机器翻译测评结果的一致性

曹冬林　李堂秋　史晓东　蔡经球
厦门大学计算机科学系　厦门　361005
E-mail:another@jingxian.xmu.edu.cn

摘要：机器翻译是人工智能领域中一项有挑战性的研究课题，而对机器翻译系统的测评也越来越受到重视。但机器翻译测评往往忽略了测评结果的一致性。为了定量地分析测评结果的一致性，本文提出了采用计算数据前后的变化趋势的方法，对测评结果进行一致性分析，从而说明测评结果的客观性和公正性。

关键词：一致性，等级差异值，句子长度差异值，句子难度值

The Consistency of Evaluation Result

CAO Dong-lin　LI Tang-qiu　SHI Xiao-dong　CAI Jing-qiu
Computer Science Department of Xiamen University　Xiamen　361005

Abstract: Machine Translation is one of the most challenging problems in artificial intelligence and recently its evaluation gets more and more attention. But most evaluation of machine translation has lost sight of the consistency of evaluation result. In order to give a quantitative analysis to the consistency of evaluation result, this paper presents a method of calculating the direction of data change. We have used it to analyse the consistency of our evaluation result and prove the objectivity and notarization.

Keywords: consistency, grade difference value, sentence length difference value, sentence difficulty value

一、引言

机器翻译的研究与开发开始于 1946 年，与计算机同龄。我国机器翻译的研究开始于 1956 年，经过近 50 年的发展，现在一大批机器翻译系统如雨后春笋般涌现出来，有的已经在翻译质量上达到了一定的水平。当然在肯定机器翻译取得长足进步的同时，我们还应看到在现今的机器翻译系统中还有许多的不足，需要进一步地深入研究。因此，为了明确机器翻译中的问题所在，对机器翻译进行测试和评估就非常重要。

目前世界上的机器翻译评估方法大致可分为三类[1~8]：

第一类为操作性评估(Operational Evaluation)，有时也称作经济评估(Economic Evaluation)。这种评估所关心的是机译系统的经济价值。

第二类为说明性评估(Declarative Evaluation)，又称质量评估(Qualitative Evaluation)。这种评估侧重通过测评译文质量评价各机译系统的性能。

本文得到国家 863 计划项目（2001AA114110），福建省科技计划重点项目（2001H023）支持.

第三类常用的评估方法为分类评估法（Typological Evaluation）。这种方法依据语言现象测试机译系统在该语言现象上的处理能力。

目前所有的机器翻译测评几乎都是以人工测评为主，自动测评也有北大机器翻译译文质量自动评估系统和NIST（National Institute of Standards and Technology）自动N元同现评分技术[9,10]。

在进行人工测评时，虽然测试的试题由专家选定，并且测评者都为专家，但由于人的思维有个渐进的过程，一开始对测评标准把握不是很准确，但随着测评的进行，后面的测评结果会比前面的相对客观、准确。因此，在人工测评中如何能说明测评结果前后是一致的是测评结果客观性、公正性的一个重要的衡量依据。而这项工作在大部分的机器翻译系统的测评中都没有进行。因此，为了说明测评结果的一致性，本文提出一个分析方案，对机器翻译的测评结果进行一致性分析。

二、测评结果的一致性分析方案

在以往的机器翻译测评中，测评人员都为专家，认为专家的意见是客观的、公正的，而忽略了人的主观因素，没有对测评结果进行一致性分析，这样就不能证明测评的标准在人的主观因素影响下前后有没有变化。更为重要的是，若测评人员不是权威的专家，而是一般的用户或技术人员，测评结果的客观性、公证性就要大打折扣。若能证明测评结果的一致性，虽然得出的绝对结果不如专家客观、公正，但是在相对结果上，在几个系统的相对比较上，能够说明测评过程的标准把握是一致的，从而该结果是客观的、公正的。

为了检验测评结果的一致性，我们对测评的数据结果进行一致性分析。首先将数据结果分成若干部分，然后计算这几个部分之间的数据的变化趋势是否符合最终的测评结果所显示的变化规律。若符合，则说明这几个部分的数据是一致的，否则不一致。按照这个思想，我们设计了如下的分析方案：

1. 将测评结果分成A、B两个测试集。
2. 统计A、B两个测试集的各个测评等级的百分比情况。
3. 计算两个测试集的差异值。
4. 依据测试集难度值和各个等级百分比估算测试结果变化是否合理。

我们曾对六个英汉机器翻译系统进行了测评，主要是针对翻译质量。测评中我们综合了以往的译文忠实度和可懂度两个等级划分标准。定义了五个译文质量等级：

（A）译文忠实反应原文内容，意义明确，用字准确，几乎不需修改。

（B）译文忠实反应原文内容，意义明确，但在用字、语法上有小毛病。

（C）基本上反应原文内容，但在某些地方较难理解，需要查看原文。

（D）少数语句可以读懂，但总体意义错误。

（E）意义完全错误，与原文完全不通。

测评等级的样例如下：

A form is the result of combining a form template with data.

