Computer Science and Application 计算机科学与应用, 2022, 12(10), 2275-2281
Published Online October 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1210232

Hans汉斯

基于高频错误类型分析的机器翻译质量标准、质量评估与发展态势

韦佑武[1]，李　娜[1*]，赵良威[2]

[1]东北电力大学，吉林 吉林
[2]中国能源建设集团天津电力建设有限公司，天津

收稿日期：2022年9月13日；录用日期：2022年10月12日；发布日期：2022年10月20日

摘　要

机器翻译作为跨文化交流、信息检索和不同语言之间迅速转换的工具，随着其使用的普及，机器翻译的质量评估问题逐渐引起广泛关注。本文在分析机器翻译发展过程中所体现的典型高频翻译错误类型基础上，综述和对比机器翻译质量评估的量化标准及评测原则，以期深入了解机器翻译质量评估在不同发展阶段的演进特征及未来发展态势。

关键词

机器翻译，发展态势，质量评估，质量标准

Quality Standards, Quality Assessment and Development Trend of Machine Translation Based on High Frequency Error Type Analysis

Youwu Wei[1], Na Li[1*], Liangwei Zhao[2]

[1]Northeast Electric Power University, Jilin Jilin
[2]China Energy Engineering Group Tianjin Electric Power Construction Co., Ltd., Tianjin

Received: Sep. 13th, 2022; accepted: Oct. 12th, 2022; published: Oct. 20th, 2022

Abstract

As a tool for cross-cultural communication, information retrieval and rapid conversion between

*通讯作者。

文章引用: 韦佑武, 李娜, 赵良威. 基于高频错误类型分析的机器翻译质量标准、质量评估与发展态势[J]. 计算机科学与应用, 2022, 12(10): 2275-2281. DOI: 10.12677/csa.2022.1210232

different languages, the issue of quality assessment of machine translation has gradually attracted widespread attention with the popularity of its use. Based on the analysis of typical high-frequency translation error types embodied in the development of machine translation, this paper reviews and compares the quantitative criteria and evaluation principles of machine translation quality assessment, with a view to gaining insight into the evolution characteristics and future development trend of machine translation quality assessment in different development stages.

Keywords

Machine Translation, Development Trend, Quality Assessment, Quality Standards

1. 引言

机器翻译是以计算机为平台通过特定的计算机程序，利用语言统计信息、世界知识图谱及相关翻译资源，将一种语言信息翻译成另外一种语言信息的计算机应用，最终把源语言转换为目标语言，其工作原理是模仿人工翻译所遵循的过程模式[1]。机器翻译质量评估是基于机器翻译系统源端句子和翻译结果，在不依赖参考译文的情况下对翻译结果的质量进行评估的机器翻译评测方法。随着机器翻译使用的普及化及翻译错误的规律性特征，针对机器翻译质量的评估研究逐渐向纵深化方向发展。

2. 机器翻译的国家标准与行业标准

机器翻译标准是衡量翻译译文质量好坏优劣的判断尺度，是衡量译文质量的尺度和翻译活动遵守法则。

2.1. 机器翻译国家标准

机器翻译的国家标准主要包括中国《翻译服务规范第 1 部分：笔译》(GB/T 19363.1-2008)、《翻译服务译文质量要求》(GB/T 19682-2005)、欧洲的《欧洲翻译服务提供商质量标准》(BS EN-15038 European Quality Standard for Translation Service Providers)、美国的《笔译质量标准指南》(ASTM F2575-06 Standard Guide for Quality Assurance in Translation)以及 2000 年欧盟、美国国家自然科学基金会和瑞士政府支持的 ISLE(International Standards for Language Engineering) ISLE/EALGLES 机器翻译评价体系。

