机器翻译评测介绍

黄瑾 刘洋 刘群

摘要:本文介绍了机器翻译评测的基本原理以及常用评测标准，包括人工评测方法、基于 n 元匹配的 BLEU 和 NIST 自动评测方法、基于准确率和召回率的 GTM 评测方法以及若干引入了语言知识的自动评测方法。文章还介绍了国内外几个著名的机器翻译评测项目，同时讨论了评测对于机器翻译的研究与发展所起的重大推动作用。

关键词：机器翻译；人工评测方法；自动评测方法；评测会议

1 引言

机器翻译领域最困难的任务之一就是对给定的翻译系统或者翻译算法进行评价，我们称其为机器翻译评测。事实上，在科学研究的任何一个领域，如果提出一个新的算法并声称其具有良好的性能，都需要使用某种通用的、被广泛承认的评价标准为这个算法“打分”。这个分数不仅要体现出该算法是好的，而且要求能够体现出与其他算法相比，该算法能够好到什么程度。与其他的评测任务面对的对象不同，机器翻译所处理的对象－语言－本身存在某种程度的歧义，无法像数学公式或者物理模型那样简单客观地描述出来，这使得为机器翻译结果进行客观的打分变得非常困难[1]。

本文结合翻译任务自身的特点，介绍机器翻译评测的常用方法、机器翻译评测任务的组织情况以及机器翻译评测对机器翻译领域研究的重要意义和巨大影响。

2 机器翻译评测技术[1]

2.1 人工评测

人是语言的使用者，是机器翻译成果的最终享用者，也就自然而然地成为了机器翻译系统质量的唯一评价标准(golden standard)。那么，什么样的译文质量才称得上好呢？我们知道翻译界推崇的“信、达、雅”这个最高标准即使是最好的翻译人员也很难做到，使用这个标准来要求现阶段还在牙牙学语的机器翻译系统实在是有些勉为其难了。因此，传统的机器翻译人工评测中使用的是诸如“忠实度”(Adequacy)和“流利度”(Fluency)这一类指标来衡量译文的质量。直观地说，这里的“忠实度”反映的是机器翻译系统生成的译文在多大程度上忠实于原文所要表达的意思，而“流利度”则用于评价译文本身是否流畅、是否符合目标语言的表达习惯等。理论上，这两个指标是相互独立的，译文可以非常通顺、很容易理解，但却与原文完全不相关。不过，对于机器翻译评测而言，这两个指标常常是相关的，一般忠实度比较差的译文也不容易理解。在人工评测的具体操作过程中，可进一步对上述指标进行分级，由双语专家对照原文判断每个译文的忠实度和流利度，并为其打分，系统的最终得分即为每个译文分数的累加[2][3][4]。

使用人工评价的方法得到的结果一般是十分准确的，但主要问题在于评测的成本太高，

[1] 这里介绍的机器翻译评测只关注机器翻译译文质量，是狭义的机器翻译评测。广义的评测还包括系统的效率、健壮性、界面的友好性和添加新资源难易程度等其他方面。

周期过长（评测过程可能长达几周甚至数月），评价结果也会随着评价人的变化和时间的推移而不同，这使得评价结果不可重复，缺乏客观性。在这种评测方式下，研究人员无法迅速得知系统改进的效果，延长了机器翻译系统的开发周期。

2.2 自动评测

人工评测的耗时耗力使得机器翻译的自动评测被提上了议事日程。如果一种语言中的每一个词都只有一种含义，只对应到另外一门语言中的一个词，任何一个句子都只有一种翻译方法，那么，不仅机器翻译评测变得轻而易举，机器翻译本身也就只需查找一一对应的符号替换表就可以了。但人类总是要发挥自己的聪明才智，尝试使用不同的方法来表达同一个含义，这造就了美丽丰富的语言世界，也给机器翻译及其评测带来了巨大的困难。我们知道，即使是一个人类专家，要对一个机器翻译的译文给出一个评分，也不是一件很简单的事情，要对源文和译文都有比较准确的理解才能做到。机器并没有办法去理解一个句子，如何能对一个译文句子进行自动评分呢？

如果一个机器翻译评测系统只根据原文就能自动地为若干译文打分并选择出其中最好的结果，那么这个评测系统本身就是一个质量更好的机器翻译系统了。因此，人们最先想到的自动评测的出发点就是给出一些标准的翻译结果，然后比较机器生成的译文与这些翻译之间的相似程度。我们称这些标准的翻译为参考译文(或者参考答案)。同一个句子可以有多个不同的参考译文，这些参考译文都表达同一个含义，但可能使用了不同的词汇，或者虽然使用了相同的词汇但在句中的词序不同。这样一来，机器翻译自动评测的问题转换为比较机器翻译系统输出的一个翻译结果和多个通过人工产生的正确的参考译文之间的相似度的问题，使用不同的相似度计算方法即可得到不同的自动评测方法。

2.2.1 基于 n 元匹配的自动评测方法

我们考虑如下两个机器翻译系统生成的翻译结果：

源语言文本: 今年前两月广东高新技术产品出口37. 6亿美元

系统译文1: The new high-tech products in Guangdong exported 3.76 billion dollars in the first two months this year