【质量 A】一种形式是把一个形式模板与数据相结合的结果。

【质量 B】形式产生于把形式样板与数据相结合。

【质量 C】一个形式是结果组合一种形式模型和数据。

【质量 D】形式是联合一种形式有数据的样板的结果。

【质量 E】安培模型是用～结合一个模型模板数据的结果。

在测评过程中我们得出了一些测评数据结果。我们使用上述方案对我们的机器翻译测评结果进行一致性分析，考察测评结果的客观性、公正性。

三、测评结果的一致性分析

1. 测试集的划分

在划分测试集时，我们考虑到两点：一、我们是从前到后对句子进行测评打分，而当人经过前面一段测评的熟悉之后，后面的句子给分相对比前面一致；二、测试集不易过小，过小则数据分布不合理，偶然性很大。所以，我们从测评结果中取前 750 个句子作为测试集 A，取后 740 个句子作为测试集 B（注：两个测试集不相交）。

2 .等级百分比

测试集 A、B 的统计结果如下：

A	系统 1	系统 2	系统 3	系统 4	系统 5	系统 6
等级 A	33.6%	17.47%	25.07%	43.2%	14.4%	23.73%
等级 B	37.33%	25.6%	30.13%	33.07%	16%	30.53%
等级 C	20.13%	27.47%	25.2%	13.73%	23.87%	21.34%
等级 D	6.67%	20%	13.87%	6.27%	22%	16.27%
等级 E	2.27%	9.46%	5.73%	3.73%	23.73%	8.13%

B	系统 1	系统 2	系统 3	系统 4	系统 5	系统
等级 A	42.03%	20.95%	30.41%	51.76%	17.97%	32.4
等级 B	29.86%	28.65%	30.54%	24.86%	24.86%	26.4
等级 C	17.03%	21.35%	21.35%	12.3%	20.27%	19.4
等级 D	7.84%	20.27%	12.3%	8.38%	20.14%	13.3
等级 E	3.24%	8.78%	5.4%	2.7%	16.76%	8.2

两个测试集的句子长度分布情况如下：

	1~9（单词数）	10~19（单词数）	20~29（单词数）	30~49（单词数）
测试集 A	62.93%	29.6%	5.6%	1.87%
测试集 B	72.3%	23.11%	3.1%	1.49%

如果我们将长度为 1 到 9 的句子认为是一般简单句，长度为 10 到 19 的句子认为是复杂简单句，长度为 20 到 29 的句子认为是一般复合句，长度为 30 到 49 的认为是复杂复合句。则我们从表中可以看到测试集 B 的简单句比例比测试集 A 上升，其中一般简单句上升了 9.37

%，而复合句比例则在下降。简单句的分析正确率是高于复合句的，因此，我们看到测试集B中等级A所占的百分比有所上升。

3．测试集差异值

定义1：设测试集A中等级i所占的百分比为a_i，测试集B中等级i所占的百分比为b_i，等级个数为n，则两个测试集之间的等级差异值D计算公式为$D=\frac{\sum_{i=1}^{n}|a_i-b_i|}{n}$

通过计算我们得出六个系统的等级差异值情况如下：

系统1	系统2	系统3	系统4	系统5	系统6
4.23%	2.72%	2.3%	4.27%	4.98%	3.52%

定义2：设测试集A中长度段i所占的百分比为a_i，测试集B中长度段i所占的百分比为b_i，长度段个数为n，则两个测试集之间的句子长度差异值L计算公式为$L=\frac{\sum_{i=1}^{n}|a_i-b_i|}{n}$

通过计算我们得出句子长度差异值为4.68%。我们看到六个系统的测评数据等级差异值D中，只有一个在句子长度差异值L之上，并且偏差为0.3%，因此，我们的等级变化接近或小于句子长度变化，我们的数据变化范围是合理的。

4．变化趋势

数据变化范围合理，我们还不能确定测评结果的一致性，我们还需要通过两个测试集计算出数据变化趋势是否合理，计算变化趋势必须定义句子的难度，在这里我们把句子长度作为难度的判断标准。于是有如下的定义：

定义3：设句子长度段个数为m，句子等级个数为n，在句子长度段i下等级j所占的百分比为$t_{i,j}$，其中$\sum_{j=1}^{n}t_{i,j}=1$；设在测试集X中的长度段i所占的百分比为X_i，则测试集X中等级j的难度值$C_j=\sum_{i=1}^{m}(t_{i,j}\times X_i)$，相应的句子难度百分比$CP_j=\frac{C_j}{\sum_{j=1}^{n}C_j}$。由此我们定义两个测试集X、Y在等级j下句子变化值$M_j=\frac{X(CP_j)}{Y(CP_j)}$，若$M_j$>1，则认为是增长；若$M_j$=1，则认为是不变；若$M_j$<1，则认为是减少。其中$X(CP_j)$为测试集X在等级j下的难度百分比，$Y(CP_j)$为测试集Y在等级j下的难度百分比。

