浅析机器翻译瓶颈及研发趋势

◆安　宁

（吉林建筑大学　吉林长春　130000）

机器翻译（Machine Translation）是翻译人员利用多种计算机翻译软件，将源语言（Source Language）文件，转换成另外一种目标语（Target Language）或者多种目标语言，翻译人员借助计算机器翻译一直是人工智慧领域中的重要研发方向。机器翻译进行翻译时涉及到自然语言（Natural Language，如中文、英文等）的加工合成，几乎已经覆盖自然语言处理的所有技术，并且有较为明确具体检测手段，可以当做自然语言处理技术的研发项目。不可避免的一个重要问题就是机器翻译若要占领翻译市场，就必须要面对来自人工翻译者的挑战。按照机器翻译的流程，翻译成品要译者人工润色和审核，这部分的人力成本将会占实际运作成本的大部分。也就意味着能够节省的时间必须多到一定程度，机器翻译才能达到实用化的阶段。在理想情况下，人工润色尽量无须参照原文，直接进行修饰即可。

机器翻译总体上可以分为直接式（Direct）、转换式（Transfer）及中介语（Interlingua）三大类，实操一般都是采用转换式。转换式机器翻译流程大致可分为三个阶段：分析、转换和生成。自然语言处理最大的难处，在于自然语言本身复杂更新较快，而且例外繁多。因此机器翻译所面对的主要问题，可以归纳为两大项：（1）文句中歧义（Ambiguity）；（2）语法不合设定（Ill-formedness）现象。自然语言的语法和语意中需通过上下文线索加以判断。以下是两类问题：

问题一，歧义。就是一个句子有多种解释。我们日常生活中不知不觉地充满了歧义句。虽然人们可以根据常识正确判断句意，但是对于依照文字规则来理解句子的计算机翻译软件来说，这就是一个歧义句。在做句子分析时，几乎在每一个环节（如断词、句法分析、语意分析等）都可能出现歧义。单个字或者词的解释往往会因前后的内容产生不同的意思。除此之外，判断句义依靠的线索在不同范围。因此在机器翻译过程中，若采用线性流水式的处理程序（Pipelined Architecture），则前面的模块则无法做出确定性的（Deterministic）判断，而须尽量多地保留候选者，而让后面的模块进行处理。所以，最终判断的时机应尽量延后，在累积足够信息后，再选择要使用的方法。这样才不会在开始就把正确的译法排除到考虑范围之外。

问题二，所谓的语法，是语言学家，依靠目前拥有的大量语料，归纳总结出的一些规则。这些规则是不完整的，往往有许多的例外。而且语言是一直在变迁的，无法要求语言的使用者，每字每句都合乎这些人为加工订定的文法，自然地也难以避免这样的情况出现在翻译稿件中。这些与设定语法的例子不符的地方包括不明的字汇，如新生的专有名词，和旧字新法的。这些现在部分来自单纯的疏失，例如错字、漏字、赘字、转档或传输时产生的乱码，或是不慎混入的标签（tag），也有些是已被大众所接受的字汇和语法。理想化的机器翻译，必须能够处理这些与设定语法不符的问题。

如何解决上述的歧义或语法问题，则需要大量知识。这些大量知识的管理，分类，储存和应用，是建立机器翻译时最大重点和难点。我们首先要将这些包含在语言学之内（intra-linguistic）、跨语言学的（inter-linguistic），以及超乎语言学之外（extra-linguistic）的知识抽取、表达出来，解释上述的语法和歧义问题，而且还要维护这个庞大的知识库。所以，我们要建立的知识库必须包罗万象，吸收涵盖各领域、各层面的知识。它本身就是一项艰难复杂的工作。也就是说知识库的建设维护是机器翻译系统开发最大的瓶颈。

一般来说知识的取得，和我们表现知识的方式有着紧密的联系。知识表现方式可以有很多不同的存在形式。其一就是加入知识库的规则，规则系统是由事先以人力建立好的大量规则所构成。在进行机器翻译时，翻译软件根据这些规则，进行二择判断，进行分析、转换和生成步骤，最后给出明确答案。这种方法被机器翻译广泛采用。它的优点在于贴近人的直觉，容易理解，遵循已有的语言学知识和规则，充分利用已有的经验和研究结果。逐渐使其参数化，不同的语言现象用几率扫描进行描述，积累量足够大时，语言模型就自然建立起来。其最大的优点在于通过参数，让计算机翻译软件在不同的条件下根据不同的偏好进行解释和加工，依靠参数估算任务给计算机翻译软件进行。

机器翻译的未来，研发高品质的翻译系统，需要的知识库是巨大并且琐碎的。对于知识的获取和管理，是机器翻译系统研发的瓶颈。近些年来，机器翻译系统的研发，已经渐渐地由规则库的方式转变成为参数化方式，并且其优越性已经得到了证明，也逐渐成为了主流。随着计算机行业的发展，计算机硬件性能的大幅提升，机器翻译软件已经突破了计算能力和记忆容量的限制。与此同时，人们生活和语言的发展使得语料库的规模也在爆炸似的增长，由译者来制定和模拟模型，利用计算机的处理优势进行语料库的加工，可以大大提高计算机学习效率，降低知识获取和管理的难度。这也是对机器翻译研发瓶颈的突破。放眼未来，如果能构建精准的模型，提高语言融合的锲合度，利用合适的规则抽取语料库中相关的知识，可以在专业特种行业领域发挥巨大的作用，提供高品质的翻译。如此一来，机器翻译最终可以拥有广泛的实用化领域和空间，也必将占有相当大的翻译市场份额。

作者简介：安宁（1980.01-），男，汉族，吉林长春人，吉林建筑大学国际合作与交流硕士，讲师，从事外国语言文学，机器翻译研究。