文章编号:1003 - 0077(2006)增刊 - 0007 - 12

2005 年度 863 机器翻译评测方法研究与实施

侯宏旭[1,3,4], 刘 群[1], 张玉洁[2], 井佐原均[2]

(1. 中国科学院计算技术研究所,北京 100080; 2. 日本情报通信研究机构,日本 京都 184 - 8795;
3. 内蒙古大学,内蒙古 呼和浩特 010021;4. 中国科学院研究生院,北京 100080)

摘要:为了能够全面了解国内外机器翻译技术的现状,促进机器翻译技术的研究,2005 年度 863 计划机器翻译评测于 2005 年 9 月举行。本次评测进行了汉英、英汉、汉日、日汉、日英、英日 6 个语言方向,两种类型的评测以及汉英词语对齐的评测。本次评测采用了网上评测的形式,利用基于 N-gram 的 NIST、BLEU 以及人工评测方法对各系统的结果进行评测。本文给出了此次评测的组织、准备、过程、结果及分析。为国内外研究单位在机器翻译方面的进一步研究提供了数据。

关键词:机器翻译;评测
中图分类号:TP391 **文献标识码**:A

Research and Implement of the 2005 HTRDP(863) Evaluation on Machine Translation

HOU Hong-xu[1,3,4], LIU Qun[1], ZHANG Yu-jie[2], ISAHARA Hitoshi[2]

(1. Institute of Computing Technology, Chinese Academy of Sciences Beijing 100080, China;
2. National Institute of Information and Communications Technology, Tokyo 184 - 8795, Japan;
3. Inner Mongolia University, Huhhot 010021, China;
4. Graduated University of Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The 2005 HTRDP(863) evaluation on machine translation was held on September 2005. The objective of this evaluation was to gain a comprehensive acquaintance with the status quo of machine translation technology at home and abroad, and to promote the research. There were 7 tracks in this evaluation: Chinese-English MT, English-Chinese MT, Chinese-Japanese MT, Japanese-Chinese MT, Japanese-English MT, English-Japanese MT and Chinese-English word alignment. The organization of the evaluation was web-based. In machine translation tasks, performance were measured using both human assessments and automatic metrics(BLEU and NIST based on N-gram). This paper will introduce the organization, preparation, process, results and analysis about this evaluation. It will also provide related information for further research on machine translation for domestic and foreign research groups.

Key words: machines translation; evaluation

1 概述

2005 年度 863 计划中文信息处理与智能人机接口技术机器翻译评测于 2005 年 9 月 20 日至 22 日举行。本次评测采用了网上评测的方式,组织单位通过网络发布评测数据,参评单位

收稿日期:2005 - 11 - 03 **定稿日期**:2006 - 01 - 13
基金项目:国家 863 计划资助项目(2004AA114010;2003AA111010)
作者简介:侯宏旭(1972—),男,博士研究生,副教授,主要研究方向为机器翻译.

通过网络在制定期限内传回结果。参加本次评测的有13个单位(国内11家,国外2家),评测有汉英、汉日、英汉、日汉、日英、英日共6个方向,对话和篇章两种类型。

2005年10月30日完成了翻译结果的自动和人工评测。

2 参评单位和语种

本次评测共有13个参评单位参加,其中国内11家,国外2家。它们是北京迈创易达科技发展有限责任公司、北京工业大学、北京赛迪翻译技术有限公司、中国科学院自动化研究所、哈尔滨工业大学语言技术研究中心、哈尔滨工业大学机器智能与翻译研究室、中国科学院计算机语言信息工程研究中心、国防科技大学、中国软件与技术服务股份有限公司、厦门大学、南京大学计算机系、Fujitsu Laboratories Ltd、Sharp Corporation。13个单位共提供了报名参加的25个系统(按语言方向分)。其中,汉英8个、英汉6个、汉日3个、日汉3个、日英2个、英日3个。另外,还包含词语对齐子项1个,参加单位2个。

表1 参评单位及参加的项目

单位	汉英	英汉	汉日	日汉	日英	英日	词语对齐
单位1	●	●	●	●			
单位2	●	●	●				
单位3	●	●	○*	○*	○*	○*	●
单位4	●	●					
单位5	●						●
单位6	●						
单位7	●						
单位8	●						
单位9		●					
单位10		●					
单位11				●			
单位12					●	●	
单位13						●	