系统译文2: This year , the former two of Guangdong , the export of hi-tech products 37.6 yi US dollars

源语言文本: 今年前两月广东高新技术产品出口 37.6 亿美元

系统译文 1: The new high-tech products in Guangdong exported 3.76 billion dollars in the first two months this year

系统译文 2: This year, the former two of Guangdong , the export of hi-tech products 37.6 yi US dollars

从直观上看，上面两个翻译结果的质量有较大的差别，第一个翻译结果明显通顺流畅易于理解，如何将这种人的直观印象与具体的客观分数统一起来呢？我们引入三个人工翻译的参考译文来进行比较：

参考译文1：Guangdong's export of new high technology products amounts to US$ 3.76 billion in first two months of year

参考译文2：Guangdong Exports US $3.76 Billion Worth of High Technology Products in the First Two Months of Th year

参考译文3：In the first 2 months this year, the export volume of new high-tech products in Guangdong Province rea 3.76 billion US dollars

可以看出，质量较好的系统译文 1 与三个参考译文共现了很多个翻译片段，与参考译文 3 共现“the”，与参考译文 3 共现“new high-tech products in Guangdong”，与参考译文 1 共现“3.76 billion”，与参考译文 3 共现“dollars”，与参考译文 2 共现“in the first two months”和“this year”（为了便于查看我们用下划线标出了系统译文中共现的部分，不考虑大小写）。相比而言，系统译文 2 与上述三个参考译文的共现片段比较少。

通过上述比较，我们可以很容易地写出一个评价算法来评价上述翻译结果的质量。通过引入一个称为 n 元匹配的概念，可对翻译结果 1 给出比翻译结果 2 更高的分数。n 元匹配的含义是：翻译结果与参考译文句子中的任意连续 n 个单词完全相同，这里的 n 值可以取任意正整数。基于 n 元匹配的策略非常类似常用的准确度的计算思想，首先统计系统译文与参考译文中共现的 n 元匹配的个数，再除以相应的系统译文中 n 元词的总数，用这个比值来表示相应的 n 元准确率。在上述例子中，系统译文 1 中除了“exported”这个单词，其他所有的单词都在三个参考译文中出现过，因此其一元准确率为 16/17，相应的二元准确率为 10/16，而系统译文 2 的一元准确率和二元准确率只有 11/18 和 3/17 而已，分数的差异与人工评测的结果取得了一致。在这个例子中我们通过引入一种基于 n 元匹配的策略，可以同时考虑多个参考译文，而且这种方法在给出的例子中表现良好。问题好像解决了，但机器翻译系统并不一定总是按规矩出牌，在下面的例子里面，系统试图瞒天过海得个高分：

系统译文: the the the the the the the

参考译文1: The cat is on the mat

参考译文2: There is a cat on the mat

按照标准的 n 元准确率的计算方法，上面这个翻译结果的每一个词都在参考译文中出现过，因此其一元准确率为 7/7!看来评测算法需要考虑这种情况并且加以处理，保证那些已经用于计数的词不被重复计算。首先统计出 n 元词在一个句子中可能出现的最大次数，即该词在系统译文和参考译文中出现的次数中的较小值。在上面的这个例子中，the 这个词在系统译文中出现了 7 次，但是在参考译文中最多也只是出现了两次，因此在计算一元准确率时只能计数两次。我们称这种策略为修正的 n 元准确率计算方法。使用该策略，上例中系统生成译文的一元准确率变为 2/7，而二元准确率为 0。采用修正的 n 元准确率计算方法的原因可以解释为：对于同一个 n 元词，如果在同一个参考译文中出现的次数多于系统生成译文中的次数，那么系统译文中可能缺失了此部分的信息；而如果情况相反，那说明系统译文中给出了多余的信息，不该重复计算。

现在的评测方法已经对系统译文中出现的正确翻译片段给予了奖励，对那些在系统译文中出现但不在任何参考答案中出现的词条进行了惩罚，而且通过修正的 n 元准确率计算方法避免了系统译文中出现的过多重复词条所带来的误判。但进一步考察可以发现：在计算 n 元准确率时是以系统译文的 n 元词数作为分母的，如果翻译系统为了追求准确度，只翻译那

些最有把握的单词，会使译文变得非常简短。此时情况会变成什么样呢？考虑如下极端的例子：

系统译文：of the

参考译文1: It is a guide to action that ensures that the military will forever heed Party commands

参考译文2: It is the guiding principle which guarantees the military forces always being under the command of the P

参考译文3: It is the practical guide for the army always to heed the directions of the party

系统译文只给出了两个无关痛痒的词，但是我们的评测系统给出的分数是：一元准确率为 2/2，二元准确率为 1/1!看来我们不得不再增加对于句子长度的限制：如果系统译文的词数少于参考译文的词数则对分数乘以一个系数进行惩罚。为了增加灵活度，这种长度惩罚是在整个测试集的层面上而不是句子级别上进行的。

