方面,评估结果还可以用于指导研究人员不断改进机器翻译结果,寻找最具潜力的技术发展方向。自 2000 年以来,世界机器翻译大会(WMT)、中国机器翻译大会(CCMT)等机器翻译领域的权威会议都设置了机器翻译质量评测任务。然而,目前的研究大多偏重自动评估技术开发,忽略了具体情境下的实践应用。鉴于此,本研究将

收稿日期:2023-10-16

基金项目:国家社会科学基金项目“神经网络机器翻译的译后编辑量化系统模型研究”(19BYY128)、教育部人文社会科学规划基金项目“基于受控语言的机器翻译译前编辑研究”(23YJAZH139)的阶段性成果

作者简介:王均松,男,西北工业大学外国语学院副教授,博士,主要从事认知翻译学、翻译技术研究。

庄淙茜,女,西北工业大学外国语学院硕士研究生,主要从事笔译实践研究。

魏勇鹏,男,北京语智云帆科技有限公司总经理,主要从事语言智能技术服务研究。

引用格式:王均松,庄淙茜,魏勇鹏. 机器翻译质量评估:方法、应用及展望[J]. 外国语文,2024(3):135 – 144.

从实践的角度探讨业界主流评估方法的优势、不足以及场景适用性,以期推动机译质量评估由理论方法向实践应用转化。

长期以来,翻译质量评估(Translation Quality Assessment, TQA)一直是翻译学领域一个非常重要的研究课题。自 20 世纪 70 年代开始,国内外学者逐渐意识到质量评估在译学中的重要性,并围绕翻译质量评估原则和模式开展了一系列研究。其中,国外比较有影响力的研究成果有“文本类型评估原则”(Reiss, 1971)“功能 – 语用翻译质量评估模式”(House, 1997, 2015)“基于论辩图式的评估模式”(Williams, 2004)等。而国内的翻译质量评估研究起步稍晚,比较有代表性的评估模式包括“语用标记等效模式”(侯国金, 2004)、“功能语言学评估模式”(司显柱, 2004, 2016)、“关联理论评估模式”(何三宁, 2015)等。上述评估原则和模式引起了学界的极大关注,并对后续的翻译质量评估理论与实践产生了深远的影响。然而,这些模型的评估对象多为人工翻译,很少涉及机器翻译。

近年来,随着机器翻译技术的迅猛发展以及翻译生产模式的变革(王均松 等, 2023),面向机器翻译质量的评估研究步伐不断加快,涌现出许多新技术和新方法。其中,机器翻译质量的自动评估是当前最具挑战性的任务。围绕这一主题,研究者们开展了大量探索性研究,并取得了丰硕的成果(Specia et al. , 2013; Turchi et al. , 2014; Graham et al. , 2017; 陆金梁 等, 2020)。然而,目前大多数研究主要聚焦于评估技术本身,试图通过改进模型或算法提高评估结果的准确性,较少涉及不同评估方法的横向对比。一些综述类研究(张霄军, 2007; 李良友 等, 2014; 肖桐 等, 2020)虽不同程度地涉及了机器翻译质量评估的各种方法,但缺乏对不同评估方法的应用场景分析。

本文首先介绍目前主流的机器翻译质量评估方法,比较各自的优势与不足,然后结合具体的应用场景分析不同评估方法的适用性,并对机器翻译质量评估的前景作出展望。

1 机译质量评估方法分类

目前,主流的机器翻译质量评估方法大致分为“人工评估”和“自动评估”两大类,而自动评估又可以分为“有参考译文的自动评估”和“无参考译文的自动预估”两种。一般而言,业界倾向于采取三分法,将机译质量评估方法分为人工评估、有参考自动评估、无参考自动预估(见图 1)。