其中，《中华人民共和国国家标准 GB/T 19682-2005 翻译服务译文质量要求》提出的“译文的忠实度、术语准确度和行文流畅度”[2]。包括以下条款：忠实原文，即：完整、准确地表达原文信息，无核心语义差错；术语统一，即：术语符合目标语言的行业、专业通用标准或习惯，并前后一致；行文通顺，即：符合目标语言文字规范和表达习惯，行文清晰易懂。归纳上述条款的要旨，可以得出以下三个属性来评价机译译文的质量：译文的忠实度、专业术语的准确度和行文的流畅度，也就是检验译文是否符合目标、语言的语法和表达习惯[3]。

除了以上标准，还有国内外译者翻译实践及理论研究的翻译标准，包括：严复提出的翻译标准“信、达、雅”，言简意赅，具有丰富内涵以及极大的包容性；鲁迅提出“易解、丰姿”双标准论；林语堂则提出“忠实、通顺、美”的翻译标准，主张对原文负责、对读者负责、对艺术负责；许渊冲提出“三美论”，即意美、音美和形美，他将文学翻译看作是两种语言和文化之间的竞赛，认为翻译的艺术贵在和

谐，文学翻译应该以“和谐”为审美标准；傅雷提出“神似”，主要包括以下四点：重神似不重形似；行文流畅、用字丰富、色彩变化；以艺术修养为根本；化为我有。国外翻译家尤金·奈达(Eugene A. Nida)基于功能对等理论(Functional Equivalence)，主张翻译时可忽略文字表面对应，而要在两种语言间达成功能上对等；彼得·纽马克(Peter Newmark)的交际翻译与语义翻译(Communicative & Semantic Translation)，认为语义翻译的重心是原文，语义翻译使译文与原文的形式更为接近，而交际翻译的重心在于译入语，注重译文读者的反应；美国著名翻译理论学家劳伦斯·韦努蒂(Lawrence Venuti)于 1995 年在《译者的隐身》中提出来异化与归化(Foreignization & Domestication)，即：归化是把源语本土化，以目标语或译文读者为归宿，采取目标语读者所习惯的表达方式来传达原文的内容，而异化在翻译上就是顺应外来文化的语言特点，吸纳外语表达方式。

以上标准的细节要求与参照，见表 1。

Table 1. Translation evaluation criteria
表 1. 译文评价标准

国家标准	GB/T 19682-2005	译文忠实、术语准确度、行文流畅	准确表达；术语符合标准；行文清晰
国内译者	严复	信、达、雅	忠实、通顺、选词得当
	林语堂	忠实、通顺、美	对原文、读者和艺术负责
	许渊冲	三美论	意美、音美和形美
	傅雷	神似	重神似不重形似
	鲁迅	信、顺	忠于原文、通顺流畅
国外译者	EugeneA. Nida	Functional Equivalence	在两种语言间达成功能对等
	Peter Newmark	Communicative & Semantic Translation	语义翻译及交际翻译
	Lawrence Venuti	Foreignization & Domestication	归化及异化

2.2. 机器翻译行业标准

与机器翻译的国家标准比较，行业标准相对数量较多，包括 LISA QA Model、SAE、MQM、TAUS DQF 等。其中，影响较大的标准 LISA QA Model、SAE J2450 和 MQM。LISA QA Model 是本地化行业标准协会(Localization Industry Standards Association, LISA)发布的，LISAQA 指标最初是为软件和硬件行业推广最佳翻译和本地化方法而设计的，但其标准化方法是语言服务行业应用最广的翻译质量评估模型。LISA QA Model 将译文质量分为语言与格式两大维度，其中语言维度分为错译(Mistranslation)、精确性(Accuracy)、术语(Terminology)、语言(Language)、风格(Style)、国家(Country)以及一致性(Consistency)七种错误类型，每种错误类型分别对应轻微(Minor)、重大(Major)、致命(Critical)三种严重程度。翻译服务企业在运用 LISA QA Model 时，通常会根据自身需要调整、细化错误分类与严重程度。因此，即使都是基于 LISA QA Model 的评价标准，仍会有不小差别。