至此，我们得到了机器翻译评测领域目前使用最为广泛的BLEU评测标准的基本方法[5]。BLEU的全称是Bilingual Evaluation Understudy，它由IBM于 2002 年提出。如上所述，BLEU方法实施起来非常简单，通过比较系统译文与多个参考答案之间字面上的相似度来为系统翻译结果的质量打分，只需要请几位不同的双语专家对原文进行翻译，计算时完全根据字面字形的相似度，不需要对参考译文和机器译文的意思加以理解。实践也证明，采用这种方法对机器翻译的译文质量进行评测，其结果与人工评测的结果相当一致。

BLEU方法是基于n元匹配的这一类方法中的典型代表，这一类基于n元匹配的评测方法综合考虑了人工评测中的忠实度和流利度指标。当n取值为 1 时，与参考译文使用更多的相同的词被认为是更加忠实于原文，而n取大于 1 的值时，可认为要求翻译结果有更长的片段与参考译文相同，体现出流利度方面的要求。类似的方法还包括NIST[6]方法，该方法由美国标准和技术研究所(National Institute of Standards and Technology，简称为NIST)提出并命名，在BLEU方法的基础上，综合考虑了每个n元词的权重，对于那些在参考译文中出现次数更少的词赋给更高的权重以体现其所包含的信息量。另外NIST采用了算术平均来代替BLEU中使用的几何平均值，这使得一元词的共现次数对于评分结果的影响更大，评价结果侧重于反映翻译结果的忠实度。NIST还改进了长度惩罚因子，减少了译文的长度对于评分结果的影响。

BLEU 和 NIST 是最常用的两种机器翻译自动评测方法，但就如我们所看到的一样，这一类评测方法并不是在真正地在评价系统译文与原文的一致程度，而是根据若干个参考译文为系统译文打分而已。系统得分似乎与待翻译的原文没有关系，参考译文的数量多寡与质量好坏才是影响评测结果的关键因素。

2.2.2 基于准确率和召回率的自动评测方法

基于n元匹配的自动评测方法是一种基于准确率的方法，与参考译文越相似的系统译文可获得越高的分数。研究人员提出了一些同时考虑召回率的自动评测方法，其中比较典型的是纽约大学提出的GTM评测方法[7]。该方法应用了图的最大匹配算法来计算词的共现次数。图 1 描述了这个计算过程：

图中的黑点表示参考译文和系统译文共现的词的位置。图中的 B 和 C 都存在两次以上的共现，这些点被认为是互相冲突的，在实际计算时应避免重复，只保留一个即可。使用图搜索算法找到最大匹配的区块，如图中灰色部分所示，并在此基础上计算最大匹配块长度(Maximum match size，简称为 MMS)，即上述若干匹配区块的长度的平方和再开方的值，准确率和召回率也通过 MMS 进行计算，在上例中分别为 4.6/8 和 4.6/10。系统最终的得分使用准确率和召回率的调和平均值 F 值来表示。

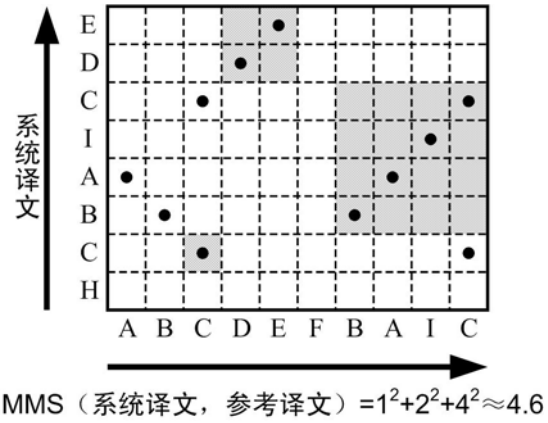

图1. GTM 评测方法使用的基于图的最大匹配的共现次数计算方法

基于 GTM 的评测标准最大的优点在于，没有人为地设置匹配时的最高阶数值（即 n 元匹配中的 n 的最大值，BLEU 方法一般只考虑到 4 元或者 5 元匹配），图的最大匹配算法会自动地寻找针对某个参考译文的最大匹配词数。据称这种基于 F 值的自动方法与人工评测的一致性可以比 BLEU 或者 NIST 更高。缺点在于，计算 MMS 本身是一个 NP 难的问题，比较费时。

2.2.3 引入语言知识的自动评测方法

BLEU\NIST以及GTM方法都是基于字面的完全匹配，并没有理解系统译文和参考译文的含义，引入多个参考译文只能在一定程度上减轻这方面的缺陷。研究人员尝试使用基于更多语言学知识的评测方法来评价翻译质量的好坏。早在十几年前，北京大学计算语言学研究所的俞士汶教授就提出了一种基于测试点的机器翻译自动评价方法[4]。这种方法并不直接评价译文句子，而是考虑系统译文在给定的测试点上的质量。其基本原理类似于在考试出题时设置考点，考生答题时答到相应的考点即可得分。我们首先确定要考察的一些测试点，例如词汇量、译词的选择、固定搭配、句法结构的理解、词序调整等，然后对每一类测试点，出一些测试句子，对于每个句子，仅仅对某一个具体的测试点进行测试。比如说，关于英汉翻译的固定搭配问题，我们可能会出一个测试句子“There are nine horses on the farm.”，目的仅仅是考察系统是否能够正确地给出与“马”搭配的量词“匹”。自动评测系统将考察机器翻译的译文中是否包含“匹”字，如果有，则认为这个译文在这个测试点上表现正确，反之，则认为这个译文在这个测试点上翻译错误，而不管这个译文的其他部分如何。这种评测方法可以综合分析机器翻译以及两种语言自身的特点，确定不同类型、不同难度的测试点，将对一个完整译文的评测问题分成一个个单独的测试点的评测，每个测试点本身只要通过简单的字符串匹配即可给出评价了。基于测试点的评测方法是一种非常巧妙的方法。使用这种评测方法，通过对评测结果进行细致深入的分析，很容易知道系统在哪些语言问题上处理得不够好，有针对性地对系统进行改进即可进一步提高系统翻译的质量。

