论机器翻译系统的评价体系

张政[1]，王贵明[2]
(1.北京工商大学 外语系，北京 100037；2.北京理工大学 外语学院，北京 100081)

摘 要：近年来,机器翻译的应用越来越广泛。本文主要对人们所关注的机器翻译系统的评价问题,就其类型与标准、系统评价的内容、以及系统评价的主要方法等方面进行较详细的介绍和评议。

关键词：机器翻译;评价体系
中图分类号：G64 **文献标识码**：A **文章编号**：1009-3370(2008)02-0112-06

与人工翻译译文质量评价不同的是，对机器翻译的评价除了评价译文质量外，还要评价机器翻译系统的其它性能。近年来,机器翻译的评测越来越受到广泛的重视。在过去几年中,国际上进行了若干次有影响的评测活动,如,信息理解评测(Message Understanding Conference,简称 MUC)评测专有名词识别问题,文本检索评估(Text-REtrieval Conference,简称 TREC)评测信息检索的发展,还有许多机器翻译和语音技术的评测活动，所有这些评测活动都对机器翻译的发展影响很大。

机器翻译适用的范围相对较小,通常是“相对限定的专业领域”,“不是用于文学性很强或文化味很浓的文本,而是用于科普文献、金融商业交易、行政管理备忘录、法律文件、说明书、农业及医学资料、工业专利、宣传册、报纸报道等”(Hutchins & Somers,1992:3)。Nagao 把机器翻译的适用范围限定在科技文献、文章题目、一般句子,而排除了诗歌、文学作品、法律文件、标书合同等(Nagao,1989)等。因此,人们对机器翻译进行评测时，限制所评测的文本的类型。尽管如此,在具体的机器质量评测上,也存在着不同的标准,它不仅仅只局限在纯粹的译文质量上,而且也涉及机器翻译系统的可操作性、机器翻译的可行性、机器翻译的类型评价、机器翻译的投入与产出比、机器翻译中的不同角色等等，鉴于这些复杂性,人们常常需要综合多个标准,以便全面、客观、公正地评价机器翻译系统。在这个意义上,正好与辜正坤的翻译多元系统互补论中通类基础标准系统（抽象标准)中的(2)(信的标准:忠实标准、准确标准、动态等值标准),和非类特殊基础标准系统(具体标准)中的(8)(科学技术著作翻译标准)标准不谋而合。阿诺德(D·Arnold) 等人曾建议机器翻译评价应考虑下述因素(Arnold et al., 1993)：

(1)机器翻译商品系统与机器翻译研究应该区分开来；

(2)从用户的角度评价机器翻译系统；

(3)从需求和结果相适应的角度看待机器翻译质量。

机器翻译评价的指标体系直接决定着机器翻译研发人员的研发路线和机器翻译的发展方向，评价标准对机器翻译的开发与研究都会产生重要影响,上个世纪 60 年代美国著名的 ALPAC 报告对 MT 作出的评价造成的影响很大（Arnold et al.,1993),所以,怎样客观、公正地评价一个机器翻译系统,这本身就构成了一个重大研究课题（Bourbeau,1993, Steiner, 1993 & Arnold et al., 1993)。

一、机器翻译类型与评价标准

1995 年,布瓦泰(C·Boitet)首先区分了机器翻译的四种类型：

(1)用于浏览者(for the watcher),称之为 MT-W，旨在为读者查阅外文资料时提供帮助,这种情况下,他们宁愿接受那种“粗糙”的译文(有时经过一定程度的后编辑),也不愿意一无所获；

(2)用于修订者(for the reviser),称之为 MT-R,旨在自动生成“粗糙”的译文,它类似于人工翻译的草稿，因而可以为专业翻译人员免去费时费力的工作,使他们变成修订者；

(3)用于翻译者(for the translator),称之为 MT-T,旨在协助翻译人员的工作,提供在线词典、同义词库、翻译实例库等；

(4)用于作者(for the author),称之为 MT-A,提供给希望作品被译成一种或几种文字的作者，该作者愿意在系统控制下写作或者帮助系统消除译本中可能产生的歧义。

收稿日期：2007-12-17
作者简介：张政(1958—),男,教授;王贵明(1960—),男,教授。E-mail:guimingwang@bit.edu.cn