SAE J2450 是由汽车工程学会(Society of Automobile Engineers)于 2002 年制定，目标是帮助汽车企业测试维修服务资料的翻译质量。SAE J2450 将译文错误分为七个类别：术语错误(Wrong Term)、句法错误(Syntactic Error)、漏译/多译(Omission/Addition)、构词错误(Word Structure & Agreement Error)、拼写错误(Misspelling)、标点符号错误(Punctuation Error)、其他错误(Miscellaneous Error)。设置严重(Serious)、轻微(Minor)两个严重级别。SAE J2450-201608 版本广泛应用于国外汽车行业和翻译服务公司，特别是在术语很重要的某些专业领域，如医疗、工业设备或制造业等。根据通用汽车公司的统计，通用汽车在采用 SAE

J2450 后的译文差错率降低了 90%，翻译交付时间提高 75%，总体翻译成本降低 80%。

MQM (Multidimensional Quality Metrics)多维度质量指标是欧盟资助的 QTLaunchPad 项目成果，由德国人工智能研究中心于 2014 年发布，最初目的是用于评价机器翻译质量。最新版本于 2015 年 12 月 30 日发布，MQM 主要是基于错误分类整合了各种翻译评价标准。表 2 为 MQM 错误架构[4]。

Table 2. MQM error structure
表 2. MQM 错误架构

Translation Quality	**Accuracy**	Mistranslation/Omission/Addition/Untranslated
	Fluency	Register/ Style/Inconsistency/Spelling/Typography/Grammar/Locale violation/Unintelligible/Monolingual Terminology/ Ambiguity/Character encoding/Nonallowed characters/Pattern problem/Sorting/Corpus conformance/ Link/cross-reference/Index/TOC
	Verity	Completeness/End-user suitability/ Legal requirements/ Local applicability
	Design	Overall design/Local formatting/Markup/Whitespace/ Graphics and tables/Truncation/ Text expansion/Length
	Internationalization/compatibility/other	

表 3 为经过简化的核心架构：

Table 3. MQM error structure simplified
表 3. MQM 错误架构简化版

Translation Quality	**Accuracy**	Mistranslation Omission Addition Untranslated
	Verity	Completeness Legal requirements Local applicability
	Fluency	Register Style Inconsistency Spelling Typography Grammar Local violation Unintelligible

MQM 提出了一般翻译质量评价标准没有涉及 Verity 维度，用于评价文本对应用环境的适宜性。可见，Verity 关注的问题超出了术语和语言的范畴，评估时不仅需要将译文与原文对比检查，保证译文忠实通顺，还需要考虑译文应用的环境。从翻译角度看，能够根据最终用途和读者调整译文，也是人类译者与机器翻译最大的区别。

另外，《翻译项目通用指南》(ISO/TS 11669 Translation Projects-General Guidance)中提出的 12 项条件选择适用的错误类型，包括：语言/地域(Language/Locale)、主题领域(Subject Field/Domain)、术语(Terminology)、文本类型(Text Type)、目标读者(Audience)、使用目的(Purpose)、语域(Register)、译入文本风格(Target Text Style)、内容对应性(Content Correspondence)、输出方式(Output Modality)、文档格式(File Format)以及生产技术(Production Technology)。

从以上质量标准可见，翻译质量评估测评的基本思路是基于对错误的多层级分类，通过统计错误评价译文质量。错误分类具有大类、小类多个层级，根据确定的错误类型对译文质量的影响，也相应设定明确的错误严重程度以及对应扣分或权重(weight)，常见的严重程度包括原文错误(Source)、偏好(Preferential)、轻微(Minor)、重大(Major)致命(Critical)。在译文检查后，根据错误数量、扣分和字数抽查计算译文得分或错误率，译文得分越高，证明错误越少严重程度越低。以上评价方法是分析式(analytic)的评价方法，其具体流程框架即：错误分类严重程度－权重个数－总扣分检查字数－得分或错误率。

3. 机器翻译的质量评估

翻译质量评价作为翻译活动的重要环节，是评测质量优劣程度，保证译文质量和提高译者翻译水平的重要途径和方法。机器翻译质量评估通常按操作方式方法分为人工评估和自动评估。对比机器翻译质量的人工评估与自动评估，人工评估的结果相对准确可靠，虽需较大的人工和时间投入；而自动评估方法，省时省力且相对稳定，被机器翻译领域广为采用，但严重依赖参考译文，所以可靠性较差，且与人工评估之间的相关性低。

