团体标准

《机器翻译服务质量评价规范—中英双向》编制说明

一、工作简况

1.任务来源

本文件由中国质量协会和华为技术有限公司联合提出，2022 年 11 月 10 日正式列入中国质量协会团体标准年度制修订项目计划。

2.编制目的

随着机器翻译（machine translation, MT）技术的发展，机器翻译已政府、教育、企业等实体的公文、教育和外贸领域广泛的应用，成为翻译领域的重要组成部分。

国际上 IBM、微软、谷歌等均在机器翻译上起步较早，特别是 IBM 首先提出了机器翻译质量测试规则 BLUE。国内主要有华为、腾讯、360、阿里、百度等龙头信息技术公司从事机器翻译的服务提供和研究。相应的，部分高校将机器翻译做为人工智能的一部分进行科研分类，如复旦大学、东北大学、哈工大等。但是机器翻译质量参差不齐，服务和交付的标准不一，如何以统一的标准进行科学、有效、高效的机器翻译质量评价，成为一个关键任务。目前机器翻译主要需求者是大篇幅翻译采购者，包括政府、公司等单位，相应的质量验收标准在合同中约定较为模糊，并由服务提供者自行编制，缺乏公信力。

因此，中国质量协会和华为技术公司期望通过整合行业的力量（包含领先的公司和专家学者）制定一个统一的机器翻译质量评估标准，用于机器翻译质量评价的方法和指标，从而提升评价的准确度、覆盖度和效率，在有效指导机器翻译质量改进的同时，提高行业的交付质量和标准。

翻译语向众多，不同语系之间存在较大差异，本文件聚焦使用最为广泛的中英双向机器翻译的质量评价，为其他语向机器翻译的质量评价提供有益借鉴。

3.机器翻译服务质量评价相关背景

《计算机科学技术名词》(第三版)将“机器翻译评价”定义为：人工或自动评价机器翻译系统译文质量的过程、技术和方法。质量评价是机器翻译研究必不可少的环节，无论是模型优化、上线、公司竞标等，都涉及机器翻译质量评价的

工作。

当前业界主流的评价方法分为自动评价和人工评价。自动评价方法，则运用特定算法和程序自动生成度量指标，对比机器翻译译文和参考译文，自动完成整个评价过程。自动评价的优点在于快速、高效、可复现。人工评价方法基于评价人员的专业能力，对机器翻译译文进行打分，准确反映出翻译的质量。因人是机器翻译的最终用户，所以人工评价更有说服力，可解释性更强。这两种评价方法，在 CCMT（中国最负盛名的机器翻译学术研讨组织，每年召开一次，http://sc.cipsc.org.cn/mt/conference）及 WMT（全球最负盛名的机器翻译学术研讨组织，每年召开一次，https://machinetranslate.org/wmt）竞赛活动、及企业对机器翻译质量自评估活动中广泛使用。

自动评价方法中，起草组选取了 BLEU[1]和 COMET[2]两个具体指标。BLEU 指标被 WMT 和 CCMT 采纳，COMET 被 WMT 采纳，用于评价每年参赛机器翻译模型的质量优劣。BLEU 是一种简单高效的统计评价方法，2002 年提出后已成为当前学术界、业界首选的自动评价方法。其论文至今被引用了 23000+次。COMET 是近年来基于神经网络技术的新评价指标，于 2020 年提出。其论文至今已被引用 400+次。COMET 算法更能衡量机器翻译译文与参考译文的语义相似度，与人工评价的相关性更高[3]。因自动评价方法对标人工翻译的参考译文，参考译文的优劣会影响指标的准确性。除此之外，测试集构成的合理性也会影响评价结果，起草组在标准内已说明测试集构建标准。

人工评价方法中我们选用了直接打分法，该方法简单高效，是 WMT 从 2016 年开始沿用至今的评测方法。评价人员的双语水平会影响打分的客观性。因此，起草组在本团体标准中，对评价人员的能力做了明确要求。除此之外，同自动评价，测试集构成的合理性也会影响评价结果的客观性。

综上，我们采用的评价指标与 WMT 设置一致，符合业界主流的机器翻译质量评价要求。

4. 主要编制过程

1）建立标准起草组

---

[1] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[2] Rei, Ricardo, et al. "COMET: A neural framework for MT evaluation." arXiv preprint arXiv:2009.09025 (2020).

[3] Konstantin Savenkov and Michel Lopez. 2022. The State of the Machine Translation 2022. In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track), pages 32–49, Orlando, USA. Association for Machine Translation in the Americas.

2022 年 11 月 10 日标准立项后，华为技术有限公司翻译中心和中国质量协会组织国内外机器翻译专家、人工翻译专家、大模型研究性院校和知名企业从业人员代表组成标准起草工作组。起草组组长刘群，华为诺亚方舟实验室主任，华为语音语义首席科学家，国内机器翻译开创人之一；起草组副组长江燕飞，华为翻译中心主任。

起草组制定了项目里程碑计划，分四个阶段完成。

2）形成标准草案

2022 年 12 月至 2023 年 7 月，标准起草组按照分工在华为内部机器翻译服务质量评价有关文件的基础上进行标准起草工作，在标准立项申报草案（华为内部机器翻译服务质量评价 SOP）的基础上形成各阶段标准 DIS 稿。起草组组织召开多次现场和在线讨论会，对相应技术内容描述、开放源归属、验证方法等进行讨论。

