载《（香港）语文建设通讯》，2009年1月，第91期，17-22页

机器翻译 任重道远

冯志伟*

我非常欢迎我的好友王永成先生对于我在《一个关于机器翻译的史料错误》[1] 中的一些观点提出批评意见[2]。

王永成先生从事中文信息处理多年，取得了很多成果，他对于语言信息处理、对于机器翻译充满了豪迈的乐观主义的精神，我是十分赞赏的。

但是，对于像机器翻译这样的世界公认的困难问题，我们在乐观的同时，还应当冷静地对待。

如王永成先生所讲，当前机器翻译正在面临着重大的技术突破，处于重大进展的前夜。这是事实。

自从20世纪80年代末期以来，机器翻译确实面临着重大的技术突破，而且已经取得了很大的技术进展，学者们把这样的进展说成机器翻译研究历史上的一场“革命”。

1993年7月在日本 神户召开的第四届机器翻译高层会议 (MT Summit IV) 上，英国著名学者哈钦斯 (J. Hutchins) 在他的特约报告中指出，自1989年以来，机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是，在基于规则的技术中引入了语料库方法，其中包括统计方法，基于实例的方法，通过语料加工手段使语料库转化为语言知识库的方法，等等。他强调地说，这种建立在大规模真实文本处理基础上的机器翻译，是机器翻译研究史上的一场革命，它将会把自然语言的计算机处理推向一个崭新的阶段。

现在我们已经进入21世纪，语料库方法已经渗透到了机器翻译研究的各个方面，一些基于语料库的机器翻译系统已经建立起来，有的系统把基于语料库的方法和基于规则的方法巧妙地结合起来，取得了可喜的成绩。

2000年，在约翰·霍普金斯大学 (Johns Hopkins University) 的暑假机器翻译讨论班 (Workshop) 上，来自南加州大学、罗切斯特大学、约翰·霍普金斯大学、施乐公司、宾西法尼亚州大学、斯丹福大学等学校的研究人员，对于基于统计的机器翻译进行了讨论，以年轻的博士研究生奥赫 (Franz Josef Och) 为主的13位科学家写了一个总结报告 (Final Report)，报告的题目是《统计机器翻译的句法》(“Syntax for Statistical Machine Translation”)，这个报告提出了把基于规则的方法和基于统计方法结合起来的有效途径。

2002年，奥赫在国际计算语言学会议 (ACL2002) 上发表论文，题目是：《统计机器翻译的分辨训练与最大熵模型》(“Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”)，进一步系统地提出统计机器翻译的方法，获ACL2002 大会最佳论文奖。

目前，统计机器翻译已经成为机器翻译研究的主流。

根据 Google 的调查，统计机器翻译论文发表的情况如下图所示：

---

* 冯志伟先生，教育部语言文字应用研究所研究员，北京。

1. 冯志伟，“一个关于机器翻译的史料错误”，《语文建设通讯》，2008年，第89期。

2. 王永成，“关于机器翻译之我见”，《语文建设通讯》，2008年，第90期。

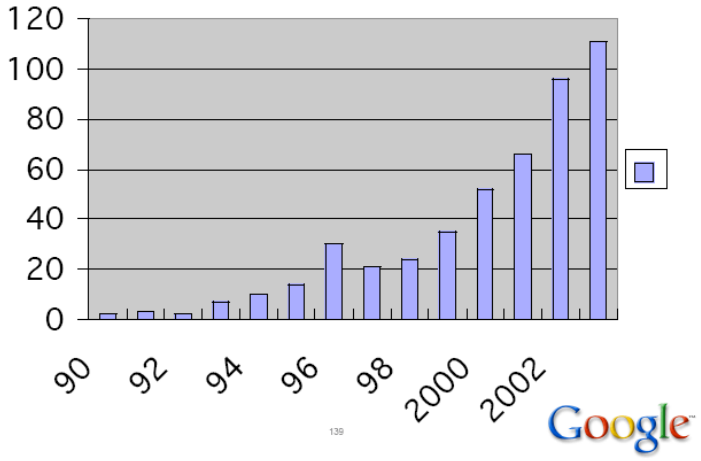

图 1　统计机器翻译论文增长情况

从图1中可以看出，统计机器翻译的论文是成线性增长的，其增长速度越来越快。

根据美国 NIST (National Institute of Standardization & Technology) 组织的统计机器翻译评测，汉语-英语机器翻译系统和阿拉伯语-英语机器翻译系统的 BLEU4 指标[3] 如下：