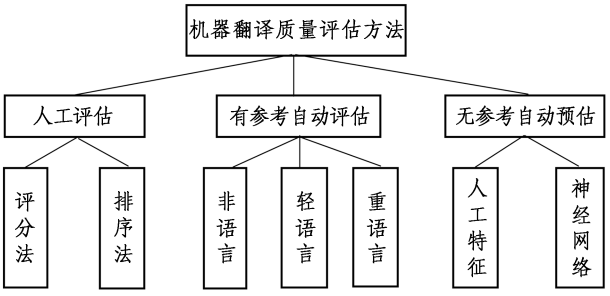

图 1 机器翻译质量评估方法分类

1.1 人工评估

人工评估是指根据一定的评估标准或指标对机器翻译译文进行人工评价,评估方法可以分为“评分法”和“排序法”两种。

(1)评分法

评分法是评价人工译文的一种常见方法,但同样适用于机器翻译质量评估。根据评分标准的不同,可以分为“直接评分”和“错误扣分”两种类型。直接评分是指评估者根据“忠实度”“流利度”等维度指标直接对译文进行打分,然后结合权重计算综合得分。而错误扣分是指评估者依据事先确定的错误类别、严重级别等对机器翻译译文中的错误进行扣分。在翻译行业中,MQM 多维质量指标体系(Multidimensional Quality Metrics)是目前应用最为广泛的错误扣分标准,它“将现有的各种资源进行整合,形成一个全面、开放、可定制的质量评估框架”(田朋, 2020: 24)。该体系将错误分为准确性、流利度、术语、区域惯例、风格、真实性、格式、国际化等八个维度,不同维度的错误又根据严重程度分为 minor(轻微错误)、major(重大错误)和 critical(严重错误)三个类别,最终的翻译质量得分 = 总分 - 译文扣分 + 原文扣分。总体而言,评分法的准确率最高,但是评估过程耗时费力,而且对评估者的能力要求比较高。

(2)排序法

排序法主要用于横向比较,对同一源语句子的不同机器译文进行宏观估计,然后按质量从高到低进行排序。如果有多个机器译文质量需要评估,且准确率要求比较高,可以选择使用评分法。但是,如果仅仅是想了解不同译文之间的相对好坏,就可以采用竞评或排序的方式,即对不同系统的每个句子根据译文质量进行排序。用户只需把不同机译系统产出的译文放入线上应用中,通过用户投票或排序结果来评估翻译质量。与评分法不同,排序法无需对译文质量给出精确的评价,而只是提供一个大概的质量高低排序,因而不仅效率高,评估结果的一致性也比较高。需要指出的是,为了获得较为准确的结果,这种方法通常需要较多的用户评估者参与。

1.2 有参考自动评估

有参考自动评估指以人工译文为依据,对机器翻译译文与参考译文之间的相似度进行自动计算,二者之间的相似度越接近,机器翻译的译文质量就越高。计算相似度的方法多种多样,根据评估方法对语言知识的依赖程度,可以分为“非语言”“轻语言”“重语言”三种。

(1)非语言的自动评估

非语言的自动评估不需要借助语言层面的分析来计算相似度,常见算法包括基于匹配率的方法(如 BLEU、SacreBLEU)、基于译后编辑距离的方法(如 TER)等。目前,业界使用

最广泛的自动评价指标是 BLEU(Bilingual Evaluation Understudy),最早由 IBM 公司于 2002 年提出。该方法的核心思想是利用 N-gram 匹配和惩罚因子对机器翻译和人工译文进行相似度计算,二者之间的相似度越高,译文质量也就越好。其中,N-gram 指 N 个连续单词组成的单元,N 值越大则表示评估时的匹配片段越大(Papineni et al. , 2002)。相比于 BLEU,SacreBLEU 采取了统一的标准进行自动分词,可以确保评估结果的一致性和可比性。

(2)轻语言的自动评估

轻语言的自动评估需要借助一定的语言信息进行计算统计,比如词性标注、同义词关联等,常见的算法有 METEOR、TER-Plus、MAXSIM 等。METEOR 算法的基本思想是基于词汇之间的语义相似度进行评价,同时利用 WordNet 等外部资源增加同义词的匹配几率。而 MAXSIM 则采用基于语言学的多种规则确定 N-Gram 的语法匹配,并且通过对每个匹配赋予一个权重,从而实现模糊匹配。