同时,布瓦泰提出机器翻译成败的标准可以分为5种标准:概念标准、工程标准、实用标准、商业标准以及交际标准:

(1)概念标准:是否提出令人感兴趣的新概念,并借助模拟或实验原型展示其可行性和优越性,它主要与研究人员有关;

(2)工程标准:是否完成技术革新,或运用更先进的编程技术来建立原形或系统,它主要与系统开发人员有关;

(3)实用标准:是否经济合理地在实用条件下运行实验原形或完备的系统并取得满意的效果,它主要涉及用户;

(4)商品标准:从经济角度来评判,而不是仅仅看安装或销售的数量;

(5)交际标准:决策者或该领域的公司是否树立了良好的形象。

以上标准都是外围的评价标准,或者说是辅助性标准。一个系统输出的译文质量对系统的评价起着关键作用,所以,译文质量的标准制定就显得非常重要。目前世界上普遍采用ALPAC报告对译文质量的评价标准,分为可懂度(Intelligibility)和忠实度(fidelity)。

ALPAC可懂度等级(Scale of Intelligibility)(ALPAC,1966:69)分为9级:

9级:完全清楚、易懂,就像读平常的文本,风格贴切;

8级:完全或几乎完全清楚、易懂,但有小的语法错误或小的风格缺陷,和/或者小的用法不当,但容易"更正";

7级:总体还算清楚、易懂,但在风格、选词和/或者句法安排上不如8级;

6级:中心大意几乎一看就懂,但是由于选词差、风格不当、非习惯的表达方法、未翻译词以及错误的语法影响全面理解,译后编辑才能使译文得到认可;

5级:只有在仔细斟酌之后才能看懂大意,用词不当、古怪的句法、未翻译词、出现上述类似的错误,但尽管有这些"噪音"干扰,可以察觉出中心大意;

4级:句子变了形,实际上,与其说是可懂,还不如说是不可懂。但是,意思可以隐隐约约地领会理解,选词、句法安排和/或者表达古怪,关键词没有译出来;

3级:总的说来不可懂,读起来不全像胡言乱语,但是,仔细推敲和思索、可以假定句子的意思;

2级:读起来不全像胡言乱语,经过大量思索和推敲也无法让人读懂;

1级:根本无法读懂,怎么思索和推敲也无法读懂句子。

遗憾的是,原始的报告中并没有附上忠实度的各项等级。可懂度和忠实度这两个指标能如实反映译文的质量。从概念上来说,这两个指标相互独立,一个译文可以清楚易懂,但缺少忠实,譬如林舒的译文和法国翻译家让-弗朗索瓦·迪西(Jean-Francois Ducis,1733-1816)(谭载喜,1991:125)的译文,"美丽而不忠实",但另一方面,也可以是"非常准确、忠实,然而可懂性差",或者说是"忠实而不美丽",比如鲁迅的"宁信而不顺"。出现后一种情况,ALPAC认为是原文的可懂度差[APLCA,1966:67]。ALPAC同时给出了信息度(informativeness)的指标,共分为10个等级。信息度的评估,当时由18名以英语为母语、科技俄语阅读能力很强的人来进行评判。具体方法是,让评判的人先读译文,然后读俄语原文句子,再决定读俄语原文句子时是否获得了新的信息。新信息最多时为9,完全没有新信息时为1,新信息增加越少,说明翻译得越准确,译文质量越高。

日本科学技术厅的机器翻译译文评估共分可懂度(共5级)和忠实度(共7级)两个标准,并对可懂度和忠实度的程度作了分级。

可懂度等级(Scale of Intelligibility):

(1)文章意义明确,没有异议,用词、语法、文体都贴切,无需修改;

(2)文章意义明确,可以理解,但用词、语法、文体上多少有些毛病,不过,这些毛病很容易更正;

(3)文章的意义在总体上可以理解。但由于用词、语法方面的原因,有些细节的理解有疑问,读者不能完全靠自己更正,而想询问懂原文的人;

(4)译文的质量差,用词、语法的问题较多,经过反复、仔细的思考之后,能够在某种程度上猜出原文的意思,让人来修改还不如人工重新翻译;