2023 年 6 月 16 日，在第三届华为机器翻译论坛期间，标准起草组进行了线下讨论和各机器翻译主要提供商的协商工作。

3）形成征求意见稿

2023 年 7 月，针对后续标准应用、测试集归属、防作弊等进行了线下讨论，形成公开征求意见稿。

5．主要起草人及所做工作

本系列标准起草单位：**中国质量协会数字化分会**，负责标准化技术要求、前言、引言部分；**华为技术有限公司**，负责范围、规范性引用文件、评价过程、质量评价和全文统稿和技术把关；**南京大学和东北大学（小牛翻译）**负责术语和定义；**腾讯技术有限公司和北京字节跳动科技有限公司**，负责评价方式；**小米技术有限公司**，负责评价过程。

在标准编制过程中，还有华为技术有限公司其他技术团队和中国中文信息学会定期举办的全国年度学术会议（CCMT）等专家参与意见。

二、编制原则和确定标准主要内容的依据

1.编制原则

按照 GB/T1.1-2020《标准化工作导则 第 1 部分:标准化文件的结构和起草规则》的要求和规定编写本文件内容。

遵循标准的先进性，系统性、可行性原则。

2.确定标准主要内容依据

本文件参考 GB/T 19363.1—2022《翻译服务　第 1 部分：服务要求》、GB/T 19682—2005《翻译服务译文质量要求》、GB/T 19000—2016《质量管理体系 基础和术语》（idt ISO 9000：2015）标准。

参考各起草单位多年在大语言模型，文本生成的约束和推理，机器翻译服务应用方面获得的能力验证做法、行业机器翻译评价的研究成果，及在业界实际应用情况，本团体标准提出人工评价和自动评价相结合，既采用最为可靠的人工评价，又通过工具快速计算出机器翻译与参考译文的相似度，同时度量语义相似性，实现科学、而准确的机器翻译的质量评价。

三、主要试验、验证分析

本文件基于评价华为内部机器翻译服务质量评价作业指导书和其他起草单位服务提供需求收集及能力验证，结合全国机器翻译大会（CCMT）以及相应国际和国内机器翻译比赛有关裁判要求，以及我国中英文机器翻译商务服务过程中实际可行做法和经验，确保本文件规范性、科学性、适用性及先进性。

本团标已设置自动评价定量指标（详见团标 7.1.1），以及人工评价定量指标（详见 7.1.2）。自动评价准入基线值取行业商用引擎的 80 分位，人工评价基线取行业惯例 80 分。

我们采用的自动评价开发集来自 WMT 2022 年的公开测试集，网络可获取，内容新，质量高，认可度高，2023 年发表的众多论文[4567]都采用了该测试集衡量机器翻译质量。而自动评价方法 BLEU 和 COMET,均已在 Github 上开源，可直接下载工具进行评价，因此自动评价方法可复制性和可行性很高。BLEU 和 COMET 评价方法的合理性在各起草单位自身工作质量评价（内测）和专业比赛中进行了验证，也可参见论文[8]和论文[9]。人工评价方法采用简单的直接打分制，我们在标准中也给出了每个分数段的错误描述和评价人员的能力要求，符合要求的评价人员根据标准快速上手。经过 WMT 2016-2022 年的实践，基于成本、效果等多方面考

---

[4] Raunak, Vikas, et al. "Leveraging GPT-4 for Automatic Translation Post-Editing." arXiv preprint arXiv:2305.14878 (2023).

[5] Raunak, Vikas, et al. "Do GPTs Produce Less Literal Translations?." arXiv preprint arXiv:2305.16806 (2023).

[6] Lo, Chi-Kiu, and Rebecca Knowles. "Data Sampling and (In) stability in Machine Translation Evaluation." Findings of the Association for Computational Linguistics: ACL 2023. 2023.

[7] Hendy, Amr, et al. "How good are gpt models at machine translation? a comprehensive evaluation." arXiv preprint arXiv:2302.09210 (2023).

[8] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[9] Rei, Ricardo, et al. "COMET: A neural framework for MT evaluation." arXiv preprint arXiv:2009.09025 (2020).

虑，直接打分法是现在行业上最佳的人工评测方法。

评价成本、投入方面，主要涉及测试集的构建、人工评价所需人力。其中测试集的构建成本主要包含数据抽取及人工翻译出参考译文，测试集构建完成后，不公开，可在同一领域内多次复用，进行某领域机器翻译服务质量评价；人工评价需要专业人员对所涉及机器翻译系统进行打分，每次人工评价活动均产生新的人员投入成本。具体成本可参考行业人工翻译、审校成本。

综上，团体标准有关技术内容和指标设定符合机器翻译目前国内（中英文互译）的通用技术水平，不包含特定服务提供商的特色服务指标。

四、采用国际标准和国外先进标准的程度

本系列标准为自主制定，不涉及国际和国外标准采标情况。

国际和国内在机器翻译服务质量评价方面尚无相关标准。采用“机器翻译”为关键字检索，国家标准 GB/T 40036-2021 《翻译服务 机器翻译结果的译后编辑 要求》等同采纳国际标准，内容为对文本的技术要求，也未见相应的质量评价要求。采用“机器翻译”、“翻译质量”为关键字检索有关论文，“机器翻译质量”论文 4 篇，“翻译质量”论文 5 篇，从名称分析，与本标准阐述内容不重复。

五、与现行有关法律、法规和强制性标准的关系

本系列标准符合国家现行法律、法规、规章要求，不涉及有关强制性标准。

六、重大分歧意见的处理经过和依据

本系列标准在制定过程中未出现重大分歧意见。

七、对公开征求意见的处理经过和依据

待公开征求意见后处理。

八、作为强制性标准或推荐性标准的建议

本系列标准建议作为推荐性标准发布实施。

九、贯彻标准的要求和措施建议

本系列标准的起草具备了良好的理论基础和实践验证，建议作为中国质量协会团体标准予以发布、实施，同时转化为认证依据，为第三方质量评估提供基本技术和流程的一致性要求。

十、废止现行有关标准的建议

本标准不涉及对现行标准的废止。

十一、其他应予说明的情况

无。