(3)重语言的自动评估

重语言的自动评估更强调从语法和语义层面对译文进行分析,通过对近义、阐释、句法结构、文本含义等语言方面的计算,考察待评价译文与参考译文之间的相似度。常见的算法包括 Asiya、ULC、DCU-LFG 等。以 Asiya 为例,该算法不仅考虑词汇之间的相似度,而且建立了一个涵盖词汇、语法、语义等多种语言学信息的数据集,从而提高计算模型的准确率。但是,由于需要考虑的因素太多,而且深层次自动语言分析的算法准确度还比较低,所以这种评估方法的代价通常比较大。

1.3 无参考自动预估

无参考自动预估从本质上来看是一种机器学习技术,它可以在不依赖参考译文的情况下对机译的产出质量进行预测(Speica, 2010),其结果可以快速地判断出机器翻译质量,对机器翻译性能的提升起着指导作用。

(1)基于人工抽取特征的质量预估

早期研究主要利用统计机器学习方法,将质量评估视为一个回归或分类问题,通过特征的抽取及选择实现对机器翻译的译文质量预估(吴焕钦 等, 2018)。其经典策略是在原文和带有得分标注的译文双语数据上,采用有监督学习方法去拟合人工质量打分或排序结果。要实现这一目标,首先需要一定的样例集合,然后基于样例集合提取翻译质量结果的有效特征作为分类或者回归模型的输入。最后,基于提取的特征,设计统计机器学习算法进行质量预测。例如,QuEst + + 基线模型(Specia, 2015)就属于基于人工抽取特征的质量预估方法,它包括特征提取模块和质量评估模块。前者借助于语言模型、IBM-1 翻译表、平行语料库等资源工具提取了 17 种质量特征,而后者基于高斯径向基函数的向量回归模型和网格搜索算法来寻找最优的超参数。其中,最核心和关键的内容是特征的抽取及选择。

(2)融合神经网络特征的质量预估

近年来,随着深度学习技术的引入,出现了不少基于神经网络模型的翻译质量自动预估方法(Kim et al., 2016)。其基本思路是先使用神经网络机器翻译模型生成一个向量来表示翻译质量,然后再使用该向量生成译文得分。利用词向量模型(如 word2vec、Fastex 等)可以将单词转化为一组稠密的向量表示,通过测量向量之间的距离来判断单词之间的语义相似性, 然后将词向量特征转换为句子的向量特征,可以进一步判断句子之间的语义和结构相似性(Mikolov, 2013)。例如,COMET 是一种基于预训练语言模型对机器翻译质量进行预测的方法。它利用神经网络模型,深入学习源语言和目标语言之间的对应关系,进而评估机器翻译的质量。前期研究结果表明,COMET 的预测结果与人工评估结果一致性较高,显示出其强大的预测能力(Rei et al., 2020)。

2　机译质量评估方法比较与适用性

2.1 机译质量评估方法比较

通过前文介绍可以看出,上述三种方法在性质上存在较大的差异,各有长处和不足。笔者基于前期相关文献的调研和分析,并结合翻译行业中的实践经验,从准确性、一致性、实时性、应用范围四个维度对三种方法的优势和不足进行归纳和总结(见表 1)。

表 1　不同机译翻译质量评估方法比较

评估方法	准确性	一致性	实时性	应用范围
人工评估	* * *	* *	*	* * *
有参考自动评估	* *	* * *	* *	*
无参考自动预估	*	* *	* * *	* * *