(5)译文完全不可理解,必须由人重新翻译。

忠实度等级(scale of fidelity):

(1)译文忠实地反映了原文的内容;

(2)译文忠实地反映了原文的内容,文章的意义也容易理解,只需进行简单的修改;

(3)基本上能忠实地反映了原文的内容,但须进行调整词序等类似的修改;

(4)原文的内容基本上忠实地译出来了,但是,短语与短语之间的关系,过去时、完成时等时态、单复数的区别、副词的位置等有错误,需要译后编辑在结构上作必要的调整;

(5)原文的内容、结构都没有很好地反映出来,而且有一部分漏译,短语、从句的搭配有错误;

(6)原文的结构、内容没有很好地译出来,短语、句子有漏译,但大体上还看得出是一个句子;

(7)译文完全不能反映原文的内容和结构,因为脱落了主语或谓语,不成句子。

欧洲共同体 EUROTRA 采用了另一套评估标准:

(1)识别方面的标准:(a)可懂读,(b)忠实度,(c)连贯性,(d)有用性,(e)读懂速度,(f)可接受性;

(2)经济方面的标准:(a)输入时间,(b)编辑修正时间,(c)誊清时间;

(3)语言方面的标准:(a)句子结构和语义的连贯性,(b)词汇评价,(c)翻译错误;

(4)系统使用的难易度。

IBM 的 BLEU(BiLingual Evaluation Understudy)评测方法认为如果翻译系统的译文越接近人工翻译,那么它的翻译质量就越高。所以,评测关键在于如何定义系统译文与参考译文之间的相似度。BLEU 采用的方式是比较并统计共现的 N 元词的个数,即统计同时出现在系统译文和参考译文中的 N 元词的个数,最后把匹配到的 N 元词的数目除予系统译文的单词数目,得到评测结果。BLEU 方法简单易行,但是没有考虑到翻译的召回率(recall,指识别出来的某种类型的未登录词的数目和文本中属于该类型的未登录词的总数之比)。

国内对机器翻译评测的研究主要由北京大学计算语言所主持进行。该所在 90 年代后研究并开发了机器翻译评测系统。该系统使用分类评估法,并建立了机器翻译测试大纲。国家 863 计划也在不定期组织专家评测,对不同时期的汉英、英汉翻译系统进行了现场评测,评测结果反映了我国当时机器翻译的发展水平。

由上述各类标准可知,机器翻译系统的评价都是以译文质量作为评价的核心,而译文评价的标准也都最后落在了忠实度、可懂度上。

二、机器翻译系统评价的内容

机器翻译译文质量是机器翻译系统研究面临的核心问题,但还有其他很多重要因素需要考虑,有些因素是系统内在的特性,不能用现行的"黑箱(black-box)"加以测定,目前,大家公认机器翻译系统评价中涉及的主要因素有以下 7 种(Arnold et al.,1993)。

(1)机器翻译译文的质量

译文质量是机器翻译评价最重要、最核心、最关键的指标。但译文质量很难量化,其评价对人来说仍然是一项十分棘手的任务。两种语言的等价性是一个模糊的概念。同一原语实际上可以有数量不受限制的不同的目标语译文,更何况原语本身也充满歧义。如 Time flies like an arrow 就有三种解释:(1)时间像箭一样飞;(2)像箭一样测量苍蝇的速度;(3)时间苍蝇喜欢箭。此外,译文中必然会有各种语病,计算机分析病句的技术还不成熟。要进行这样的分析,必然会涉及系统内的核心内容,这方面的资料难以获得。就译文本身的质量而言,评价的标准也不尽相同。目前 ALPAC 报告中提出的"可理解性(comprehensibility)"和日本机器翻译系统中采用的"忠实度(fidelity)"影响最广,详见上一节。

(2)应用效率

机器翻译能否提高工作效率是用户,特别是专业翻译公司所关注的另一个重要指标。但是进行这种评价时应将机器翻译置于语言文字信息处理的全过程,如检索(retrieval)、识别(recognition)、输入、前编辑(pre-editing)、翻译、后编辑(post-editing)、输出、排版(type-setting)、印刷、远程通讯(distance communication)等,检验机器代替人进行翻译是否节省了大量的时间。