近几年来研究人员又提出了一些融合了语义知识的机器翻译评测方法，由美国南加州大学信息科学研究所（ISI/USC）[2]的Zhou Liang等人提出的ParaEval方法通过引入语义对BLEU

[2] Information Sciences Institute, University of Southern California

方法进行改进，在计算n元匹配时除了完全匹配的情况，还同时考虑具有相同含义的词对的匹配情况。这里具有相同含义的词对是通过在一个训练语料库中自动学习得到的。以汉英翻译为例，首先在一个双语语料库中进行自动抽取，将具有相同汉语翻译的多个英文短语收集起来，认为他们表示相同的语义。在进行翻译结果的评测时，系统译文与参考答案相比匹配的数目除了BLEU中采用的n元词，还可以包括那些虽然词形不同，但出现在上述相同含义数据集中(即语义相同)的词汇[8]。计算所的刘洋等人在GTM自动评测方法的基础上，引入模糊匹配的策略，在不使用其他语言资源的情况下，尝试了基于词形等方法的相似度计算方法[9]。卡耐基梅隆大学(CMU)[3]提出的Meteor自动评测方法在计算 1 元匹配时引入了含有丰富同义词知识的WordNet[10]来增加匹配的数目[11][12]。这些融合了语义方法的机器翻译自动评测方法，试图利用更多的语言学知识，更好地评价机器翻译系统译文的质量，使其结果与人工评价更为接近。

2.3 自动评测方法的评测

应该如何评价机器翻译自动评测方法同样是一个有趣的问题[4]，一方面，由于人是翻译质量的最终裁定者，因此好的自动评测方法的结果应该尽可能地与人工评测的结果一致。这种一致性不仅体现在与人工评价结果排名的一致性上，而且还要求分数所体现出的各个系统之间的差异与人工评价结果存在一定的正相关关系。有很多统计方法可以用来评价这种相关性，例如斯皮尔曼等级相关系数(Spearman Rank-Order Correlation Coefficient)可通过计算两个排序序列之间次序的差值的平方和来描述两种排序方法之间的一致程度。另一方面，评测方法的易操作性也成为需要频繁进行评测的研究者们考虑的重要因素。BLEU方法虽然只是基于词形字面上的相似度匹配，但是操作简单、所需资源很少，使其得到广泛使用；与之对应的测试点评测方法，非常细致地考虑了翻译时的多种语言现象，却面临制作测试题很难的困境，没有被广泛采用；引入了语义判断的ParaEval方法，需要预先通过训练得到一个含义相同词汇的数据集，操作起来不太容易，评测结果也会与训练语义集所使用的数据、方法等相关，不利于比较。

当然，人们对于自动评测方法的评价和认可程度与机器翻译发展的程度是息息相关的。目前已经比较成熟的基于短语的机器翻译，由于使用统计的方法，并不真正理解待翻译文本的含义；以 BLEU 为代表的基于 n 元匹配的评测方法，在计算时也只采用字面上的相似性度量，BLEU 方法基本可以满足人们当前的需求。但随着机器翻译系统翻译质量的提高，人们发现针对机器翻译某些特别困难的部分所做的改进并不总是能在 BLEU 指标上有所提高。以机器翻译中较为困难的实体名翻译为例，如果翻译系统对原文中的某个人名给出了音译名，但是与参考答案中的音译人名略有差别，虽然系统译文给出了人名信息使得句子更加完整更易理解，但在基于字形匹配的 BLEU 标准看来，是添加了一个错误的翻译结果，系统译文变长却没有新增加匹配的 n 元词数目，打分反而有可能下降。机器翻译自身的发展将促使人们研究并且使用更为复杂的、能够更好地评价机器翻译质量的自动评测标准。

3 国内外著名的机器翻译评测项目

机器翻译评测项目主要有两种组织形式，一种由官方机构组织，比如由美国标准和技术研究所组织的NIST机器翻译评测[13]、欧盟主办的TC-STAR机器翻译评测[14]以及由中国科学院计算技术研究所组织承办的 863 机器翻译评测[15]等。这一类型的评测活动通常具有一定