*报名但没有提交结果

3 评测过程

3.1 数据准备

本次评测的测试语料包含6个语言方向(汉英、汉日、英汉、日汉、英日、日英)、2种类型(对话、篇章),其中对话语料为奥运相关领域(体育、交通、旅游、餐饮、天气等),篇章语料为新闻领域。根据国外相关评测及具体分析,我们制订了相应的语料规模。

考虑到在一个系统中同时处理汉、英、法、日三种语言,我们在评测中采用了Unicode(UTF16, little endian)作为标准编码方式,评测大纲中要求系统的输入和输出都采用这种编码,这有效地解决了编码格式不统一而导致的需要做编码转换的工作。针对新的编码方案,我们对相应的评测程序进行了改造。

(1)测试语料

本次评测的数据分为对话和篇章两种,分别涉及奥运相关和新闻领域,汉英、汉日方向采用基本相同的汉语语料,日英、日汉方向采用基本相同的日语语料,英汉、英日方向采用基本相同的英语语料。语料的规模见下表(其中单词和字符数是粗略统计的数目)。在评测中,实际发送给被测单位的语料大约是实际测试集规模的5倍,包括测试集及干扰集。干扰集的结果在评判时被舍弃。

其中,汉语和日语是字符数,英语是单词数。

测试语料采自网站、书籍、教材。其中汉语和英语语料由中科院计算所制作,日语语料由日本情报通信研究机构(NICT)制作。

经过纠错、格式转换后,混入了大约4倍的干扰集数据后形成若干XML文本文件。本次测试的语料文件一律采用utf-16(little endian)编码以适应多语言编码的需要。

表2 测试语料规模

语种		合计	
		句子数	字符/单词数
汉语	对话	约460	约9400
	篇章	约490	约21000
英语	对话	约450	约4700
	篇章	约490	约12000
日语	对话	约460	约11000
	篇章	约490	约21000

(2)参考译文

本次评测的每个句子都包含4个参考译文。制作参考译文时选择了4个以目标语言为母语的翻译者,他们各自独立地进行翻译。

其中,汉日、英日、日英方向的参考译文由日本情报通信研究机构(NICT)制作。

3.2 测试流程

本次863评测采用了网上评测的方式,根据时间安排,9月20日发布测试集,9月22结束。各评测单位共提交结果23个(含词语对齐的2个结果,另有4个结果未提交)

3.3 结果评测

各系统翻译的结果首先需要转换为评测软件能够处理的内部格式。

现场评测结束后开始进行自动和人工评测。其中人工评测在10月21~22日进行,自动评测的结果到10月30日全部完成。

其中汉日、英日、日英方向的人工评测在日本情报通信研究机构(NICT)进行。

4 评测结果

4.1 汉英方向

(1)对话

表3 汉英对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.1392	0.2506	0.7158	0.6192	0.4843	0.4091	65.38	64.25
系统2	6.2097	0.1747	0.6677	0.6717	0.5351	0.3357	57.42	52.49
系统3	5.7794	0.1524	0.6277	0.6942	0.5602	0.3197	51.56	47.06
系统4	5.8981	0.1544	0.6472	0.6881	0.5485	0.3155	56.96	53.72
系统5	5.5226	0.1454	0.5795	0.7357	0.6078	0.3509	53.41	51.59
系统6	5.9216	0.1814	0.6478	0.7134	0.5514	0.3518	50.42	57.16
系统7	6.0509	0.1714	0.6161	0.7175	0.5813	0.3589	55.58	55.02
系统8	4.2273	0.0710	0.5179	0.7683	0.6437	0.2024	39.74	33.02

(2)篇章

表4 汉英篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	6.9015	0.1843	0.7053	0.7228	0.5337	0.2343	61.72	55.90
系统2	6.2120	0.1361	0.6452	0.7560	0.5727	0.2090	53.97	47.28
系统3	5.3211	0.1073	0.5946	0.7860	0.6121	0.1743	43.90	38.72
系统4	5.9200	0.1287	0.6645	0.7612	0.5702	0.1851	52.81	46.97
系统5	4.9876	0.0718	0.5268	0.8412	0.6729	0.1863	41.23	32.30
系统6	5.7906	0.1188	0.6463	0.8307	0.5936	0.2087	37.33	39.33
系统7	5.5238	0.1056	0.5745	0.8077	0.6297	0.1926	40.65	36.08
系统8	4.1341	0.0550	0.4944	0.8385	0.6946	0.1292	36.52	30.31

