面向用户的 MT 系统评测方法

傅 爱 平

（中国社会科学院语言研究所）

摘 要：本文从用户的角度出发，分析了与 MT 系统评测标准相关的因素，认为它们有与系统的内部结构相关和无关之分，有与译文质量相关和与系统运行相关之分，还有用户的主观和客观之分。为了在评测标准中体现这些特点，本文提出了“双主体，多项目，客观描述”的原则，并根据这些原则给出了评测的标准和实施方法。

The User-Oriented Evaluation of MT Systems

Fu Aiping

(Institute of Linguistics, Chinese Academy of Social Sciences)

ABSTRACT: The paper first make a general survey of user-concerned factors relevant to the evaluation of MT systems. The factors are different in the following aspects: To depend on linguistic principles and algorithm of a MT system; To come from the processing enviorment of the system; To affect the quality of translation; To be relevant to the user's subjective wishes. In order to cover all the factors discussed, a principle of evaluation is presented, which is "Double-Subjects, Multi-Items and Objective Descriptions". The criteria and approach of the evaluation on the principle are also described.

评测一个实用型(或称商品化)机器翻译系统的目的是什么？从不同的角度出发会有不同的理解。系统的研究人员希望评测能够肯定他们的设计思想和处理技术，并且能为下一步的研究工作提供信息；系统的开发人员希望他们对系统所作的改进能够得到证实；而系统的用户关心的则是，这个系统能否有令人满意的翻译质量和工作效率。不同的目的会导致不同的评测方法，需要不同的评测标准。其中适用于用户的评测方法涉及的相关因素最多，因素间的相互关系也相当复杂，最不容易找到确切的标准。本文将讨论这种方法。

1. 与评测标准有关的因素

人们通常认为，评测一个 MT 系统的标准首先是译准率，然后是翻译所用的时间。有了这两个参数就可以刻划一个系统的基本状况了。实际上，从用户的角度考察一个实用型的 MT 系统，其情形要复杂得多。一方面我们会发现许多影响 MT 译文质量的因素，不是仅仅用一次测试得到的译准率就可以概括的；另一方面我们还会发现一些与 MT 系统运行有关的因素，翻译时间仅仅是其中之一。这两方面的因素都与评测的标准有关，它们主要是：

1.1 影响 MT 译文的因素

影响 MT 译文质量的因素有些来自系统内部，有些来自系统外部。

1.1.1 系统内部的因素

(1)系统的语言处理能力(Linguistic Coverage)

语言处理能力能够体现系统的语言学理论基础和算法技术，主要表现在各种句型的识别和生成，歧义的处理，词汇量等方面。它与 MT 译文的质量直接相关，是评价 MT 系统的最重要的依据。人们常常用特定的测试集评测一个系统的语言处理能力(俞士汶等，1992；Arnold et al. 1993)。在测试集的句子当中，分布着许多体现词汇、词法、句法、语义等问题的测试点，综合各个点的得分评价被测试的译文，就可以对系统处理这些问题的能力有一个比较全面的认识。几个国内的 MT 系统在研制阶段的后期都采用这种方法进行了测试。这种方法最受 MT 研究人员的欢迎，因为它能直接验证系统的设计思想、方法和技术，还能对改进系统的性能提供有价值的数据。

(2)系统的改进能力(Extensibility and Up-grading capability)

任何一个实用型的系统，当面对用户时，都会面临改进的问题。用户常常要根据特定的翻译需求调整他的系统。譬如扩充词典、增加术语、修改词的用法规则等等。笔者曾经作过一个统计，一个有一定翻译能力的 MT 系统，在翻译科技文献时，由于缺乏专业术语而译错的句子约占全篇总句数的四分之一到三分之一。换句话说，如果系统能够为用户提供增加术语的功能，那么译文的质量将会提高二十到三十个百分点。再考虑到每一个系统的用户对专业领域的要求是相对固定的，那么象增加术语这样的改进能力对翻译质量的作用就更不可低估了。

对于用户来说，系统的改进能力如何，可以从以下几个问题考虑：什么样的改进是用户可以作的？作这样的改进需要多少语言学知识或对系统的了解？系统提供的更新词典的手段是否简明、实用，以保证用户可以完成这些改进？是否可以控制或者追踪这些改进的项目？有无较好的用于改进系统的界面？等等。显然，这些问题很难用测试集的方法评定。

(3)系统的结构特性(Structural Characteristics)