[3] Carnegie Mellon University
[4] 希望我们不会陷入证明算法有效性的无限循环当中

的项目背景或者研究背景，受到一定的政府基金的资助。另一种评测由学术机构组织，这种学术机构可能是一些比较长期从事此类研究的学术机构，也可能是一些较大型的、参与单位众多的合作项目。此类评测大多是由研究兴趣驱动的，参与者对某个领域具有共同的兴趣，非常愿意在一起组织评测进行交流，比较典型的例子就是IWSLT(the International Workshop on Spoken Language Translation)机器翻译评测[16]和费利普.科恩（Philipp Koehn）组织的机器翻译研讨会中的评测项目[17]。TC-STAR和IWSLT评测都是面向语音翻译的，而NIST评测主要面向篇章类文本的翻译，特别是新闻翻译。下面我们主要介绍目前机器翻译领域最为权威、参加单位最多、影响最大的NIST机器翻译评测。

从 2002 年起，在美国国防部高等研究计划局(DARPA)[5]资助的项目TIDES[6]的框架下，美国国家标准和技术研究所出面组织了NIST机器翻译评测。NIST评测每年举办一次，主要考察的语言对是汉语到英语以及阿拉伯语到英语，并且只对各参评系统的机器翻译结果的质量进行评测，对系统本身的其他方面不做评价。评测采用自动评测和人工评测相结合的办法，其中自动评测主要采用BLEU指标。自动评测的结果在参评系统提交译文之后的几分钟之内即可得到，人工评测则在随后几个月中进行，并在评测后的研讨会中在各参评单位内部公布自动评测和人工评测的结果。2002 年的第一次NIST机器翻译评测一共有 8 个单位提交了 30 多个系统译文，最终德国亚琛工业大学(University of Technology Aachen.)基于统计的机器翻译系统在所参加的评测中取得了最好的成绩。2006 年 NIST 机器翻译评测的参赛单位以及系统已经增加到 46 个。图 2 显示了从 2002 年到 2006 年 NIST 评测中在两个语言对上的最好成绩的变化趋势。

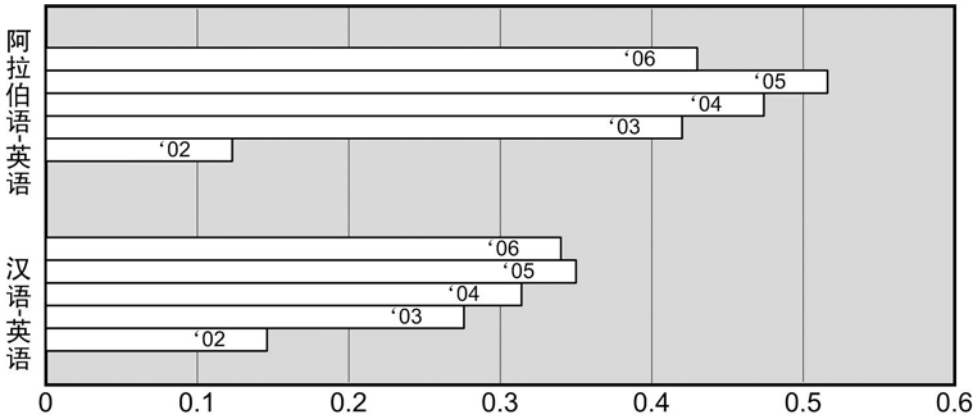

图2. NIST 评测 2002 年-2006 年中最好系统在中文和阿拉伯语上最好成绩

从图中可以看到，2002-2005 年统计机器翻译系统的性能几乎是呈直线上升的，但 2006 年的成绩略有降低。严格地说，每次评测的结果是不具有可比性的，因为每次的测试集都不同。不过，由于 2002-2005 年的测试集无论从规模和类型来说都基本一致，因此可以认为这几次的测试题难度是大致相同的（实际上每年难度略有提高），而这几年评测的最好成绩逐年上升反映了机器翻译研究取得的进展。2006 年的成绩略有下降，并不是因为参评系统的性能有所降低，而是因为 2006 评测的题目与往年相比有了较大的不同，评测语料不仅规模大大增加，而且语料的来源比较分散，类型比较多样。往年的评测数据基本上都是新闻(newswire)文本，而这次增加了网络新闻组(newsgroup)文本、广播新闻(broadcast news)和广播谈话(broadcast conversation)几种类型文本。实际上，2006 各单位的参评系统比 2005 年也都有了普遍的提高。可以看到，目前统计机器翻译系统的性能还在明显地上升，这也是统计机器翻译引起人们普遍兴趣的一个重要原因。

在 2002 年的 NIST 评测以前，很多人对统计机器翻译到底能达到什么水平，还心存疑虑。在 1990 年代初，IBM 公司开发的世界上第一个统计机器翻译系统 Candide 在 DARPA

[5] Defence Advanced Research Projects Agency
[6] Transligual Information Detection, Extraction and Summarization

的评测中，取得的成绩与著名的商用系统 Systran（采用规则方法）大致相当。研究人员虽然相信 DARPA 的评测是客观的，但对 Candide 和 Systran 水平相当这个结果，还是有很多人持怀疑态度，觉得 IBM 取得的这个成绩可能有一定的偶然因素。但 2002 年 NIST 评测的结果出来以后，Systran 的成绩只排在了所有参评单位的中游，参加评测的前几名全部是采用统计方法的系统，这个结果出乎很多人的意料。虽然那次评测的结果并没有公开，只是在参评单位内部之间交流，外部并不了解，但潜在的变化已经发生，机器翻译领域的研究重点开始转向了统计机器翻译。这次评测可以说是一个很重要的分水岭。在后面几年的评测中，统计机器翻译的水平逐年提高，新方法不断涌现，而以 Systran 为代表的传统机器翻译系统水平几乎停滞不前，统计机器翻译的优势更加得到巩固。