(3)工作方式

安装和使用机器翻译系统必然造成工作方式的改变,特别是译者工作方式的改变,他不必进行草稿翻译,而大部分时间用于对译文的修改。由于这些机器翻译自动生成的译文必然含有大量不同于人工翻译的错误,就要求译者掌握与以往不同的修改技巧。另一方面,机器翻译系统安装时,译者会接受不同的培训,但培训时间越短越好,系统越"友好"越好。因此,工作方式也是机器翻译评价中要考虑的因素之一。

(4)实用环境

使用环境包括机器翻译系统对硬件的要求,对其它软件的依赖,对输入文本的要求,用户界面的质量以及兼容性等等。人机界面是决定机器翻译系统的一个主要因素,一般情况下,用户愿选择译文质量稍差但便于译文修改的系统,而不愿选择译文质量稍好但不便于修改的界面。

人机界面分面向用户与面向开发维护人员(包括语言学家)两种。面向用户的界面应具有方便的前编辑、后编辑、辞典扩充等功能。面向开发维护人员的界面应能方便地修改词典、规则、语言模型,并能提供词频、句型、错误类型等统计数据。

(5)维护性和扩展性

维护性是指一个机器翻译系统能否方便解决实际应用中出现的问题,或者弥补系统的不足。而扩展性涉及该系统能否容易地扩展它的词汇或语言结构覆盖范围,它包括系统的扩展能否在用户工作现场

完成、完成扩展所需的语言学知识以及说明文档是否清楚地揭示用户应该怎样做，有无词汇添加的用户接口、性能情况,是否可以引进其他语言资源等。

(6)机器翻译系统性能价格比

机器翻译系统运行的速度、前后编辑时间、容量、外部配置等诸多因素,以求获得最佳的性能价格比。

(7)健壮性(即 robustness,音译为鲁棒性)

健壮性,指系统在处理系统认为是非法的(包括处理范围以外的)输入时的性能。鲁棒性高的系统,遇到处理范围以外的内容不容易崩溃，但鲁棒性与质量存在一定反比关系。

除了这七个因素外，阿诺德还讨论了其它几个相对抽象的深层系统特性,这些因素“可能无法让用户直接判断,但却相当重要,在有关的评价论述中容易被忽视。列举如下:

(1)模块性

系统的划分应与所涉及任务的逻辑和经验性的特点相统一,而且各个部分之间的接口应该清晰,例如词法分析要与转换分开,数据与算法混合等。模块好的系统容易维护、便于扩展。

(2)陈述性

数据和算法划分明确，算法处理的数据应以独立的解释而存在,即不依赖算法数据也可以被理解,因而可以预测系统的变更或更新的效果。

(3)单调性

系统升级后的每一个优点都不会使系统“退步”,即独立的升级可以成功地结合在一起,彼此间能够避免冲突,“和谐共处”。

(4)概念支持

一个系统能否实现了一些较清楚的理论原则,尤其是关于语言结构的一些理论。

由于计算机技术性能大幅度的提高和计算机技术的成熟,系统之间的运算速度、实用环境、维护性和可扩展性差别不大,机器的价格性能比大同小异,译文质量的比重显得更加重要，实际上很多机器系统的测评只测评译文质量，这也是近几年来机器翻译系统测评发展的新趋势。

三、机器翻译系统评价的主要方法

目前世界上机译评估异彩纷呈，就评估方法而言,机译评估大致可归纳为以下四类。

1.操作性评价

操作性评价(Operational Evaluation)也叫实用性评估,有时也称作经济评估(Economic Evaluation)。这种评估所关心的是机器翻译系统的经济价值,考虑机器翻译的成本，每个字的价格和译后编辑的开销,侧重机译与人译的花费以及时间比。基于这种思想,斯赖坡(G·V·Slype)于 1982 年提出了机器翻译系统的“效益评价”(Evaluation of effectiveness)的观点。这一方法也被人们称为实用性评价或经济性分析(Economic Analysis)。但该评价方法会受到以下几个方面的限制:(1)待测样本的限制;(2)译者水平训练;(3)质量控制;(4)词典更新。另外,操作性评价对一个用户来说代价过高,耗时较长;又很难使评价环境完全模拟实际情况，一般的用户都忽略这种评价方法。