系统的结构特性包括系统的模块化程度，陈述性(算法与数据分开)，单调性(规则与规则之间互不干扰)，还有系统的鲁棒性(Robustness)。这些特性很抽象，无法作定量分析，但是它们与译文质量间接有关。因为它们既制约着系统的语言处理能力，也制约着系统的改进能力。不言而喻，这些问题也很难用测试集的方法评定。不过用户在评价一个系统时，注意的往往是这些性能提供给他们的方便，而并不直接关心这些性能如何。

1.1.2 系统外部的因素

前面三点是系统的设计者可以把握的因素，称之为系统内部的因素。以下是系统的设计者无法把握，却又与译文质量密切相关的因素：

(1)翻译的目的(Purpose of translation)

用户使用 MT 系统的目的可以分成以下四种：①浏览译文，估计与主观需要的相关程度；②通过译文了解文献的内容；③出版译文(Jordan et al. 1993)；④人工翻译的辅助手段。

目的①和④对翻译质量的要求不高。目的②要求译文首先在语义上忠实于源文。目的③除了语义上的要求以外，还在语法、修辞、文体、风格等方面有所要求。不同的翻译目的要求不

同的翻译质量，就需要不同的评测标准。

⑵源文所属的领域(Application Domain)

一个长于翻译生活用语的系统不一定能译好科技文章。同样，用于科技文献翻译的 MT 系统在处理不同专业的源文材料时，译文质量也会有很大的差别。另外，源文的体裁和行文风格也会对译文有所影响。譬如，某些学术论文经常使用公式、符号，而在技术说明书中，符号、插图、无谓语句、祈使句出现较多。这些都会对分析源文的句法结构产生影响。可以肯定地说，同一个 MT 系统在翻译不同类型的源文时，会表现出不同的能力。

1.2 与 MT 系统运行有关的因素

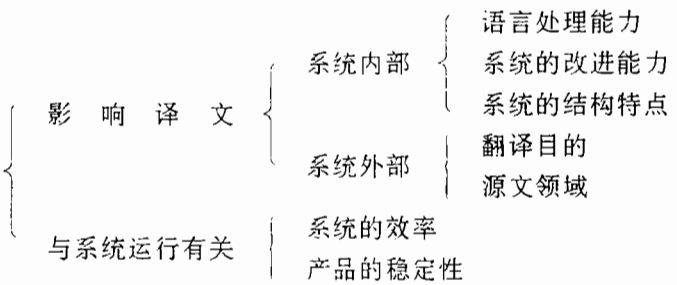
⑴系统的效率(System Efficiency)

翻译所用的时间是效率的重要参数。在某一硬件环境的限定下，可以测出翻译的速度。再加上设备、译前/译后编辑的费用，就可以计算出翻译的成本及系统的效率。在考虑 MT 系统的评测标准时，系统效率的重要程度仅次于系统的语言处理能力。显而易见，效率越低的 MT 系统越不容易被用户接受。由于效率与系统的实际运行密切相关，这个问题最好从用户的角度出发考虑，而不应仅仅由系统的研制人员确定。

⑵产品的稳定性(Product Stability)

从系统运行的角度看，MT 系统要成为一个稳定的产品，至少应该具备以下条件：①友好的用户界面、简明的帮助文件和实用的辅助工具，②用户培训和售后服务，③系统的维护和升级。产品的稳定性对用户来说十分重要，因此也应该成为与评测标准有关的因素。

综上所述，与评测标准有关的因素可以归纳如下：



2. 确定评测标准的原则

在评测 MT 系统时，如果把用户作为主体，那么可以认为我们在上一节归纳的与评测标准有关的七个因素都是客观因素。用户对系统的评价还要受其主观因素的制约。这些因素包括：用户使用翻译的源语言或目标语言的能力，对译文所属专业的熟悉程度，对 MT 的了解和期望值，对系统改进或译后编辑的方法掌握得如何，甚至还有用户对使用计算机工作的习惯程度。主观因素因人而异，同一个系统，同一篇译文，不同的人会有不同的评价。因此在评测 MT 系统时，若不考虑用户的主观因素，难免无的放矢；可要由 MT 的评测人员去考虑用户的主观因素，又太难为他们了。

关于 MT 系统的评测问题，近年来国外曾有数篇论文讨论。有人曾把评测的方法归为三类(Arnold et al. 1993)：

(1)操作法(Operational Evaluation)：从特定的系统、特定的用户、特定的应用环境出发，人工评价运行该系统的经济效益。这种方法能够照顾到用户的实际要求，也能反映出系统的语言处理能力、系统的效率等主要评测因素。但是用这种方法评测占用时间长，费用高。而且对某些相关因素处理不当会影响评测结果的可靠性，如测试的源文材料、对用户或译后编辑人员的训练、词典更新，等等。