在 2002 年 NIST 机器翻译评测的参评系统中，也有我们当时开发成功的一个汉英机器翻译系统。这是一个基于规则方法的系统，是我们历经多年开发的成果，虽然还只是一个实验系统，规模有限，但我们自信，在当时也是国内最好的汉英机器翻译系统之一。但是，在这次评测中的成绩却非常差。不仅排名非常落后，而且结果评分也是差距非常大。这个结果同样让我们感到极为震惊，我们没有想到我们历经这么多年研究的成果在统计机器翻译方法面前会这么不堪一击。也正是这次评测，让我们看到了统计机器翻译的前景，促使我们痛下决心，改弦更张，走向了统计机器翻译研究之路。当然我们的统计机器翻译研究之路也并不平坦。刚开始由于势单力薄，我们只是在广泛阅读文献的基础上，开展了一些基础性的研究工作，比如说汉语词语切分、句法分析、语料库收集和加工、词语对齐等等。直到 2004 年，我们才开始真正全面开展统计机器翻译研究，而到 2005 年，我们开发的第一个统计机器翻译系统再次参加 NIST 评测，结果依然不理想。再经过一年的艰苦努力，到 2006 年，我们的系统在 NIST 评测中名列前茅（在参评的 24 个单位中名列第 5），我们的统计机器翻译研究才总算是取得了阶段性的进展。现在，统计机器翻译研究竞争日趋激烈，各国竞相投入大量资金开展研究工作，一些原来不在这个领域的著名研究机构纷纷参加进来，一些其他领域的成名人物也陆续进入这个研究领域。这个领域的研究越来越激动人心，而我们的对手也更多，更为强大。这激励着我们付出更多的努力，去取得更好的成绩。

分析近几年 NIST 评测的排名可以发现一个有趣的现象。这就是汉英机器翻译的水平远远落后于阿英机器翻译。由于测试集不同，这两类结果理论上是不可比的。但基于上面同样的理由，由于这两项评测所采用的数据类型、数据规模都非常相似，而且评测的时候，主要是将机器翻译的结果与多位人工译员翻译的结果相比，因此可以说，这个比较还是能说明一些问题的。汉语理论研究界有一种观点，认为汉语由于缺乏形态上的变化，导致汉语的自然语言处理比英语等形态丰富的语言会困难得多。这个评测结果，从一定程度上证明了这种观点是有根据的。

在NIST评测中有一个颇具传奇色彩的人物，就是毕业于德国亚琛工业大学的博士生弗朗茨.约瑟夫.欧赫（Franz Joseph Och）。在 1999 年约翰霍普金斯大学夏季研讨班（JHU Summer Workshop 1999）上，他开发出了著名的IBM模型训练工具Giza。在 2002 年NIST评测中，取得第一名的亚琛工业大学的机器翻译系统就是由他开发的。欧赫 2002 年从亚琛工业大学毕业后进入美国南加州大学信息科学研究所(ISI/USC)工作，同时作为Language Weaver公司的顾问，后来于 2004 年加盟了谷歌（Google）公司。他所到的每一个地方都稳拿当年NIST机器翻译评测的第一名。尤其是 2005 年的NIST评测中，他所在的谷歌公司开发的汉英机器翻译系统取得了 0.35 的BLEU值，比第二名的南加州大学(即他原来所在的单位)的系统的性能提高了近 5 个百分点。2005 年在汉语到英语方向取得前四名的单位分别是谷歌公司、美国

南加州大学信息科学研究所(ISI/USC)、马里兰大学(UMD)[7]和德国亚琛工业大学(RWTH)；2006 年汉英翻译的这个排名变成美国南加州大学信息科学研究所（ISI/USC）、谷歌公司、美国Language Weaver公司（LW）和德国亚琛工业大学。其中 2006 年的这四个研究单位的技术都有一定的渊源关系，全部都是欧赫曾经或者正在工作的地方。在 2006 年评测中，除了汉英机器翻译的受限语料项目，其他所有项目的第一名都是谷歌公司。欧赫不仅仅是在评测中成绩绝对领先，而且在研究方面也是非常出色的。他这些年来发表的很多论文，包括博士论文，都成了统计机器翻译研究领域的经典，被人广泛引用和验证。更难得的是，他对自己的研究工作持一种非常开放的态度，一点都不保守。目前统计机器翻译研究领域一些著名的开源软件，如IBM模型训练工具Giza++、最大熵模型训练工具YASMET都是他开发的。这一切显示了欧赫不愧为统计机器翻译研究的第一人。