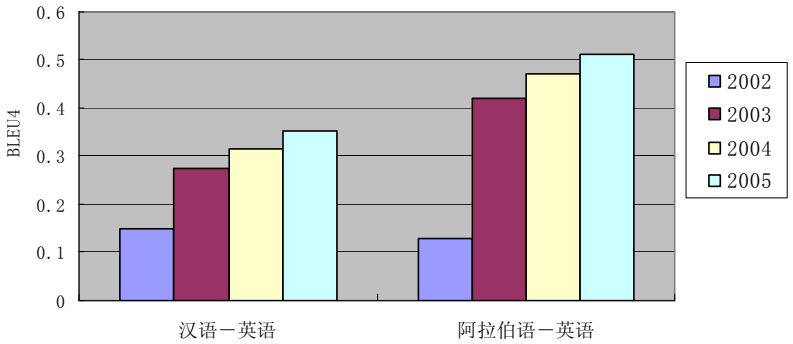

图 2 统计机器翻译系统的 BLEU 指标逐年提高

从图2中可以看出，统计机器翻译的质量正在逐年提高。

统计机器翻译的质量与语言模型的规模有密切关系。机器翻译的研究者们兴奋地发现，随着语言模型训练数据的增大，机器翻译的译文质量也相应提高了。如图3所示：

3. BLEU4 是评测机器翻译系统质量的一个数学指标，BLEU4 的数值越高，机器翻译系统的质量越好。

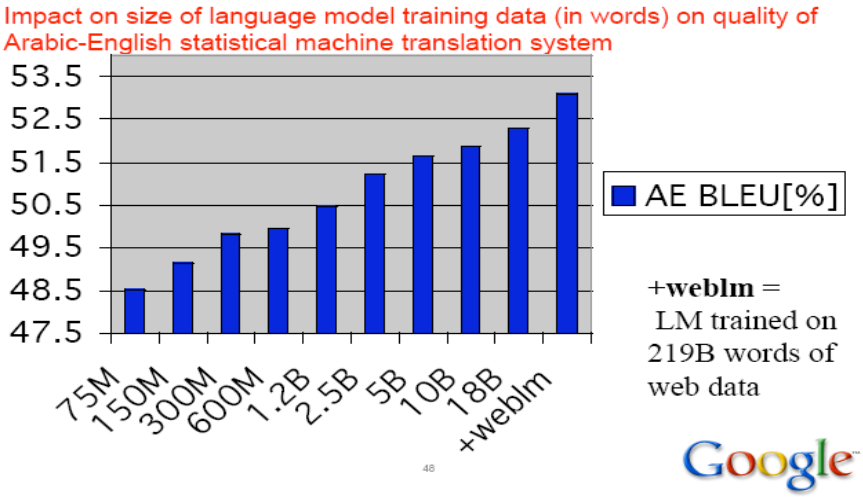

图 3 英语-阿拉伯语机器翻译系统的质量随着语言模型训练数据的增大而提高

图3所描述的这个结论是很有启发性的，它意味着，只要我们扩大语言模型训练数据的规模，就有可能提高机器翻译系统的质量。这似乎为提高机器翻译系统的质量找到了有效的方法。

当前机器翻译研究的这个重要结论与王永成先生的看法是有差异的。

王永成先生在他的文章中提出：“如果能找到字符、词、辞、短语、句子、字串等与概念之间的种种最简映射关系，你就能更方便、更智能、更快捷与更高质量地完成翻译的任务。此事，对机器来说，也是如此！”[4]，王永成先生显然主张，只要不断加深语言的分析深度，找到种种最简映射关系，就一定可以提高机器翻译系统的质量，从而实现全自动、高质量的机器翻译。

王永成先生的主张与当前机器翻译研究得出的结论是不同的。当前机器翻译研究的大量事实证明，在机器翻译中，对语言的分析并非越深越好！目前，人们更加倾向于通过扩大语言模型训练数据规模的方法，从大规模真实的语料中获取对于机器翻译有用的语言知识，并适当地进行一些浅层的语言分析，把基于统计的机器翻译与基于规则的机器翻译结合起来，争取得到最好的机器翻译结果，而这种最好的机器翻译结果，可以是全自动的，但却不一定是高质量的，而只是具有较高参考性的译文。如果一定要高质量的机器翻译译文，那么，就必须经过人工的加工，进行“译後编辑”(post-editing)，这样，就不是全自动的机器翻译，而是“机助翻译”(machine-aided translation) 了。这意味着，在目前的技术条件下，既要做到“全自动”，又要做到“高质量”，一箭双雕，是非常困难的，因此，我认为，在目前的技术条件下，全自动高质量的机器翻译还是不可能的。

