关于1995年度机器翻译评测的总结报告

段慧明　俞士汶

关键词：机器翻译评测、英汉翻译、汉英翻译、日汉翻译

一. 准备工作与出题指导思想

1995年1月7日，在863智能机专家组成员高文教授与专家组办公室的主持下召开了智能接口评测工作会议。与会人员一致表示愿意为提高国家科委与863组织的全国智能接口评测的权威性、知名度与规范化而努力。会议指定俞士汶担任自然语言处理评测组组长。会后由组长约请了国内中文信息处理学界知名学者担任自然语言处理评测组成员，经专家组办公室批准。

由于1995年度自然语言处理评测包括了一些从未评测过的项目，如中文文本切词、标注、自动文摘，而机器翻译也只在 1994年进行过一次。为探索这个领域的评测方法与评测标准，在专家组办公室的支持下， 3月25日在北京召开了一次专家研讨会，会议确定了专家的分工。

本文只对机器翻译的评测作一个总结。关于自动文摘评测的总结，另行撰文。关于自动切分与词性标注评测的总结，则由刘开瑛教授和冯志伟教授撰写。

会后，北大计算语言学研究所提出了“1995年863智能接口评测关于机器翻译评测大纲、评测方法及评测数据的说明”，制订了英汉、汉英、日汉三套机器翻译测试大纲，并于5月份将上述文件提交给专家组办公室。按照机器翻译测试大纲，北大计算语言学研究所为英汉、汉英、日汉机器翻译各出了一套测试题，每套题均包含源语言的400个句子，并提供了参考译文。

我们出题是在以下基本原则指导下进行的。

⑴ 评测的成绩显然与题目的难易有关。为了测试机译系统的质量，对作为源语言的英语和日语，基本上以大学教学大纲为水平测试的主要依据，对作为源语言的汉语则以常用句型为出题的主要根据。

⑵ 词汇选自一般领域，侧重在社会、生活、政治、经济、常识等方面，不出冷僻的词，也不出专业性很强的术语。

⑶ 少量固定词组也是常用的，不选俗语、谚语。

⑷ 选现代的书面的规范的英语、日语和汉语句子。

⑸ 以句子为测试单位。尽管理论上早已阐明，要提高机器翻译译文的质量，上下文信息的利用是必要的。但当前实用机译系统仍是以句子作为处理单位。因此，以句子作为测试单位也是合适的。

⑹ 出题考虑了机器翻译的特点。例如：下面的英语句子

You will work in this office under Mr. Pitt .

对于人来说，很容易确定“under Mr. Pitt”是修饰谓语动词的状语，而不是修饰office的定语。但对于机译系统来说，正确分析这样的歧义结构就不容易。事实上，参加评测的四个英汉系统有的译对了，有的就未译对。

我们以为按照这样的原则出的三套题大体上能测出机译系统的质量。当然实际评分还受其它因素影响。这样的题目与实际文本还有较大的出入，可以将这样的题目译得很好的系统处理真实文本并不一定能让用户满意。不过，反过来，如果一个系统自我评价其对真实文本有很高的译准率，但这样的题目却考得很糟，那它的自我评价是难以让人信服的。

二. 测试概况

参加1995年度机器翻译评测的有四个英汉系统,两个汉英系统,一个日汉系统。测试分现场测试与专家评分两个阶段进行。

2.1 现场测试

现场测试于1995年12月27日进行。用一个上午的时间对七个系统完成了从装系统、装题目到出结果的全部测试过程。大体上说还是顺利的，这说明评测大纲等材料的发布是起了作用的，出题者与参测者之间基本上是沟通的。但也产生了一些技术性的问题。从出题者角度检讨，由于题库先是在Windows环境中编辑的，而有的机译系统只能在DOS环境中运行，出题者需要将题库在DOS环境中重新整理，结果有些编辑标志不相容，造成机译系统处理失败。幸好出题者在现场进行了处理，这些问题都解决了。今后可相应多准备几套题目。从参测者的角度检讨，有些系统的适应性可能要差一些。因为，对同一套测试文件，有的机译系统可以顺利通过，有的系统则遇到了麻烦。要求出题者作适当的修改。现场测试时，除记录了每个系统翻译一套题目所需的时间外,也观察了包括界面、适应性等在内的运行情况。

2.2 专家评分

2.2.1 准备工作

应邀担任机器翻译与自动文摘的阅卷工作的几位专家都是国内自然语言处理领域的知名学者，并且都懂英语与日语两门外语。

北大计算语言所将三种机器翻译的结果按语种将源语言句子、参考译文、机器译文（英汉4个单位，汉英2个单位，日汉1个单位）合并成三个文件。在这些文件中，只注明“单位1，单位2，…”，没有具体给出单位的名称。

评分前，为每位专家准备了以下文件：英汉机译译文质量评分标准，英语原文、参考译文及四个单位的测试结果，汉语原文、参考译文及两个单位的测试结果，日语原文、参考译文及测试结果。

2.2.2 阅卷情况

实践表明评阅工作量非常大。考虑到英汉有4个单位参加，汉英有2个单位参加，日汉只有1个单位参加，专家们经过磋商，决定英汉抽出200道题进行评审，汉英只评150道题，日汉只评100道题。为了评分标准的掌握尽可能一致，四位专家参照英汉机译译文质量评分标准(限于篇幅，这里不能详细介绍这个标准,拟另撰文介绍)，对于英汉翻译，对挑选出来的单位1 的前10题分别进行了试验性的评分，结果列表如下：