欧赫是统计方法的忠实信徒。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.”）。欧赫模仿阿基米德的口吻说：“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构造出一个机器翻译系统。”（原话是“Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.”。）在欧赫的研究中，数据规模总是第一位的。他也尝试过使用一些句法知识，但他的最后结论是，句法知识对统计机器翻译毫无用处，甚至有反作用。因此，欧赫总是试图用最简单的模型和最大量的数据取胜。到谷歌公司以后，谷歌公司对海量数据的驾驭能力使得欧赫如鱼得水。他把谷歌公司在 Internet 上采集的所有英语文档都用来训练英语的语言模型，动用了谷歌公司数千个 CPU 组成的计算机集群进行计算。如此巨大的语言模型，使得他所代表的谷歌公司在 NIST 评测中取得了其他单位难以撼动的优势地位。他这样做，也是利用了 NIST 评测规则中的一个不太合理的规定。在 NIST 评测中，有两类项目：受限语料项目和不受限语料项目。在受限语料项目中，参评者只能使用评测组织者提供的训练语料进行训练。而在不受限语料项目中，参评者可以使用任何语料进行训练。研究人员一般比较关注受限语料项目的评测，因为只有在语料受限的情况下，参评单位之间的结果才是可比的。大家比的是算法的好坏，而不是数据的规模和质量。但 NIST 评测规则不太合理的地方在于，对于受限语料项目，NIST 评测只限定了用于训练翻译模型的双语语料必须受限，但对于训练语言模型的单语语料却没有任何限制。这使得谷歌通过这种方式训练出来的语言模型也可以参加 NIST 的受限语料项目评测。但他这种做法也遭到了越来越多研究人员的质疑，也许作为谷歌公司的企业行为，这样做是无可非议的，但作为研究人员来说，他这么做对其他研究人员来说，无疑是不公平的。另外，欧赫到谷歌以后，虽然系统做得非常强大，但他现在已经很少发表论文，通常只是在大会上做一些特邀报告。这样做的原因可能是因为谷歌公司要保守商业秘密吧。但这无疑也是让人觉得非常遗憾的。好在统计机器翻译领域现在人才辈出，出现了很多新的重量级人物，大家并不会因此感到寂寞。

NIST评测见证了统计机器翻译研究取得的各种进展，一些新的方法都是在NIST评测中得到验证，才被其他研究人员普遍接受的。早期的统计机器翻译系统都采用IBM提出的基于词的统计翻译模型。后来，很多研究者试图提出各种基于短语的模型，但都没有被普遍接受。直到后来，欧赫所在的德国亚琛工业大学（RWTH）小组采用了他们提出的新的基于短语的统计翻译模型，并被其他研究者证明有效，这种模型才被人普遍接受。另一位著名的统计机器翻译青年才俊费利普.科恩（Philipp Koehn）推出了开源的基于短语的统计机器翻译系统Pharaoh[18]，基于短语的模型终于成为了研究的主流。

[7] University of Maryland

虽然欧赫对统计机器翻译中使用句法知识没有兴趣，但随着研究的深入，基于短语的统计翻译模型逐渐走到了尽头，很难再有进一步提高，越来越多的研究人员开始考虑在模型中引入句法结构知识。这些方面的研究工作目前已经开始崭露头角。2005 年的评测中，马里兰大学的系统获得了第三名。他们的系统就采用了一种引入了句法知识的统计翻译模型——“层次短语模型”[19] 。这是由一位叫蒋伟（David Chiang）的华人研究人员提出的。在 2006 年的评测中，欧赫在几乎所有的项目中都取得了第一，而唯一一个失手的项目就是汉英机器翻译的受限语料项目。在这个项目中，取得第一名的是南加州大学信息科学研究所（ISI/USC）。ISI的负责人是凯文.奈特（Kevin Knight）教授。他也是统计机器翻译的倡导者和领袖人物之一。1999 年的霍普金斯大学夏季讨论班开发出了著名的统计机器翻译开源工具集Egypt（其中的训练工具就是Giza），凯文.奈特就是这次研讨会的发起人之一。凯文.奈特一直对在统计机器翻译中引入句法知识抱有很强的信念，他的一个博士生山田（Yamada）在这方面做了一些工作。后来，欧赫到ISI以后，一味强调大规模数据训练，凯文.奈特的句法模型方法受到了抑制。欧赫离开ISI去谷歌以后，凯文.奈特所领导的ISI统计机器翻译小组重新开始重视句法模型研究，终于在 2006 年NIST评测的汉英翻译受限语料项目中一举击败谷歌公司，这也标志着基于句法的统计翻译模型开始在评测中体现出优势。在 2006 年NIST评测中，获得第 5 名的中科院计算所也采用了基于句法的统计翻译模型，这也给基于句法的模型提供了有力的支持。

从上面的介绍可以看到，机器翻译评测不仅仅是给各参评单位提供了一个可以公平比较的平台，也为各种新的机器翻译理论提供了一个很好的实验场，为机器翻译的各个研究单位和研究人员提供了一个充分展示自己水平的舞台，这也使得机器翻译研究更加精彩纷呈。

4 机器翻译评测的重要意义

机器翻译评测对机器翻译研究的影响是十分巨大的。美国国家科学院下属的语言自动处理咨询委员会(Automatic Language Processing Advisory Committee，简称ALPAC委员会)在上世纪六十年代发布了一个题为《语言与机器》的报告(简称ALPAC报告)[20] 。该报告对机器翻译采取否定态度，将全球的机器翻译研究打入了冷宫。这是历史上最著名的一次机器翻译评价活动改变机器翻译发展状况的事件。