*　代表低, ** 代表中, *** 代表高

人工评估是目前反映机器翻译质量最可靠的一种评价方式。在评价译文时,评估者不仅可以识别发现译文中的显性错误,而且还可以根据经验判别那些形式对应但功能不对等的隐性错误。只要评估者的经验水平比较高,译文评估的结果一般也会比较准确。但是,这种评估方法的不足之处在于实时性较差,“需要耗费人力物力,而且评价的周期长,不能及时得到有效的反馈”(肖桐 等, 2020:32)。另外,理想的评估者是源语和目的语能力以及专业背景都很强的双语专家,但是在很多情况下,实际招募的评估者在双语能力方面可能参差不齐。由于个人偏好、翻译理念差异以及状态疲劳等因素,不可避免地会出现一定的主观偏差,导致评估结果一致性不高。

有参考自动评估具有速度快、一致性高的优点,但是在评估准确率和精细度方面不及人工评估。相较于人工评估,该方法的最大优势在于可以帮助用户快速对比多个机器译文

的质量。评估者只需提供对应的参考译文,系统就可以自动统计出译文得分。由于采取了统一的参考译文和统计方法,评估结果的一致性往往比较高。然而,有参考自动评估也存在明显的不足:这种方法严重依赖有限的参考译文,而现实生活的语言表达丰富多样,用户提供给系统的参考译文不可能覆盖所有可能正确的译文。而且,仅从匹配的角度考察译文质量的好坏很容易忽视语义内容和结构上的准确性,因而与人工评估结果的相关性不是很高(Freitag et al. , 2020)。此外,这种方法只能以数值形式提供一个机器译文的整体质量评价,无法像人工评估那样标注出译文中存在的具体问题。

无参考自动预估的优势在于实时性高,应用范围广,但不足也十分明显:依赖大规模人工标注数据,评估结果可靠性较低。在上述三种方法中,无参考自动预估的应用场景最为广泛,尤其是面对海量的译文时,人工评估就显得力不从心。而有参考自动评估方法需要一定的人工译文作为参照,因而在翻译行业中的应用也十分有限。无参考自动预估方法既不需要人工译文作为参照,也无须手动标注错误和统计得分,因而应用场景广、评估效率高,拥有巨大的发展潜力。尽管如此,在评估实践中应用该方法仍然面临一些现实挑战。比如,无论是机器学习还是训练神经网络都需要大量人工标注的数据集,而目前能满足大规模训练需求的数据集相对较少。此外,虽然自动预估的准确性有所提高,但是离实际应用需求还相去甚远。

2.2 机译翻译质量评估场景适用性

在语言服务行业,由于目的和需求不同,质量评估的应用场景也多种多样。任何一种评估方法都有其优势与不足,“不存在一种放之四海而皆准的标准或方法”(王均松, 2019:29)。在翻译质量评估实践中,应当针对特定的应用场景,灵活多样地选择评估方法(见表 2)。

表 2 不同场景下评估方法的选择

评估方法 / 应用场景	人工评估	有参考自动评估	无参考自动预估
机翻多引擎对比	√	√	
机翻引擎升级优化	√	√	
双语语料句对筛选	√		√
译后编辑内容筛选			√
翻译项目质量控制	√		√
翻译教学与译员培训	√	√	√

(1)机翻多引擎对比

在过去 20 年,随着机器翻译技术的迅速发展,国内外的机器翻译引擎如雨后春笋般涌

现出来,如 Google 翻译、DeepL、百度翻译、有道翻译、小牛翻译等。这些翻译引擎的质量评测受到广泛的关注和重视,对比不同机翻引擎的译文质量成为当前的一个研究热点(李奉栖, 2021)。从 2006 年开始,世界机器翻译大赛(WMT)就启动了机器翻译引擎的评测任务,评测方法主要采取自动评估和人工评估相结合的方法。有参考自动评估虽然准确率不及人工翻译,但评估效率和一致性高,因而在机翻引擎质量对比和评测方面备受青睐。而人工评价虽然需要耗费大量的人力和时间,但是准确率和可靠性高,二者互为补充,相辅相成。

