机器翻译评估方法评述

北京大学英语系　罗爱荣　北京大学计算语言所　段慧明

邮编: 100871 电话: 2501892

Assessment of Machine Translation Evaluation Methods

Dept. of English LUO Airong Inst.of Computational Linguistics DUAN Huiming

Peking University, Beijing, 100871 Tel: 2501892

机译评估与机译相辅相成，五六十年代机译研究蓬勃兴起，机译评估也由此诞生。1959、1960年，Bar-Hillel连续发表了“美国与英国机器翻译现状的报告”与“自动翻译语言研究的现状”两篇报告，在第二篇中他评价了美国、英国、前苏联、意大利以及以色列机器翻译研究的情况，最后得出悲观的结论，即完全自动的高质量翻译（FAHQT）不可能实现(Bennet, 1994)。因为Bar-Hillel的报告中没有提到任何评价标准，所以真正的机译评估始于1966年美国的ALPAC报告。这个报告制定了几个评估标准（Pierce & Carroll, 1966)。ALPAC中所采用的评估方法现在还具有很高的实用价值，但这个报告同样否定了机译研究，导致机器翻译停滞不前。七十年代末至八十年代，机器翻译又蓬勃发展，人们逐渐认识到ALPAC报告的偏见性，认识到机译评估应能客观比较各个机译系统而不是单纯比较人译与机译(Nirenburg, 1987)。但到目前为止，要开发这样一个理想的评估系统还是困难重重。首先，翻译没有绝对答案，所以翻译评估难以摆脱主观性。其次，不同的用户对机译及机译评估要求不同，因此很难制定出统一的评估标准。另外，机译评估还难在它通常需要把体系结构完全不同的各系统进行比较。鉴于这些困难，目前世界上机译评估色彩纷呈。单就评估方法而言，大致分三类：第一类为操作性评估(Operational Evaluation)，也称经济评估(Economic Evaluation)。这种方法是比较机译与人译每字或每页的花费以及耗时，其评估结果较直观，但没有涉及译文质量。第二类为说明性评估(Declarative Evaluation),又称质量评估(Qualitative Evaluation)。这种评估一般以译文的可理解性与忠实度为质量标准。ALPAC报告和我国的专家评测都使用这种方法。它的优势在于能直接表明译文质量,但其评估过程带有主观性。第三种评估方法为分类评估法(Typological Evaluation)。实现分类评估大致有两种途径，一是类似于语言教学中的“错误分析法”，即根据译后编辑中发现的错误类别与多少为系统评分，二是预先制定语言点覆盖面广的测试集，然后根据系统对测试集翻译情况评分。分类评估能测出各系统的弱点。有时，根据评估所采用的技术，机译评估可分为自动评估与非自动评估。目前世界上能够实现评估与评分过程全部自动化的只有两个系统：一个是Thompson的实验系统，(Thompson, 1991)；一个是北京大学计算语言所在俞士汶教授领导下于七五期间开发的世界上第一个自动评估系统—MTE(Thompson, 1991)。尽管七五期间MTE已初具规模(俞，1991; Yu, 1993; 俞,1994)，但还留有许多工作要做。近两年来，我们对MTE进行进一步改进，测试点做了较大调整与扩充，同时增设了双语语料库作为试题集的基础，而且增加了自动描述测试点程序,改造了辅助自动生成试题程序。今后我们要不断总结描述的规律,丰富双语语料库,争取扩大自动生成题库与描述测试点的范围,实现测试集的完全自动生成。