在上世纪八十年代机器翻译研究复苏以后，人们又重新开始关注机器翻译评测。以NIST评测为代表的一系列机器翻译评测，对机器翻译研究工作起到了非常重要的推动作用。机器翻译评测对于机器翻译研究的影响是多方面的。有了自动评测工具，研究人员对系统进行任何一点小的改进后，随时可以使用自动评测工具测试系统总体上翻译质量是否有所提高。这无疑会缩短研究周期，对机器翻译的研究起到极大的推动作用；各种由官方机构或者学术机构组织的评测活动，给出可使用的数据集，制定统一的测试方法和评价标准，公开相应的测试数据、参考答案和评测工具，通过评测构造了一个共同的实验平台、为不同的研究方法之间提供一个可以比较的基准、加强了不同研究队伍的合作与交流。这使得不同的研究者也可以在相同的条件下进行实验比较，获得可比的数据，随时了解自己研究的工作水平，促进研究的进展。另外，在现在组织的各种机器翻译评测活动中，除了评测本身以外，一项非常重要的活动就是评测之后进行的学术研讨会。各个参评者，尤其是那些在评测中表现出色的参评者报告的研究方法，特别是他们所采用的新技术、新思想，成为同行们最为关心的内容。这也是技术评测的最终意义所在[21][22]。

从更高的层次来看，机器翻译评测本身对于机器翻译领域的研究和发展起到了一种引导作用。例如，由于 NIST 机器翻译评测的项目主要是汉语到英语和阿拉伯语到英语，使得汉

英和阿英成为国际上机器翻译研究最多的语言对。机器翻译所需要的大量数据资源、相关的词法分析、句法分析等工具也多是针对这几种语言的；由于 NIST 评测主要采用 BLEU 作为评测指标，机器翻译中使用到的一些参数优化算法都直接使用 BLEU 分值作为优化方向。随着机器翻译本身质量的不断提高，一些旧的评测项目逐渐退出历史舞台，新的评测任务不断出现。这些新任务反映出研究者的兴趣所在，或者是实际应用需求，引导着研究者去进行相关的研究。可以说，机器翻译评测极大地促进了机器翻译相关研究的发展。

5 总结

在机器翻译研究领域，技术评测不仅推动了领域向前发展，同时还对研究发展和技术进步起着非常重要的引导作用。本文介绍了机器翻译评测的基本原理和几种常见的自动评测方法，希望有助于读者较为全面地了解国内外机器翻译评测的基本方法和最新进展，也希望我国的机器翻译事业能蒸蒸日上、更上一层楼。

参考文献

[1] Simon Zwarts. Machine Translation Evaluation. http://www.ics.mq.edu.au/~szwarts/MT-Evaluation.php

[2] Final report on workshop 1999 at Johns Hopkins University. http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps

[3] 刘群. 机器翻译评测综述. http://www.ict.ac.cn/diffusive/channel/detail2532.asp

[4] 俞士汶. 计算语言学概论. 280:285

[5] Kishore Papineni, Salim Roukos, Todd Ward and Wei-jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp311-318

[6] NIST Report (2002) Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

[7] Joseph P. Turian, Luke Shen, and I. Dan Melamed, Evaluation of Machine Translation and its Evaluation.

[8] Liang Zhou, Chin-Yew Lin and Eduard Hovy. Re-evaluating Machine Translation Results with Paraphrase Support. Proceeding of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006). 77:84

[9] 刘洋，刘群，林守勋.机器翻译评测中的模糊匹配. 中文信息学报. 2005. 第 19 卷第 3 期.45:53

[10] WordNet: http://wordnet.princeton.edu/

[11] Lavie, A., A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, To appear in Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)

[12] Banerjee, S. and A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)

[13] NIST Machine Evaluation: http://www.nist.gov/speech/tests/mt/

[14] TC-Star Machine Evaluation: http://www.tc-star.org/

[15] 863 评测: http://www.863data.org.cn/index.php

[16] IWSLT Machine Evaluation: www.slt.atr.jp/IWSLT2006/

[17] http://www.statmt.org/

[18] http://www.isi.edu/publications/licensed-sw/pharaoh/

[19] David Chiang. A hierarchical phrase-based model for statistical machine translation. 2005. In Proceedings of ACL. 263:270

[20] Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966.

[21] 钱跃良，刘群，林守勋，褚诚缘. 863 计划中文信息处理与智能人机接口技术评测综述

[22] 刘群.基于模板的统计翻译模型研究及汉英机器翻译系统实现. 北京大学计算语言学研究所博士论文开题报告

[23] 张霄军，陈小荷. NIST2005 机器翻译评测(MT-05)简评.南京师范大学文学院学报. 2006 年第 3 期.166:168

作者简介：

黄 瑾： 中科院计算技术研究所，硕士研究生

刘 洋： 中科院计算技术研究所，博士研究生

刘 群： 中科院计算技术研究所，博士，研究员