在计算之前我们需要的六个系统的句子长度与等级分布情况如下：(我们给出系统1和系

统 2 的相关数据表)

系统 1	1～9（单词数）	10～19（单词数）	20～29（单词数）	30～49（单词数）
等级 A	44.19%	27.23%	12.31%	12%
等级 B	32.57%	36.13%	38.46%	24%
等级 C	13.9%	26.21%	36.92%	40%
等级 D	6.46%	7.89%	10.77%	20%
等级 E	2.88%	2.54%	1.54%	4%

系统 2	1～9（单词数）	10～19（单词数）	20～29（单词数）	30～49（单词数）
等级 A	25.52%	7.12%	1.54%	0%
等级 B	29.59%	25.7%	6.15%	4%
等级 C	21.45%	32.57%	24.61%	16%
等级 D	15.49%	25.2%	53.85%	40%
等级 E	7.95%	9.41%	13.85%	40%

有了上面的数据，则由定义 3 我们计算出的变化趋势如下：

	系统 1	系统 2	系统 3	系统 4	系统 5	系统 6
等级 A	升	升	升	升	升	升
等级 B	降	升	降	降	升	降
等级 C	降	降	降	降	降	降
等级 D	降	降	降	降	降	降
等级 E	升	降	降	降	降	降

比较 2 中的测试集 A、B 的等级分布情况，我们看到在升降趋势上有以下几个与我们的计算值不同。

	测试集 A	测试集 B	差值（变化趋势）
系统 1	等级 D（6.67%）	等级 D（7.84%）	1.17%（升）
系统 2	等级 D（20%）	等级 D（20.27%）	0.27%（升）
系统 3	等级 B（30.13%）	等级 B（30.54%）	0.41%（升）
系统 4	等级 D（6.27%）	等级 D（8.38%）	2.11%（升）
系统 6	等级 E（8.13%）	等级 E（8.24%）	0.11%（升）

从该表的数据可以看到系统 2、系统 3、系统 6 的上升变化不足 0.5%，而系统 1 和系统 4 的变化略大，分别为 1.17%和 2.11%。考虑到人的主观因素在内，前后两个测试集必然会出现与我们预计的变化趋势不同的发展趋势。这些异常变化在 2.11%之内，我们认为这种异常变化是可接受的，总体的变化趋势是和预计的相同。

5 .分析结果

由上面的分析我们看到，在测试集 B 相对于测试集 A 有所变化的情况下，测试集的等级

差异值 D 小于或接近测试集句子长度差异值 L，测试集的各等级平均变化量是符合要求的。在这前提下，我们计算的变化趋势也大体符合我们的统计数据。因此，测试集 B 和测试集 A 的测评结果是一致的。

四、结论

机器翻译测评对机器翻译的系统开发来说是非常有意义的事情，可以及时地指出系统存在的问题。因此为了保证测评结果的客观性、公正性，必须对测评结果进行一致性分析。本文在原先的机器翻译测评数据基础之上，采用计算不同数据集之间的变化趋势是否合理的方法进行一致性分析，试验结果表明，我们的测评数据是比较客观的，我们的分析方法是可行的。

参考文献

【1】罗爱荣，段慧明 《机译评估方法评述及一个基于测试集的自动评估系统—MTE 的进展》《计算语言学进展与应用》1995 年

【2】段慧明，俞士汶 《关于 1995 年度机器翻译评测的总结报告》 《计算机世界》报 1996 年 3 月 25 日评测版

【3】俞士汶，姜 新，朱学锋 《机器翻译译文质量评价的实践与分析》中文电脑国际会议 ICCC '94(新加坡)论文集，PP26~32

【4】晶合实验室 《五颜六色——国产词典翻译软件之横向对比评测》 《大众软件》2001 年第 07 期

【5】徐剑，梁茂成 《对几种英汉机器翻译系统的测评》 《语言文字应用》1999 年第 2 期，PP97~102

【6】Eduard Hovy, Maghi King, Andrei Popescu-Belis 《An Introduction to MT Evaluation》 Workshop at the LREC 2002 Conference 2002,27(5):1-7

【7】Marianne Dabbadie, Anthony Hartley 《A Hands-On Study of the Reliability and Coherence of Evaluation Metrics》 Workshop at the LREC 2002 Conference 2002,27(5):8-16

【8】Andrei Popescu-Belis, Margaret King, Houcine Benantar 《Towards a corpus of corrected human translations》 Workshop at the LREC 2002 Conference 2002,27(5):17-21

【9】《The NIST 2002 Machine Translation Evaluation Plan (MT-02)》 http://www.nist.gov/speech/tests/mt

【10】《Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics》 http://www.nist.gov/speech/tests/mt