(2)机翻引擎升级优化

为了提升机器翻译引擎的性能,机器翻译公司会持续不断地对其机翻引擎进行优化升级,定期推出新版本。在此之前,他们需要对机器翻译的产出质量进行评测,以确定新版本的机翻引擎质量较之前有所提升。由于系统优化对机译产出质量的分析要求更加细致和具体,通常需要采取可靠性比较高的人工评价方法。通过人工评估升级前后两个机器翻译引擎在不同层次(词汇、短语、句子等)的机译产出质量,尤其是前一版本中出现的典型错误是否得到纠正或完善,评估方可以判定新的机翻引擎能否上线。此外,部分机器翻译公司也会考虑使用有参考自动评估方法,以避免人工评估的主观偏差。

(3)双语语料句对筛选

机器翻译系统需要以大规模的双语语料库为基础,语料的质量会在很大程度上影响机器翻译的产出质量。因此,对双语语料进行筛选是一项非常重要的任务。由于语料库的规模通常比较大,常采取无参考自动预估和人工评估相结合的方法。首先通过自动质量评估筛选出有可能存在错误的语料句对,然后通过人工评估进一步确认,最后进行剔除或修改,从而为后续机器翻译引擎优化提供质量保障。人工评估所积累下来的质量评分或修订记录,可以为无参考自动预估补充训练语料,从而逐步提升自动评估的准确性。

(4)译后编辑内容筛选

随着机器翻译质量的不断提升,机器翻译译后编辑已经成为语言服务行业的新业态。据美国权威语言服务研究机构 CSA(Common Sense Advisory)的调查,从 2017 年到 2019 年全球提供译后编辑的语言服务企业从 29.86% 上升到 36.53%。这表明“机器翻译的译后编辑方式已经成为越来越多企业的翻译方式”(崔启亮, 2020: 29)。然而,并非所有的文本材料都适合进行译后编辑。如果机译产出的质量很差,那么译后编辑的效率就会大大降低。因此,在进行译后编辑前,通常需要通过评估对翻译材料进行筛选。鉴于待译材料通常数量多且没有参考译文,翻译公司会采取无参考自动预估方法。这样既可以完成筛选任务,还可以提升效率、降低成本。但是,采取这种方法的前提是必须有大量同类型文本语料作为训练数据集。如果文本语料不够充分,那么机器学习的效果就会大打折扣,评估结果

的准确性也就难以保障。

(5)翻译项目质量控制

在语言服务行业,翻译活动通常以项目的形式开展。其中,项目质量控制对于翻译项目的顺利进行至关重要,而译文质量评估又是项目质量控制的关键环节。一般情况下,翻译公司会首先使用一些 CAT 软件自带的质检工具修正一些低级错误,如拼写、单复数、术语一致性等。然而,这些软件工具只能起到辅助作用,要确保最终交付客户的译文质量达到要求,必须依赖人工评价。评估者可以依据特定的行业评估标准(如 LISA QA Model、SAE 等),对译文质量做出具体分析和整体评价。由于多数项目的翻译量都比较大,在具体实施过程中,人工评估通常以抽检的方式进行。另外,对于有一定历史数据积累的译后编辑项目,无参考自动预估也可以起到一定的辅助作用。项目经理可以根据自动评估结果,及时排查可能出现质量问题的环节,从而提升评估效率。

(6)翻译教学与译员培训

除了上述场景,机器翻译质量评估在翻译教学和译员培训中也发挥着非常重要的作用。客观准确的评估可以为译员培训提供及时、有效的反馈,从而提升教学效果。在教学实践过程中,教师和培训者可以根据不同情境灵活选择评估方法。比如,当学习者译文数量不多,而又需要在翻译策略和方法上进行指导时,教师应采取人工评价的方法,提供细致具体的反馈和修改意见。而当待批改译文数量较多且时间紧迫时,教师可以考虑采取有参考自动评估,根据一定数量的参考译文自动统计学生的译文得分。然而,如果学习者自主完成了大量的翻译实践,但是又缺乏相应的参考译文,教师则可以考虑采取无参考自动预估方法进行反馈和修改。