	题号	专家1	专家2	专家3	专家4	一致性
	1	B	B	B	B	4
	2	E	E	D	D	2－2
单	3	B	C	B	C	2－2
	4	A	A	A	A	4
位	5	A	A	A	A	4
	6	B	B	C	C	2－2
1	7	A	B	A	A	1－3
	8	A	A	A	A	4
	9	D	E	E	E	1－3
	10	B	A	B	A	2－2

从上表可以看出：4位专家评卷结果的一致性是相当好的。10道题中4人一致的有四道题，3人一致的有2道题，其余4道皆是两两一致，且不一致的评分等级也只相差一级。

在上述工作基础上，4位专家都对评分标准的掌握有了信心，然后各人分别对4套英汉200道题、2套汉英150道题、1套日汉100道题的译文独自进行评分，最后汇总成评分表。人们习惯于用百分制给出成绩，也为了便于直观比较，4位专家针对今年试题及译文的实际情况，决定了各个等级的权值：

A＝95分 ，B＝80分，C＝65分，D＝45分，E＝20分，F＝0分

这些权值对英汉、汉英、日汉都是一样的。经加权平均，计算出各个系统的百分制成绩为：

	英　　汉				汉　英		日汉
	单位1	单位2	单位3	单位4	单位1	单位2	单位1
译文成绩	72.53	60.57	53.83	71.90	57.89	53.55	60.84

若将运行情况（包括翻译速度）的成绩考虑在内，一种可供参考的综合成绩（即译文质量占90％，运行情况占10％）如下：

	英　　汉				汉　英		日汉	百 分 比 率
	单位1	单位2	单位3	单位4	单位1	单位2	单位1	
运行成绩	80	80	60	100	100	80	80	10％
译文成绩	72.53	60.57	53.83	71.90	57.89	53.55	60.84	90％
总成绩	73.28	62.51	54.45	74.71	62.10	56.20	62.77	

三. 机器翻译的进展

1995年参加评测的系统有七个，比1994年增加了近一倍。其中五个系统都是首次参加。语种也增加了日汉。随着社会信息化的进展，对机器翻译的需求越来越强烈。国内机器翻译研究与实用系统的开发方兴未艾。实际上，参加评测的与国内已有的相比较，所占比例很小。希望今后有更多的系统参加评测，这样会更有利于机器翻译的发展。

中国的机器翻译系统都是在微机上开发的，这与微机的发展与中国的国情都有关系。这也是中国开发的机器翻译系统的优势之一。

考虑到汉语的特点，即汉语没有丰富的形态变化，机器分析比较困难，专家们预测汉外翻译比外汉翻译更难。现在大量的实践证实了这种预测。应该说，国内已有若干个英汉机译系统已做得相当好，达到或接近实用程度。参加评测的汉英系统要能实用，尚需付出艰苦的努力。

现在实用机译系统的开发已不再只追求译文质量的局部提高，更多地把注意力放在使用环境、用户界面的改进上。现在的系统一般都比从前的好用。用户也在探求如何利用机译系统以提高语言信息处理全过程的效率。

四. 评测工作小结

与国内通常采用的鉴定会形式的评测相比较，863智能机专家组组织的机器翻译评测，具有客观性与公正性的显著特点。国际上虽然存在多种形式的机器翻译评价，特别重视用户效益和用户意见的调查，但是将若干个同一类型的系统集中在一起进行公平的“竞赛”，也只在中国才有这样的举措。

与94年相比较，评测工作的公正性做得更好。在正式测试前，题库做到了严格保密。评分的专家是临时聘请的。参加评测者完全不知道哪些专家参加阅卷评分。由于测试结果文件上只使用参加评测单位的临时代号，专家也不知道哪些单位参加了评测，更不知道某个代号指的是哪个单位。这样专家们在评分时完全排除了“先入为主”等主观因素、感情色彩的影响，只对译文质量进行公正的评判。除了专家们事先认同的“绝对”标准外，由于同一道题的不同系统的译文都并列在一起，也能一目了然地看出不同系统的“相对”差异，这样也有助于专家运用“绝对”标准。

与94年相比较，评测的内容也有了发展。原先打算评测译文质量、速度、界面三项内容。由于现场测试那天专家们均未到场（这样做对保证公正性是有好处的），对界面不易给分，改为“运行情况”与“译文质量”两项，运行情况成绩占10％，译文质量成绩占90％。机译系统的使用环境与用户界面确实很重要，应该作为评测的内容。但它们的评测如何模型化、规范化、定量化，还是一个需要探索的课题。

现在评测结果出来了,这个结果应该有一定的参考价值，特别是同类型的机译系统之间，大致上反映了客观情况。但也不能绝对地对待这个结果。作为源语言的英语句子与汉语句子的难易很难比较。前面已经说了，95年的评测内容有所发展，但毕竟还是很有限的。不同系统的规模很不一样，适用领域也有所不同，这些都不能用一个百分制的成绩衡量出来。

这项评测工作是有意义的。我们希望今后有更多的系统参加评测，特别希望一些有影响的系统能够参加评测。这样也有利于评测研究的发展，可以不断提高863专家组组织的评测工作的权威性。

随着参测系统的增多和测试题目的增加，评测工作量将是人力所不能胜任的。因此寄希望自动评测的实现。北大计算语言学研究所在机器翻译译文质量自动评测方面已经做了多年的工作，建立了原型系统，题库正在扩大，向实用化方向发展。我们衷心希望这项工作能继续得到支持，实现规范化，那么在今后的大规模评测中，它作为一种辅助评测的自动化工具，就可以发挥重要作用。

参加阅卷的专家为这次评测作出了重要贡献，北大计算语言学研究所还有一些同事做了工作，恕不能一一公开致谢。

本文发表于《计算机世界》报1996年3月25日评测版，P183，题目“机器翻译评测报告”，内容略有改动。