2003年7月，在美国 马里兰州 巴尔的摩 (Baltimore, Maryland) 由美国 商业部 国家标准与技术研究所 NIST/TIDES (National Institute of Standards and Technology) 主持的评比中，奥赫获最好成绩。他使用统计方法，在很短的时间之内就构造了阿拉伯语和汉语到英语的若干个机器翻译系统。伟大的希腊科学家阿基米德 (Archimedes) 说过：“只要给我一个支点，我就可以移动地球。”(“Give me a place to stand on, and I will move the world.”) 而现在奥赫也模仿着阿基米德说：“只要给我充分的并行语言数据，那么，对于任何的两种语言，我就可以在几小时之内给你构造出一个机器翻译系统。”(“Give me enough parallel data, and you can have translation system for any two

4. 王永成，“关于机器翻译之我见”，《语文建设通讯》，2008年，第90期，第13页。

languages in a matter of hours.”) 这反映了新一代的机器翻译研究者朝气蓬勃的探索精神和继往开来的豪情壮志。看来，奥赫似乎已经找到了机器翻译的有效方法，至少按照他的路子走下去，也许有可能开创出机器翻译研究的一片新天地，使我们在探索真理的曲折道路上看到了耀眼的曙光。

在过去的数十年中，我曾经研制过汉译外和外译汉的若干个机器翻译系统[5]，我研制一个机器翻译系统往往需要几年的时间，而现在采用奥赫的方法构造机器翻译系统只要几个小时就可以了，可见，研制机器翻译系统的速度已经大大地加快了。

在研制速度加快的情况下，机器翻译的译文质量也在不断地提高。

例如，美国研制的阿拉伯语-英语的机器翻译系统的译文质量有明显的改善。下面是一段阿拉伯语的短文从阿拉伯语到英语的机器翻译的实例。稍微比较一下译文我们就可以看出，2002年的英语译文是很差的，而到了2003年的英语译文已经通顺可读了。

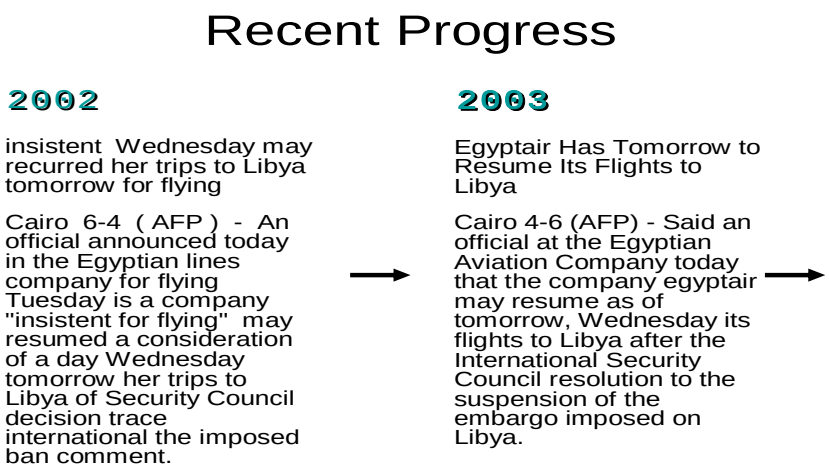

图4 机器翻译译文质量的比较

机器翻译确实已经取得了令人高兴的成绩。然而，奥赫只是说他可以在几小时之内给我们构造出一个机器翻译系统，并没有保证这样的机器翻译系统是全自动高质量的，因此尽管机器翻译取得了令人兴奋的成绩，我们仍然不能够说，现在就可以实现“全自动高质量的机器翻译”(Full-Automatic High-Quality Machine Translation，简称 FAHQMT) 了。

从已经推出的实用化机器翻译系统的译文质量来看，还不十分令人满意，对于一些简单的句子，译文一般不会有大问题，但对于一些稍长的句子或结构稍复杂的句子，译文质量就不能令人满意，有时简直是不可卒读；有的系统为了保持一定的译文质量，不得不将输入语言的範围加以严格的限制。因此，有许多商品化系统虽然卖出去了，但使用情况并不理想。例如，日本 富士通的 ATLAS 系统已售出300多套，但是据说只有10%的用户在使用。国内一些商品化的机器翻译系统，虽然也有一定数量的销售额，但用户使用的实际情况并不十分理想。带有探索性的大型机器翻译计划 EUROTRA 和 ODA，投入了巨大的人力和财力，但是至今仍然没有达到预期的目的。

我们认为，当前机器翻译系统的实用化和商品化问题仍然面临着严峻的考验。我们对于机器翻译产品的实用化和商品化，还不能估计得过分乐观。42年前（1966年）美国 ALPAC 报告[6]指出的