3 机器翻译质量评估研究展望

过去 10 年中,伴随着机器翻译技术的快速发展,机器翻译质量评估取得了长足的进步,展现出巨大的应用潜力。本文认为,未来机器翻译质量评估研究应着力在以下三方面加强:

(1)提高自动评估的准确率

首先,机器翻译质量评估的首要任务是提高自动评估的准确率与可靠性。目前,机器翻译质量自动评估方法的实时性、一致性较高,因而备受青睐。然而,自动评估方法,尤其是无参考译文的自动预估技术目前尚不成熟,准确率亟待提高。因此,未来研究的一个重点是改进机器学习算法、通过大规模预训练模型提高自动评估方法的准确率和可靠性。

(2)增加翻译质量评估维度

其次,机器翻译质量评估需要增加评估维度,拓展评估对象和评估内容。一方面,机译

评估的对象不能仅仅局限于译文,而应将原文也纳入其中,因为如果原文的质量较差,那么机器翻译很难产出质量较高的译文。另一方面,目前的自动评估方法主要关注语言形式上的相似度或语义内容的相关性。未来的研究重心应是如何将译文的形式特征和语言特征相结合,进行综合评估。为此,我们需要探索新的评估方法和技术,以更全面地评估机器翻译的质量。

(3)深耕垂直细分领域评估

最后,机器翻译质量评估的一个重要发展趋势是深耕垂直细分领域。目前,大多数机器翻译引擎主要面向通用领域进行开发,因此在应用于特定专业领域的翻译时,其质量往往无法满足需求。为了弥补这一缺陷,许多机器翻译公司已经开始致力于打造面向特定垂直细分领域的语料库和术语库,以提升机器翻译的产出质量。例如,百度机器翻译引擎已经提供了学术论文、生物医药、信息技术、金融财经等九个领域的在线翻译服务。未来,针对特定专业领域的机器翻译质量评估将成为一个极具潜力的发展方向,并呈现出蓬勃发展的趋势。

4　结语

本文围绕机器翻译质量评估这一主题,详细介绍了三种主流机译质量评估方法的优势、不足以及适用性。研究发现,人工评估、有参考自动评估、无参考自动预估三种方法各有优势与不足,在特定应用场景下应根据不同的目的和需求灵活选择评估方法,提高评估结果的准确性和可靠性。文章最后对机译质量评估研究前景作出展望,指出研究者应在提高自动评估的准确率、增加翻译质量评估维度、深耕垂直细分领域评估三方面深入挖掘,以期为后续研究提供借鉴和参考。

参考文献:

Freitag, M. ,D. Grangier & I. Caswell. 2020. BLEU Might be Guilty but References are not Innocent [G] // B. Webber, T. Cohn, Y. He & Y. Liu. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing*. Online Conference.

Graham, Y. , Q. Ma T. Baldwin, Q. Liu, C. P. Escartín & C. Scarton. 2017. Improving Evaluation of Document-Level Machine Translation Quality Estimation [G] //M. Lapata, P. Blunsom & A. Koller. *Proceedings of the* 15*th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia: ACL.

House, J. 1997. *Translation Quality Assessment*: *A Model Revisited* [M]. Tübingen: Gunter Narr Verlag.

House, J. 2015. *Translation Quality Assessment*: *Past and Present* [M]. New York: Routledge.

Kim, H. & J. H. Lee. 2016. A Recurrent Neural Networks Approach for Estimating the Quality of Machine Translation Output [G] // K. Knight, A. Nenkova & O. Rambow *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*. San Diego: ACL.

Mikolov, T. , Q. V. Le & I. Sutskever. 2013. *Exploiting Similarities Among Languages for Machine Translation* [EB/OL]. [2021-08-25]. http://arxiv. org/pdf/1309. 4168. pdf.

Papineni, K. , S. Roukos, T. Ward & W. J. Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation [G]

// P. Isabelle, E. Charniak et al. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: ACL.

Rei, R. ,C. Stewart, A. C. Farinha & A Lavie. 2020. COMET: A Neural Framework for MT Evaluation. arXiv preprint arXiv: 2009. 09025.