4.2 英汉方向

(1)对话

表5 英汉对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.6444	0.3506	0.7302	0.5631	0.4261	0.4581	78.01	71.61
系统2	6.6385	0.2657	0.6917	0.6129	0.4644	0.3690	70.41	64.47
系统3	7.0142	0.2958	0.7096	0.5914	0.4535	0.4123	73.55	67.36
系统4	7.8703	0.3776	0.7470	0.5321	0.4156	0.4677	82.59	78.24
系统9	5.6119	0.2063	0.5972	0.6795	0.5651	0.2563	74.64	69.17
系统10	6.8419	0.2913	0.7135	0.5853	0.4529	0.3912	73.62	68.16

(2)篇章

表6 英汉篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	8.4334	0.3447	0.7537	0.6544	0.4170	0.3051	51.24	43.57
系统2	8.2600	0.3246	0.7629	0.6519	0.4191	0.2834	51.22	42.47
系统3	7.7755	0.2876	0.7333	0.6840	0.4435	0.2632	47.05	37.95
系统4	8.7453	0.3709	0.7930	0.6162	0.3934	0.3137	55.78	47.85
系统9	5.8304	0.1804	0.6205	0.7523	0.5581	0.1267	43.16	33.90
系统10	6.6745	0.2281	0.6998	0.7236	0.4946	0.1959	41.16	31.45

4.3 汉日方向

(1)对话

表7 汉日对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.1158	0.3512	0.7792	0.6483	0.4421	0.3197	53.44	44.87
系统2	6.9879	0.3069	0.7637	0.7071	0.4771	0.2782	47.56	37.28
系统3	未提交							

(2)篇章

表8 汉日篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.6785	0.3036	0.7851	0.6904	0.4363	0.2321	42.37	33.02
系统2	8.5858	0.3750	0.8265	0.6450	0.3886	0.2788	44.74	35.29
系统3	未提交							

4.4 日汉方向

(1)对话

表9 日汉对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.7098	0.3302	0.7302	0.6030	0.4430	0.4767	67.94	67.03
系统3	未提交							
系统11	6.3052	0.2292	0.6656	0.6626	0.5019	0.3781	58.44	56.88

(2)篇章

表10 日汉篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统1	7.9797	0.3007	0.7170	0.6748	0.4636	0.3249	50.41	44.58
系统3	未提交							
系统11	6.7836	0.2277	0.6862	0.7066	0.4969	0.2591	43.84	37.00

4.5 日英方向

(1)对话

表 11 日英对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统 3	未提交							
系统 12	5.3656	0.1529	0.5878	0.7392	0.5983	0.3495	65.81	52.77

(2)篇章

表 12 日英篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统 3	未提交							
系统 12	5.5193	0.1309	0.6139	0.8213	0.5984	0.2295	55.58	38.09

4.6 英日方向

(1)对话

表 13 英日对话评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统 3	未提交							
系统 12	8.0045	0.4875	0.7934	0.5320	0.3818	0.4314	63.31	46.69
系统 13	7.1239	0.3915	0.7663	0.5995	0.4377	0.3562	63.80	46.81

(2)篇章

表 14 英日篇章评测结果

ID	NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
系统 3	未提交							
系统 12	9.1112	0.4581	0.8167	0.6406	0.3766	0.3071	45.70	29.25
系统 13	8.6910	0.4464	0.8223	0.6463	0.3938	0.2831	44.93	27.66

4.7 词语对齐(汉英)

表 15 词语对齐评测结果

ID	准确率	召回率	F1 值	对齐错误率
系统 3	0.4993	0.5186	0.5088	0.4918
系统 5	0.8087	0.7220	0.7629	0.2348