(2)陈述法(Declarative Evaluation)：不针对某一特定的用户，采用可以广泛应用的标准(如正确性、精确性、智能性)。与(1)相比，是一种具有普遍性的方法。然而照顾到了普遍性，就往往忽略了特定的因素。譬如，测试材料的选择就很难作到公平(不同的系统适合不同的领域)。另外这种方法也没有体现出与系统运行有关的因素(如系统的效率)。

(3)类型法(Typological Evaluation)：使用测试集测算 MT 系统对各种语言现象的处理能力(详见 1.1.1)。得到的数据在研制阶段后期对系统的设计人员很有意义。但是用户常常会感到他实际应用系统时的情况与这样测试的结果有相当的差距。这是因为除了语言处理能力这个因素以外，用户关心的其他各种与评测标准有关的因素几乎都很难用这种方法反映出来。

在与 MT 评测标准相关的诸因素当中，有系统内部的，有系统外部的；有与译文质量有关的，有与系统运行有关的；还有用户的主观和客观之分。可以看出，上述三种方法都想尽量地把这些因素全面、准确地反映出来。所谓"全面"，即照顾到各个因素；所谓"准确"，即或给出统计数字，或给出图象图表。然而，这些因素有的可以定量分析，有的只能定性分析，有的对于研制人员来说，连定性分析都困难(譬如用户的主观因素)，同时这些因素又在相互作用，使人在评测的时候难免顾此失彼。

那么，针对相关因素的主、客观之分，能不能按照"各负其责"的原则，让系统的研制人员和用户分别在各自有发言权或关心的问题上扮演评测主体的角色。

其次，针对相关因素复杂的情况，能不能设立多个评测项目，使每个项目对应单个因素。

此外，对相关因素的评测，能采用定量分析的尽量采用定量分析。不能采用定量分析的，则采用定性分析。在定性分析当中，根据"各负其责"的原则，系统的研制人员以提供客观描述为主，给出定论为辅。

以上就是我们确定评测标准的三条原则：双主体，多项目，客观描述。

从这三个原则出发，对系统的语言能力可以以研制人员为主进行定量分析，对系统的改进能力可以以研制人员为主进行定性分析，翻译目的和源文领域可以用适应用户需求的多个子样品集反映，系统的效率和稳定性则应由研制人员提供相关数据，以用户为主给予评价。

3. 评测的方法

根据双主体，多项目和客观描述这三个原则，可用以下标准和方法评测一个 MT 系统：

(1)翻译基本测试集的成功率

基本测试集用来测试系统的语言处理能力，与前一节所叙述的类型法的思想类似。测试点的分布侧重于基本句型，没有特定的专业背景(例如北京大学的 MT 译文质量自动评估系统

的测试集)。

由几位评测人员(要求他们熟练掌握翻译的源语言和目标语言)对系统输出的译文进行分析,这种分析应依据某一预先制订的标准,譬如表1(Nagao et al. 1985)。对于表1,可以选择单一标准(规定1-4级是翻译成功的译文)记分,也可以用复合标准记分(给不同的级别加权)。综合全部译文的分析结果,就得到翻译的成功率(例如百分制下的一个分数)。可以认为这个分数是对系统处理基本语言现象的能力的比较客观的定量分析。

(2)翻译高级测试集的一次成功率和二次成功率

高级测试集仍以句子为单位,收集合乎规范的难句和长句,譬如具有相当复杂程度的复合、同等、省略等语言现象(一般程度的这类问题应归入基本测试集),没有特定的专业背景。

用计算基本测试集成功率的方法计算这个测试集的一次成功率。然后由研制人员针对第一次翻译不成功的句子对系统进行调整,同时记录调整所需的人力和时间。调整以后再第二次翻译整个测试集,用与前一次相同的方法计算得到二次成功率。一次成功率定量描述了系统处理复杂语言现象的能力,一次和二次成功率之间的差别以及调整所耗费的人/时反映了系统的改进能力。虽然用户看不到这种改进是如何具体实施的,但可以通过客观的描述得到一种定性的分析。这种分析还能使用户间接了解到系统的某些结构特性。