Reiss, K. (1971). 2000. *Translation Criticism: The Potentials & Limitations* [M]. Manchester: St. Jerome and American Bible Society.

Specia, L. , G. Paetzold & C. Scarton. 2015. Multi-Level Translation Quality Prediction with Quest + + [G] // Chen, H. H. and Markert, K. *Proceedings of ACL-IJCNLP* 2015 *System Demonstrations*. Beijing: ACL.

Specia, L. , K. Shah, J. G. De Souza & T. Cohn. 2013. QuEst-a Translation Quality Estimation Framework [G] // M. Butt & S. Hussain. *Proceedings of the Conference System Demonstrations of the* 51*st Annual Meeting of the Association for Computational Linguistics*. Sofia: ACL.

Turchi, M. , M. Negri & M. Federico. 2014. Data-Driven Annotation of Binary MT Quality Estimation Corpora Based on Human Post-Editions [J]. *Machine Translation* 28(3): 281-308.

Williams, M. 2004. *Translation Quality Assessment: An Argumentation-centered Approach* [M]. Ottawa: Ottawa University Press.

崔启亮. 2020. 人工智能在语言服务企业的应用研究 [J]. 外国语文(1): 26-32.

何三宁. 2015. 翻译质量评估模式研究 [M]. 北京: 中央编译出版社.

侯国金. 2005. 语用标记等效值 [J]. 中国翻译(5): 30-35.

李奉栖. 2021. 基于神经网络的在线机器翻译系统英汉互译质量对比研究 [J]. 上海翻译(4): 46-52.

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

陆金梁, 张家俊. 2020. 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) (2): 151-158.

王华树, 李智. 2019. 人工智能时代笔译员翻译技术应用调查——现状、发现与建议 [J]. 外语电化教学(6): 67-72.

王均松. 2019. 翻译质量评估新方向: DQF 动态质量评估框架 [J]. 中国科技翻译(3): 27-29.

王均松. 2023. 积极防范机器翻译的伦理风险 [N]. 中国社会科学报(003)2023-03-21.

王均松,肖维青,崔启亮. 2023. 人工智能时代技术驱动的翻译模式:嬗变、动因及启示 [J]. 上海翻译(4):14-19.

司显柱. 2007. 功能语言学与翻译研究——翻译质量评估模式建构 [M]. 北京: 北京大学出版社.

司显柱. 2016. 翻译质量评估模式再研究 [J]. 外语学刊(3): 84-94.

田朋. 2020. 翻译多维质量标准 MQM 模型介评与启示 [J]. 东方翻译(3): 23-30.

吴焕钦, 张红阳,李静梅,等. 2018. 基于伪数据的机器翻译质量估计模型的训练 [J]. 北京大学学报(自然科学版) (2): 279-285.

肖桐, 朱靖波. 2020. 机器翻译统计建模与深度学习方法 [EB/OL]. [2021-08-27]. https://opensource. niutrans. com/mtbook.

张霄军. 2007. 翻译质量量化评价研究综述 [J]. 外语研究(4): 80-84.

张霄军, 邵璐. 2021. 构建可信赖机器翻译系统的基本原则——一种基于工程伦理的观点 [J]. 外国语文(1): 1-8.

Machine Translation Quality Assessment: Methods, Applications, and Outlook

WANG Junsong ZUANG Congxi WEI Yongpeng

Abstract: The widespread adoption of machine translation post-editing underscores the importance of translation quality evaluation. This paper examines three primary evaluation methods (human evaluation, reference-based automatic evaluation, reference-free automatic estimation), delineating their strengths and weaknesses. It advocates adaptable selection of evaluation methods according to specific requirements. Future research should prioritize in enhancing the accuracy of automatic evaluation, broadening evaluation dimensions, and conducting domain-specific evaluations.

Key words: machine translation; translation quality evaluation; human evaluation; automatic evaluation

责任编校:龙丹