5 评测方法与分析

在评测方法方面,采用以人工评测为主的方法。人工评测方法采用了国际上常用的忠实度和流利度为指标。同时,除采用了大纲中规定的 NIST 和 BLEU 方法外,我们还给出了国外评测中常用的评测指标 GTM、mWER、mPER 的结果,这些评测指标可以作为结果的参考。另外,我们也开发了自己的自动评测方法 ICT。从这些数据上,我们可以更全面的评估翻译结果的好坏,有助于参评单位找到参评系统的优缺点。

词语对齐的评测采用了准确率、召回率、*F*1 值及对齐错误率。

5.1 人工评测方法

(1)打分标准

本次评测的人工评测指标采用了忠实度和流利度两个指标。下表为打分档次的参考标准。

表 16　人工评测的忠实度打分标准

等级分	得分标准
0	完全没有译出来
1	译文只有个别词符合原文
2	译文有少数内容符合原文
3	译文基本表达了原文的意思
4	译文表达了原文的绝大部分信息
5	译文准确完整地表达了原文信息

表 17　人工评测的流利度打分标准

等级分	得分标准
0	完全不可理解
1	译文晦涩难懂
2	译文很不流畅
3	译文基本流畅
4	译文流畅,但是在地道性方面有所不足
5	译文是流畅而且地道的句子

表格给出了 5 个打分档次的译文忠实度和流利度水平,最低为 0 分,最高为 5 分,具体实施的时候打分可以包含一位小数。事实上,不同的评判者对以上标准的理解和具体处理会有一定的差距,所以做不同评判者之间的横向比较(评判者间、语言方向间)是不合适的,但是,这不影响人工评判结果的纵向比较(参评单位间)。

人工评测选择 4 个以目标语言为母语的评价者,分别对人工评测的句子打分。最后对所有打分计算平均值,然后乘以 20,即为总的得分。

忠实度 = 所有句子忠实度得分之和/总句数/5 × 100

流利度 = 所有句子流利度得分之和/总句数/5 × 100

(2)人工评测的实施

本次评测选取了每种语言方向和类型的 100 ~ 200 句测试语料进行人工评测,占语料的约 20 ~ 40%。根据评测结果可以看到,这样的人工评测规模基本满足了评测的要求。但是,如果从对机器翻译的技术推动方面,全部人工评测是一个完美的办法,不过,这样需要更多的时间和资金。

表 18　人工评测的任务量

语言方向	参评单位数目	数　量	完成时间(小时)
汉英	8	200	12
英汉	6	300	11
日汉	2	400	7

2004 年 10 月 20 ~ 22 日,汉英、日汉、英汉四个方向的人工评测在中国科学院计算技术研究所举行,同时,日本情报通信研究机构(*NICT*)也进行了汉日、英日、日英方向的人工评测。下表是具体的时间安排和实际耗时。由于每组评判者的熟练程度不同,这个时间仅作参考。