(3)翻译特定专业子样品集的一次成功率和二次成功率

对于一个翻译科技文献的MT系统,用户最关心的是它能否胜任本专业的翻译工作。因此系统的评测标准就不可避免地要体现专业的特点。专业文献的样品集是一篇或数篇在专业上有代表性的文章。为了体现源文体裁和行文风格的特点,还应该在同一专业的样品集中分设几个子集,分别收入科普文章或论文类、技术说明类、文摘类等各类文献材料。

象标准(2)那样,分别计算系统翻译某个子样品集的一次成功率和二次成功率,就可以得到一个定量的分析,说明系统翻译某个专业的某类文献的能力,以及一种定性的分析,客观地描述系统在翻译某个专业的某类文献时的改进能力。

需要说明的是,这里在计算二次成功率之前对系统进行的调整,与标准(2)有所不同。标准(2)的调整工作既可以把系统看成一个整体,只根据输出的译文表面提供的信息,确定修改的目标和策略(称为基于"黑箱"(Black Box Based)的调试),也可以深入系统的内部结构,追踪运行的情况,了解翻译的过程,然后确定修改的目标和策略(称为诊断(Diagnose 或 Glass Box Based))。而在这里,一般只允许进行基于"黑箱"的调试。因为这种调试主要是修改译文表面可以发现的短语的结构和语义问题,在科技文献的翻译中,则主要是增加专业术语。本条标准把系统的调整限制在术语处理的范围内,是基于笔者用MT系统调试大量科技文献后形成的两个观点。其一,对一个具有一定翻译能力的MT系统来说,处理科技文献失误的主要原因是缺乏特定专业的术语,因此用户能否方便地更新词典、增加术语,是评测的一个重要标准。其二,对用户来说,基于"黑箱"的方法是更新词典、增加术语的最方便的方法。在这样的限制下,用户通过一次成功率和二次成功率之间的差别,就可以清楚地看到,经过他们自己作得到的调整之后,系统有了多大程度的改进。

(4)与系统效率有关的参数和对系统稳定性的说明

系统的研制人员能够给出的参数主要是翻译速度:在某一特定硬件环境的限定下,单位时间内翻译的词的个数。关于系统稳定性的说明主要是对用户界面、帮助文件和辅助工具的使用说明,用户培训和售后服务的安排以及系统维护和升级的承诺。实际上在这方面最有发言

权的是用户，例如效率与他们实际使用系统的情况就有很密切的关系。

4.结论

我们分析了与MT系统评测标准相关的因素，认为它们有与系统的内部结构相关和无关之分，有与译文质量相关和与系统运行相关之分，还有用户的主观和客观之分。为了在评测标准中体现这些特点，本文提出了"双主体，多项目，客观描述"的原则，并根据这些原则给出了评测的标准和可操作的实施方法。操作的结果得到一组数据，称为评测相关数据。对与评测标准有关的因素来说，它们有的是定量分析的结果，有的是定性的客观描述，有的是有关的参数。评测的最终结果应该由用户根据这一组评测相关数据结合自己的主观需求和使用系统的情况综合考虑而定。对于准用户来说，这组评测相关数据与其主观需求的比较将成为他们确定购买意向的主要依据。当然，系统的研制人员也可以在缺少用户参与的情况下，从评测相关数据得出自己关心的评价结果。

表1：

级别	译文是否忠于源文(fidelity)	译文语义上的清晰度(clarity)
1	传达了源文句子的内容，不需修改	意思清楚，不需修改
2	传达了源文句子的内容，需要某些修改	意思清楚，需要修改
3	传达了源文句子的内容，但有些词序错误	意思清楚，需要修改
4	一般来说传达了源文句子的内容，但有从属关系、时态或数方面的问题	意思清楚，需要修改
5	没有恰当地传达源文句子的内容，丢失词或词组，有从属关系方面的错误	意思不清楚，但可以猜出来
6	没有传达源文句子的内容，丢失从句或短语	意思不清楚
7	没有传达源文句子的内容，丢失主语或谓语	意思不清楚

参 考 文 献

[1] 俞士汶 等 (1992) 基于测试集与测试点的机译系统评估，机器翻译研究进展，电子工业出版社，524－537。

[2] D. Arnold, L. Sadler and R. L. Humphreys (1993) Evaluation: An Assessment, Machine Translation, (8) 1－2, 1－24.

[3] P. W. Jordan, B. J. Dorr andJ. W. Benoit (1993) A First-Pass Approach for Evaluating Machine Translation Systems, Machine Translation, (8)1－2,49－58.

[4] M. Nagao et al. (1985) The Japanese Government Project for Machine Translation, Computational Linguistics, 11, April-September.