依据评价方法，翻译质量评价分为质化评价和量化评价。翻译质量评价研究总体以质化评价研究为主[5]。

3.1. 机器翻译质量的自动评估

自动评估基于结果的量化对比，即：机译结果与标准人工参考译文或译后编辑结果之间的相似度对比，包括诊断性评价(Diagnostic Evaluation)、评分(Scoring)和排序(Ranking) 三种评价方式，常用方法包括 BLEU、NIST、GTM、METEOR、TER、HTER 等。其中，目前机器翻译领域常用的自动测度方法是 IBM 研究人员提出的 BLEU 测度，其原理是计算机器翻译译文与参考译文之间的距离，即：文本间 n 元相似度的平均(n = 1, 2, 3)。如果二者间的 2 元(即连续词对)或 3 元相似度较高，则该译文得分就更高。BLEU 的技术优势是评测简单快捷，可测算译文词汇的精确度和译文的连贯性进，能够比较机器翻译与标准译文间的分差，分值范围在 0 到 1 之间，并与人工评价高度相关。但由于 BLEU 仅关注词语搭配关系而忽略句子的整体结构，评估比较粗略，不适用于需要精确评估翻译文本质量的情况，因此在评估时也会用到一些改进方法，如基于词汇相似度进行评价的 METEOR、TER 等。

翻译自动化用户协会(TAUS)于 2011 年推出了动态质量评估框架(Dynamic Quality Framework，简称 DQF)。相比传统的静态翻译质量评估模式，动态质量评估框架提供了更具灵活性和适应性的评估方法，可根据评估者的需要构建能满足客户不同需求的质量评估体系和参数[6]。

自动评估还有以下发展趋势或特点：句法、语义等语言学知识的使用，语言学在质量评估方面提供了更加抽象的语言表现形式，使得评估更加灵活和准确；单个方法往往只能从有限的角度评估译文质量，所以倾向于从多个方面进行组合评价。

3.2. 机器翻译质量的人工评估

人工评估方法包括忠实度/流利度评分、等级排序、错误分类，如翻译自动化用户协会(The Translation Automation User Society, TAUS)的动态质量评估框架(DQF)及目前采用最广的翻译质量多维评估模型(Multidimensional Quality Metrics) [7]。

人工评测主要是依据不同测评体系，利用人工打分的定性方法和利用模糊数学的定量方法。孙逸群与周敏康构建 12 个指标、3 个评价层面与 5 个等级的指标体系，其中 12 个指标包括词义搭配、修辞、专门术语、方言使用、语法、衔接性、连贯性、意图性、可接受性、信息性、语境性和互文性，3 个层面分别为词汇、语法、和语篇，并通过问卷法获取数据，利用模糊数学理论进行定量分析[8]。徐雪惠曾利用错误分析理论对谷歌、百度以及科大讯飞三个主流机器翻译平台的译文进行评估，将机器翻译出的译文与正式译文对比并记录错误个数，但是该方法会因个人判断的误差而对结果产生影响[9]。程维库和梁洁曾提出一种质量综合评价方法，提出 12 种质量评价指标，采用问卷调查的方式获取机器翻译评估数据，后采用层次分析法确认评价指标的加权向量，然后利用模糊理论构建质量评价模型，该方法能够改善质量指标提取不精准的问题，但受到单一模型的限制，无法对深度学习等特征进行提取，质量评价准确性欠佳[10]。李奉栖采用错误记分法，从忠实度、流利度、术语翻译、风格、文化接受度 5 个维度对比研究人工译文与机器译文的英汉翻译质量，但也只是从宏观上进行较为模糊的定性评价[11]。梁伟玲和穆雷从定量角度对质量评估展开研究主要包括错误扣分法、标准参照模式、综合性评分法三种，并以相关研究为例，详细剖析了各评分方法的操作步骤[12]。总之，人工评测的结果虽然在统计学上有一定的意义，但评价过程需投入大量的精力和物力。