评判时,所有参评单位的结果作为原文的答案按照随机的顺序罗列出来,评判者对每个答案进行评分。

(3)人工评测软件

在中国科学院计算技术研究所进行的人工评测采用的方式和上年相同,将评测语料打印成试卷,由评判者在试卷上填写得分,然后录入成 excel 表格。

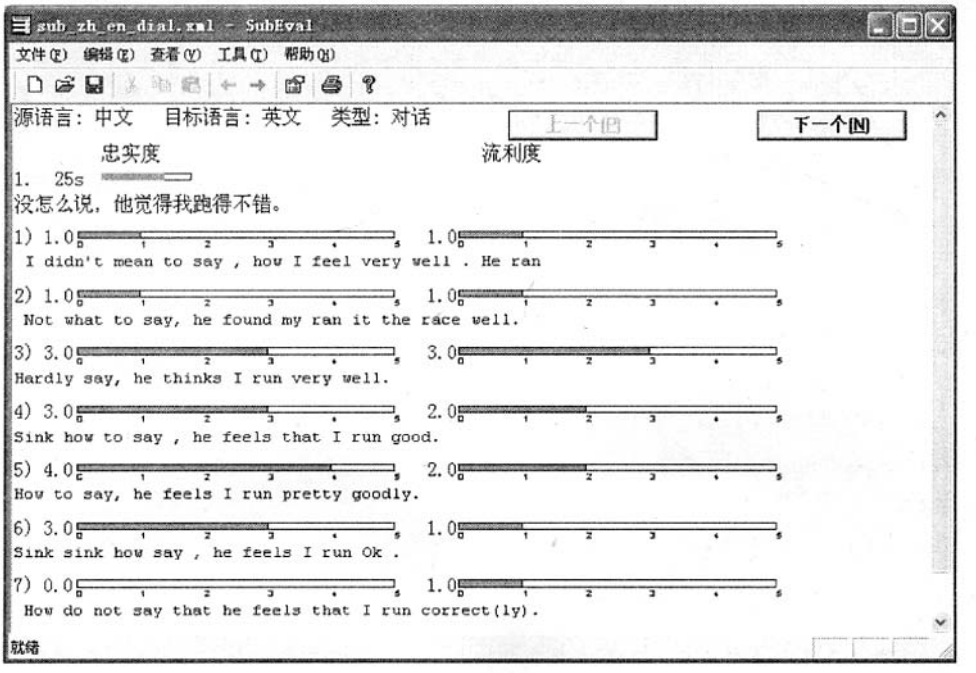

图1 人工评测程序

在 NICT,由于条件限制,不便将大量的语料全部打印。为此,我们编写了一个用于人工评测的软件。

5.2 自动评测方法

(1)方法

按照评测大纲,本次评测采用了 NIST 和 BLEU 两个指标,这两个指标是目前国际上机器翻译评测最常用的评测指标。除此之外,我们还给出了 GTM、mWER、mPER 得分,供参评单位参考。本次评测的自动评测方法都是大小写敏感的,这和往年的评测不同。

BLEU:基于 n 元语法准确率几何平均值的方法,得分越高越好。得分范围为[0,1]。

$$score = BP \times \exp(\sum_{n-1}^{N} w_n \log P_n), BP = \min\{1, \exp(1 - \frac{Lref}{Lsys})\}$$

其中,P_n 为被测译文中与参考答案匹配的 ngram 总数/被测译文中 ngram 总数,BP 为长度惩罚因子,$Lref$ 为被测句子长度最接近的答案长度,$Lsys$ 为被评测句子的长度,N 为最大 ngram 长度,W_n 为 ngram 的权重;

NIST:基于 n 元语法准确率算数平均值的方法,得分越高越好。得分总是大于 0 的,上限不定,大约在 12 至 14 左右。

$$score = \sum_{n=1}^{N} \{ \sum_{\substack{all\ w_1 \cdots w_n \\ that\ co-occur}} Info(w_1 \cdots w_n) / \sum_{\substack{all\ w_1 \cdots w_n \\ in\ sys\ output}} (1) \} \times \exp\{\beta \log^2 [in(\frac{Lsys}{Lref}, 1)]\}$$

$$Info(w_1 \cdots w_n) = \log_2 (\frac{number\ of\ occurrences\ of\ w_1 \cdots w_{n-1}}{number\ of\ occurrences\ of\ w_1 \cdots w_n})$$

β 是一个常数，是一个经验阈值，使得在 $Lsys/Lref = 2/3$ 时，β 使得长度罚分率为 0.5，$\bar{L}ref$ 是参考答案的平均长度，其实参数同 BLEU。

GTM：基于调和平均值的文本相似度方法，得分越高越好。范围在 0 ~ 1 之间。

$$score = \frac{2 \times Precision \times Recall}{Precision + Recall}, Precision = MMS/Lsys, Recall = MMS/\bar{L}ref$$

其中 *MMS* 为最大匹配长度。

mWER：基于编辑距离的最小单词错误率方法，得分越低越好。值大于 0，通常是小于 1 的。

$$score = \min_{all\ refs}\{edit\ distance/Lref\}$$

mPER：位置无关的 mWER 方法，得分越低越好。值大于 0，通常是小于 1 的。计算方法和 mWER 类似，但是并不考虑顺序。

ICT：有中国科学院计算技术研究所研制的以熵为基础的自动评测方法。它利用匹配片断计算加权熵。详情我们将另外发表论文加以介绍。

词语对齐的评测采用了准确率、召回率、*F*1 值及对齐错误率。

词语对齐任务的自动评测采用准确率和召回率，并根据准确率和召回率计算 *F*1 值和对齐错误率。

参评系统的运行结果给出的对齐结果集合记为 *A*。参考答案中的对齐结果分为确定的对齐结果，记为 *Gs*，和可能的对齐结果，记为 *Gp*。其中，*Gs* 是 *Gp* 的子集。