2.说明性评价

说明性评估(Declarative Evaluation)又称质量评价(Qualitative Evaluation),这种评估又叫输出评价(Output Evaluation) [赵铁军,2000:361],也被称为是一种标准评价方法。这种方法不拘泥个别用户的特殊情况,而是力图给机器翻译系统更一般、更广泛的评价,它往往集中于译文质量上,侧重通过评测译文质量,评价机器翻译系统的性能。这一类评价的经典范例是卡洛尔(J·B·Carroll)在 1966 年进行的研究,其结果后来成为 ALPAC 报告的一部分。这种方法经济、实用、直观、快捷,还能测评译文质量的各个基本特性,得出的结果具有普遍性,与文本生成的过程无关。这种方法还可使第三方对各种机器翻译系统的质量进行比较测试，通过对机器翻译译文质量的评价,达到对机器翻译系统的评价。但这种评价只注重译文质量,容易忽视机器翻译系统的内在性质、不容易说明什么因素影响到译文质量，也不能说明某一个特定的语言现象、一个机器翻译系统的表现如何等。因此,使用这种方法时,至少有三个方面需要考虑:

(1)用于测试的语料:选材要慎重、公平,不同的语料对同一个机器翻译系统的评价影响很大;

(2)实验设计中的打分:打分的依据具有主观色彩因素,而且常常环境、时间、工作劳动强度的影响,因此，需要较多的评测人员对更多的样本进行评分以取得统计学上的意义;

(3)结果分析:结果只表明译文的质量情况,得到的数据难以深加工。

3.分类评估

分类评估法(Typological Evaluation)的核心是全面探究机器翻译系统的语言现象覆盖范围 [赵铁军,2000:363]。实现分类评估大致有两种途径:第一种途径类似于语言教学中的“错误分析法”,即根据错误多少为系统评分,有时也根据错误类型进行加权评分；第二种途径是预先制定覆盖面广的系统测

试集，测试集中每一个测试项目代表机译系统可能遇到或者它应该了解的语言现象，然后根据各机器翻译系统对测试集中句子的翻译情况记录下它生成的“好译文”、“差译文”或是“未翻译”的现象，予以评分。

机器翻译系统的评估，一般由人来进行，难免会带有测试者个人的主观因素，造成评估数据不准确，因此，一些机器翻译研究人员使用计算机来进行评估，建立机器翻译的计算机评估系统。乔姆斯基认为语言是有限规则的无限应用，英语原文的句子是无限的，汉语译文的句子也是无限的，为了测试英汉机器翻译系统的质量，可以将测试点约束在有限的范围内，并将测试点划分若干等级，当机器翻译系统通过了某个登记的测试点的集合，则认为该系统达到了相应的水平。北京大学设计的MTE系统还设计了一种测试文本描述语言（Text Description Language简称TDL）。利用TDL瞄准测试点，可以描述一个或若干个英语句子所得到的各种水平的答案及其所属科目的得分。TDL解释程序的工作就是将这种预先准备好的各种答案同机器翻译的译文进行比较，从而决定译文分值。

这种方法对研制者非常有用，不仅能评估系统的译文质量，而且能发现机译系统对哪些语言点处理不好，提供系统的弱点所在，而且反复应用同一测试集可以比较出后继的改进是否奏效，测出系统的改进程度，还可以针对机器翻译过程的各个阶段来进行评估，即针对形态分析、结构分析、词汇转换、结构转换、结构生成、形态生成等不同阶段，准备不同的测试数据，从而判断错误是在哪一个阶段造成的，为规则的调整、辞典的补充、算法的改进提出可靠的数据。