---

5. 冯志伟，《机器翻译研究》，中国对外翻译出版公司，2005年。

6 ALPAC, Language and machines: computer in translation and linguistics, A report by the

机器翻译遇到的“语义障碍”至今仍然存在，机器翻译技术至今似乎仍然没有取得突破性的进展。因此，今後进一步加强机器翻译基础理论和应用技术的研究，仍然是非常必要的。

由此看来，为了实现机器翻译系统真正的实用化和商品化，我们还要走相当长的一段路程。至于全自动、高质量的机器翻译，在目前还是做不到的，至于在今後一个相当长的时期内能否做到，还需要事实来证明。

机器翻译是自然语言处理 (Natural Language Processing，简称 NLP) 的一个应用领域，机器翻译处理的对象是自然语言，由于现实的自然语言极为复杂，不可能直接作为计算机的处理对象，为了使现实的自然语言成为可以由计算机直接处理的对象，在机器翻译领域中，我们需要根据处理的要求，把自然语言处理抽象为一个“问题”(problem)，再把这个问题在语言学上加以“形式化”(formalism)，建立语言的“形式模型”(formal model)，使之能以一定的数学形式，严密而规整地表示出来，并且把这种严密而规整的数学形式表示为“算法”(algorithm)，建立自然语言处理的“计算模型”(computational model)，使之能够在计算机上实现。

在自然语言处理中，算法取决于形式模型，形式模型是自然语言计算机处理的本质，而算法只不过是实现形式模型的手段而已。

从理论上说，建立这样的形式模型的时候，问题的解一般应当满足三个条件：第一，存在性：要证明问题的解是存在的；第二，唯一性：要证明问题的解是唯一的；第三，稳定性：要证明问题的解是稳定的。

这种建立形式模型的研究是非常重要的，它应当属于自然语言处理的基础理论研究。由于自然语言处理的复杂性，机器翻译形式模型的研究往往是一个“强不适定问题”(strongly ill-posed problem)，也就是说，在用形式模型建立算法来求解机器翻译这样的“强不适定问题”时，往往难以满足问题解的存在性、唯一性和稳定性的要求，有时是不能满足其中的一条，有时甚至三条都不能满足。因此，对于像机器翻译这样的强不适定性问题的求解，应当加入适当的“约束条件”(constraint conditions)，使问题的一部分在一定的範围内变成“适定问题”(well-posed problem)，从而顺利地求解这个问题。

机器翻译是一个多边缘的交叉学科，我们可以通过计算机科学、语言学、心理学、认知科学、人工智能等多学科的通力合作，把人类知识的威力与计算机的计算能力结合起来，给机器翻译的形式模型提供大量的、丰富的“约束条件”，从而解决机器翻译中的各种困难问题。机器翻译这个学科的边缘性、交叉性的特点，为解决这样的“强不适定问题”提供了有力的手段，我们有可能把机器翻译形式模型的研究这个“强不适定问题”变成“适定问题”，从而在某些局部的问题上获得比较完满的解决，以满足社会对于机器翻译的需求。但是，像机器翻译这样的“强不适定问题”是否一定可以实现全自动、高质量，在理论上还是一个需要认真地、严格地加以证明的问题，在实践上也需要通过强有力的事实来证明。

我研究机器翻译已经将近五十年的时间了，五十年前，我还是一个不谙世事的十九岁的小青年，现在，我已经是白发苍苍的老人了，我们这一代人正在一天天地变老；然而，我们如痴如醉地钟爱着的机器翻译事业却是一个新兴的学科，她还非常年青，充满了青春的活力，尽管她还很不成熟，但是她无疑地有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵常青的参天大树相比较，显得多么地微不足道，有如

Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Publication 1416, Washington，1966.

沧海之一粟。想到这些，怎不令我们感慨万千！“路漫漫其修远兮，我将上下而求索”，机器翻译的探索者任重道远，不论在理论方面还是在应用方面，我们都需要加倍地努力。我常常对我的研究生们说，“机器翻译道路崎岖，充满了艰险，在研究中往往会感到无路可走，但是，我们机器翻译的研究者就像侦察兵，对于侦查兵来说，没有道路的路才是最好的路”。在研究机器翻译的艰险道路上，虽然我们经常会陷入“山穷水尽疑无路”的困境，但是，当我们有所发现的时候，也会尝到“柳暗花明又一村”的愉快，顿时感到心旷神怡。这就是我们这一代机器翻译研究者的真实感受。当前机器翻译仍然面临诸多的困难，我们还要继续奋战，才能度过难关，走向一马平川的坦途。至于在将来机器翻译能否既做到全自动又做到高质量，还是一个需要严肃地加以探讨的科学问题，从目前我们所积累的知识和经验来看，我们还没有把握仓促地做出结论。 □