a）准确率

$$P = \frac{|\ AI\ Gp\ |}{|\ A\ |}$$

b）召回率

$$R = \frac{|\ AI\ Gs\ |}{|\ Gs\ |}$$

c）*F*1 值

$$F1 = \frac{2 \times P \times R}{P + R}$$

d）对齐错误率 *AER*

$$AER = 1 - \frac{|\ AI\ Gp\ | + |\ AE\ Gs\ |}{|\ A\ | + |\ Gs\ |}$$

（2）软件

本次自动评测采用的软件是基于 NIST 评测采用的 mteval-v11a. pl[1]（http://www. nist. gov/ speech/tests/mt/resources/scoring. htm），但是在这个基础上作了一些改进。

a）修改文件处理，以使其可以处理 Unicode 编码的输入文件；

b）增加了对中文、日文的处理。

计算 GTM 得分的程序来自 http://nlp. cs. nyu. edu/GTM/[2]，原来的程序是 java 编写的，我们将其移植到了 perl 上，并将其和 NIST 评测程序结合在了一起。

mWER 和 mPER 的算法也被加入到了评测程序中。

今年，我们将以上程序以及 ICT 评测指标加入了评测程序中，并将其改写成了 C ++ 程序。在开发集发布时，评测程序也同时发给了各参评单位。

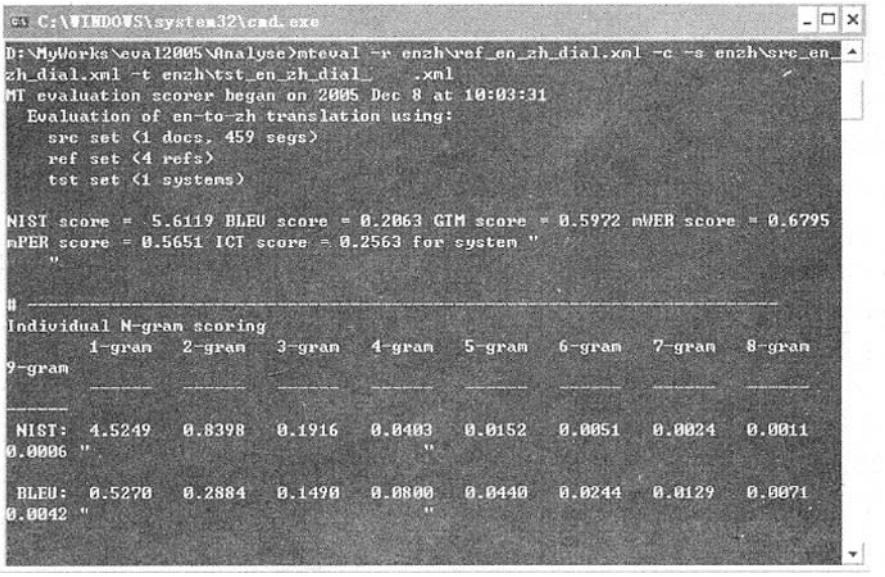

图2　自动评测程序

5.3　评测结果的对比和分析

(1)参评系统的种类和性能

在提交的21个系统中,RBMT或者RBMT结合EBMT的19个,EBMT的1个,SMT的1个。

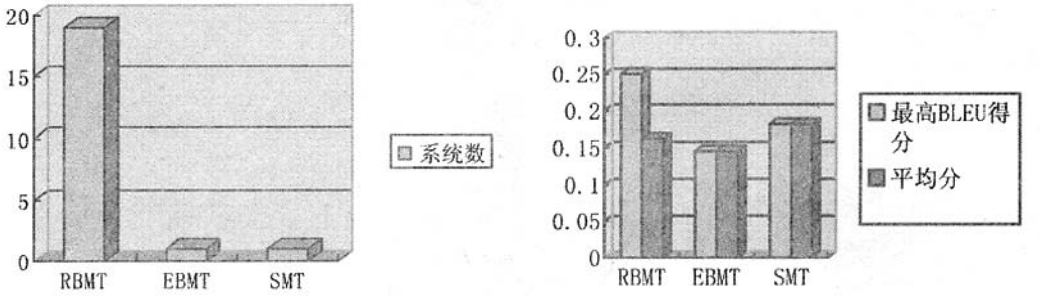

图3　汉英对话系统种类分析和得分比较

从这里我们可以看出,目前国内主要是RBMT为主流,这些RBMT方法的机器翻译系统经过几年的开发改进已经日趋成熟,并且步入了实用阶段。其中获得各项最高分的系统仍然是基于规则的系统,但是我们也可以看到,利用SMT或者EBMT的系统虽然少,但是仍然取得了不错的成绩。

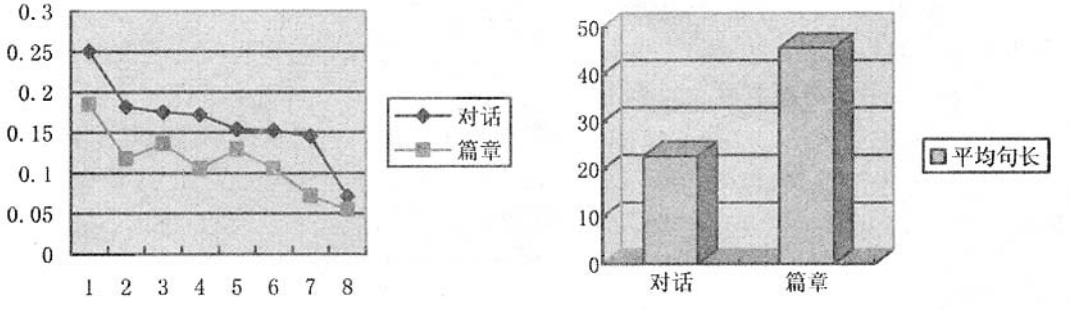

图4　汉英对话和篇章得分的比较