基于语言学语义对应检测点的方法，另有相关学者从词汇、句子、语义及语用四个层面进行质量评估。

一、词汇层面，词的使用频率和词语的搭配是词汇级层面的两个重要评估指标。在对词的评估上，通过对源语言文本、人工翻译文本以及机器翻译文本中词的使用频率的统计以及词语搭配的分析，以高频词的匹配度和词语搭配理论为评价标准考量机器翻译系统对词语翻译的忠实程度。二、句子层面，句子级的评估主要集中在平均句长和句法分析两个指标。在对句子的评估上，通过对源语言文本、人工翻译文本以及机器翻译文本中句子长度的计算以及句法结构的分析，以平均句长差距和句法生成理论为评价标准考量机器翻译系统对句子翻译的忠实程度。三、语义层面，语言意义上的评估主要表现为文本相似度和主题词的计算。在对语义的评估上，通过对人工翻译文本和机器翻译文本的语言意义的相似性和语言内容匹配性的计算，以语言内容的关联性、逻辑性和一致性为评价标准考量机器翻译系统对语言意义翻译的忠实程度。四、语用层面，语用级层面的评估主要表现为文本的情感分析，包括情感状态和情感倾向。在对语用功能的评估上，通过对人工翻译文本和机器翻译文本的情感概率的计算以及情感倾向性分析，以情感数值和语用效果为评价标准考量机器翻译系统对语言交际功能翻译的忠实程度[13]。

4. 结语

随着机器翻译技术的发展，机器翻译质量评估发展态势更加趋向于语言学知识的使用，以及由单方面评估向多方面评估转变，从而对机器翻译译文进行组合评价。本文仅在研究主题和研究内容方向进行了梳理与归纳，对于相应翻译案例及具体实例有待于在以后的研究中进一步提供和充实。

基金项目

本论文部分研究获得全国教育科学教育部青年项目(EIA160479)和吉林省高教研究课题(JGJX2021D116)的支持。

参考文献

[1] Poibeau, T. (2017) Machine Translation. The MIT Press, Boston.

[2] 国家质量监督检验检疫总局, 国家标准化管理委员会. GB/T 19682-2005. 翻译服务译文质量要求[S]. 北京: 中国标准出版社, 2005.

[3] 黄海英, 冯剑军. 英汉专业翻译软件翻译质量的人工测评[J]. 中国科技翻译, 2008, 21(1): 28-32.

[4] 熊志远. 2017-2022 年谷歌神经机器翻译系统英汉翻译质量对比研究[D]: [硕士学位论文]. 北京: 北京外国语大学, 2022.

[5] 王金铨, 于香, 吴万能. 基于词汇计量特征的翻译质量评价研究[J]. 中国翻译, 2021, 42(5): 113-120.

[6] 王均松. 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, 2019, 32(3): 27-29.

[7] O'Brien, S. (2012) Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialised Translation*, 55-77.

[8] 孙逸群, 周敏康. 机器翻译质量综合评价方法研究[J]. 中国科技翻译, 2017(2): 20-24.

[9] 徐雪惠. 神经网络机器翻译汉译英质量评价[D]: [硕士学位论文]. 北京: 北京外国语大学, 2018: 8.

[10] 程维库, 梁洁. 基于 Markov 网络的双语翻译译本质量评价方法[J]. 自动化与仪器仪表, 2022(6): 27-31.

[11] 李奉栖. 人工智能时代人机英汉翻译质量对比研究[J]. 外语界, 2022(4): 72-79.

[12] 梁伟玲, 穆雷. 应用翻译学研究的测试学视角——评介《翻译测试与评估研究》[J]. 外国语言文学, 2022, 39(4): 114-120+136.

[13] 王青, 马萧. 问题意识视域下的机器翻译质量评估方法研究[J]. 湖南社会科学, 2020(6): 144-151.