电 信 终 端 产 业 协 会 标 准

TAF-WG2-AS0059-V1.0.0:2020

面向智能终端的智能翻译测试库构建方法

The Construction Method of Intelligent Translation for Smart Mobile

2020－06－17 发布　　2020－06－17 实施

电信终端产业协会　发布

目　　次

前　　言

本标准按照 GB/T 1.1-2009给出的规则编写。

本标准由电信终端产业协会提出并归口。

本标准起草单位：中国信息通信研究院，中译语通科技股份有限公司，维沃移动通信有限公司

本标准主要起草人： 曾晨曦，林瑞杰，李欣杰，施艳蕊，宋佳明，苏兆飞，马雱阳，马蓁蓁，吴寒冰，高立发

面向智能终端的智能翻译测试库构建方法

1 范围

本标准建立的测试库适用于针对智能翻译系统进行的自动评测和人工评测。本标准对测试库建立时原文和译文的来源以及质量等进行了要求，并提出了相应的评估方法。测评时可根据产品的特性和测试需求从本测试库中选择测试集，测试集可以是本测试库的子集。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本适用于本文件。

3 原文要求

3.1 来源

在原文的选取原则上应保证原文语法正确、语序合理、用词恰当。基于以上原因，数据可来源于知名的开放媒体、会议或期刊、政府或者国家机构网站等。

原文需经过人工二次确认保证原文的准确性方可作为自动评测原文。

3.2 内容

评测数据集应保证不含粗俗以及政治敏感内容，价值观正确，所有参与评测的数据应使用正确的语法、合理的语序、恰当的用词等，不应存在错误信息。

评测数据应具有普遍适用性及代表性。对于通用领域评测数据集，所选评测数据应包含众多领域，可侧重在社会、政治、经济、口语、科技、体育、教育和医疗方面，领域应至少涵盖以上八个领域中的五个，所选的题材、句型、词汇应为这些领域常见而非生僻的，选择有代表性的真实文本。对于特定专业领域评测数据集，所选的题材、句型、词汇应体现出不同专业领域的普遍特点，选择有代表性的真实文本。

评测数据应涵盖特定语言不同于其他语言的词、短语、句法等数据。

3.3 数据规模

自动评测中，可采用以下两种数据规模：

——条件允许的情况下，为了减少同义词、同词干等问题对评测结果造成的影响，推荐使用采用一比四的原文译文句数比作为最终测试数据。例如使用1000句测试规模，应包含1000句原文及4000句对应译文。

——考虑到成本问题，也可使用一比一的原文和译文配比。数据集规模建议不少于1000句。

人工测评中，数据集规模不宜少于500句。

3.4 句子类型

不同智能翻译系统对不同长度的数据处理能力不同，为均衡评测数据本身带来的影响，在评测数据集制作过程中，遵循以下的规则：

其中，原文数据长度及对应长度数据占比如下表所示：

语料长度	百分比
15个字（词）以下	20%
15-30个字（词）	30%
30-75个字（词）	40%
75-120个字（词）	10%

3.5 数据格式

评测数据均使用UNIX纯文本格式进行存储，评测数据编码需为UTF-8无BOM 编码，若翻译系统中的中文为简体则中文评测数据需保证为中文简体。

4 原文质量评估

4.1 词语级

词汇级评测关注的重点是词语的含义以及词性的问题。比如评测点根据英文单词的词性划分为动词、名词、形容词、副词、代词、介词、连词、专有名词和新词、数字和倍数、冠词等10类。每一类中又将容易译错的内容细分为小类，如：因单词的多义、形态变化等引起的不同译法，并列出了部分典型测试点。

4.2 短语级

短语一级评测关注的重点是词语的搭配、成语、短语结构歧义等翻译中容易引起的问题。短语和搭配、修饰关系的测试点包括：名词短语、动词短语、成语和习语、并列结构、语法歧义结构等。

4.3 句子级

句子一级评测关注的重点是句法结构、从句和特定的句式的翻译。举例英文语料中测试点包括：否定形式、强调句、倒装句、省略句、there be结构、the...the从句、各种从句的翻译以及复杂长句的语义层次划分等评测点。

4.4 篇章级

段落和语篇级评测重点是句子之间的衔接和连贯、语义逻辑关系的合理性。

5 译文要求

5.1 来源

自动评测数据集的参考译文需由以目标语言为母语的翻译者或者精通目标语言的专业译员翻译得到。由于单条评测原文需对应四条参考译文，需保证四条参考译文由四名不同的译文翻译者翻译。同时在翻译过程中需在忠实度与流利度两个维度上保证译文的准确性与专业性。在忠实度方面，要求原文能够与参考译文内容完全对应，没有错译、漏译和多译等翻译错误现象产生；在流利度方面要求译文较地道、流利、无语法错误、无错别字、无错误标点符号使用等。

参考译文初步准备完成后，需进行二次人工检查，确认无重大问题方可作为最终参考译文加入到自动评测数据集中参与后续评测。

5.2 数据格式

评测数据均使用UNIX纯文本格式进行存储，评测数据编码需为UTF-8无BOM 编码，若翻译系统中的中文为简体则中文评测数据需保证为中文简体。

6 译文质量评估方法

参考译文初步准备完成后，应进行二次人工检查，确认无重大问题方可作为最终参考译文加入到自动评测数据集中参与后续评测。

附 录 A
（规范性附录）
标准修订历史

修订时间	修订后版本号	修订内容

附　录　B
（资料性附录）
附录

参 考 文 献