因此，这种评估方法简便易行、节省人力、物力，但这种测试系统的开发前期投入大、要求测试点的选择合理、科学、全面。

4.自动评价和人工评价

自动评价和人工评价(Automatic Evaluation and Manual Evaluation)是根据评价方法来划分的。所谓自动评价就是利用计算机的评价译文。汤普森(H·Thompson)系统是目前世界上能够实现评价与评分过程全部自动化的一种，该系统以段落为评价单位，而北京大学计算语言所研制的MTE评估系统则以句子为单位，IBM的BLEU采用的自动评测是以词为单位，这些评估系统对机器翻译系统的质量能做出比较客观的评价。所谓人工评价就是利用人工根据制定的评分标准对机器翻译系统输出的译文进行评分。这种方法，简单适用、前期投入少，不足之处是主观因素大，同一个译文，不同的参评人员的评分也不同。ALPAC报告的评分就是采用这种方法，欧洲一些国家和日本大多采用这种方法。

综上所述，不难看出，机器翻译的质量完善以及对机器翻译系统评估的体系完善都是复杂而漫长的过程。如果说，现在机器翻译的结果还不尽人意，计算机辅助翻译的方式有待革命，那么我们应该从什么方面寻找突破口，使计算机翻译能够更真正智能化？当今的语言交流信息的量度和快捷度，要求一种适应全球交流速通的新技术，建构一种真正能适应快捷而且大容量语言信息交流的多文种数字化交互平台。也许我们可以从一种新的语码转换方式中找到明天的曙光。

参考文献：

[1] ALPAC. Language and machines: computers in translation and linguistics[R]. A report by the Automatic kj,.language Processing Advisory Committee. Division of Behavioral Science, National Research Council. Washington, D.C.: National Academy of Sciences - National Research Council, 1966.

[2] Arnold, D. & Louisa, S. Evaluation: An Assessment [A]. Arnold, D. & Lee, H. eds. Machine Translation: Special Issue on Evaluation of MT Systems[C], 8(1-2). Dordrecht: Kluwer Academic Publishers. 1993,1-24.

[3] Arnold, D. & Dave, M. Automatic Test Suite Generation[A]. Machine Translation: Special Issue on Evaluation of MT Systems[C], 8(1-2). Dordrecht: Kluwer Academic Publishers. 1993,29-38.

[4] Bar-Hillel, Y. The present status of automatic translation of languages[J]. Advances in Computers, 91-163. 1960(1).

[5] Hutchins, W. J. Machine Translation: Past, Present, and Future[M]. Chichester, England: Ellis Horwood Limited, 1986.

[6] Hutchins, W.J. & Somers, H. L. An Introduction to Machine Translation[M]. London: Academic Press, 1992. [Hutchins & Somers, 1992, p3]

[7] Nagao, M(trans by Cook, N. D.) Machine Translation, How Far Can It Go?[M]. Oxford: Oxford University Press, 1989.

[8] 辜正坤. 中西诗比较鉴赏与翻译理论[M]. 北京：清华大学出版社，2003.

[9] 李晓敏，等. 译文评价标准新探索[J]. 上海科技翻译，2003，(3).

[10] 谭载喜. 西方翻译简史[M]. 北京：商务印书馆，2000.

[11] 赵铁军. 机器翻译原理[M]. 哈尔滨：哈尔滨工业大学出版社，2002.

On the Evaluation Systems of Mechine Translation System

ZHANG Zheng[1], WANG Gui-ming[2]

(1.Beijing Industrial and Commercial University, Beijing 100037; 2. Beijing Institute of Technology, Beijing 100081)

Abstract: With the wider use of machine translation in recent years, more people are concerned about its evaluation system. This paper introduces, in detail, the evaluation systems of machine translation from the types and standards, the main items of the system, and the main modes of the systematic evaluation, and meanwhile it makes some comments on the relevant conceptions.

Key words: machine translation; evaluation system

[责任编辑:萧姚]

(上接第 86 页)

Loss of Cultural Implications for Relational Vocatives in the English Version of A Dream of Red Mansions

SI Dong-hong, CAI Zhong-yuan

(Department of Foreign Language, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: As there are so many characters and many complex relational vocatives which are best reflected by Chinese traditional culture in A Dream of Red Mansions, it is necessary to analyze the substantial differences between Chinese and English relational vocatives. This thesis examines the loss of cultural contents by comparing the English and Chinese relational vocatives, and attempts to find out some feasible remedies as from the English version of A Dream of Red Mansions by Yang Xianyi.

Key words: loss of cultural contents; a dream of red mansions; translating process; relational vocatives

[责任编辑:萧姚]