从近几年的国际评测的情况来看，参评系统都是以 SMT 为多，基本上能占到参评系统的 60% 以上，因此，国内对 SMT 的研究还有待深入。

(2)语料种类和得分的关系

根据各单位反馈的系统说明，大多数单位对于对话语料和篇章语料采用了相同的翻译引擎，因此各系统的对话得分和篇章得分的趋势基本趋于一致。

表 19　各自动指标和句子长度的相关度

NIST	BLEU	GTM	mWER	mPER	ICT	忠实度	流利度
-0.06	-0.148	0.031	0.0368	0.213	-0.4463	-0.316	-0.451

我们计算了汉英对话语料中句子长度与各评测指标的相关度，相关度通过计算两个序列的相关系数得到。

$$correl(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2 \sum_{i=1}^{n}(y-\bar{y})^2}}$$

从表格中可以看出，各指标同长度都没有太大的关系，其中 NIST、BLEU 和 GTM 基本和长度无关，而其它指标和长度负相关(长度越长评价越差)，尤其是主观评测中，句子的长度对人们评价句子的好坏产生了很大的影响。

(3)各自动评测指标之间的相互关系

表 20　各自动评测指标总体得分的相关度

	NIST	BLEU	GTM	mWER	mPER
BLEU	0.9818				
GTM	0.9576	0.9348			
mWER	-0.9203	-0.8898	-0.9517		
mPER	-0.9355	-0.9201	-0.9913	0.9734	
ICT	0.9275	0.9356	0.8123	-0.7288	-0.7633

表 21　各自动评测指标句子得分的相关度

	NIST	BLEU	GTM	mWER	mPER
BLEU	0.6248				
GTM	0.8858	0.4990			
mWER	-0.3534	-0.3979	-0.2019		
mPER	-0.5854	-0.4593	-0.4944	0.7765	
ICT	0.7590	0.6842	0.6756	-0.8135	-0.7985

这两个表给出了汉英对话语料各自动评测指标得分之间的关系。第一个表是总体得分的关系，从这个表可以看出，各个得分之间的相关对都接近于 1，即它们之间的相关度是相当高的，基本上通过一种得分就可以大致地评价系统的水平。第二个表给出了在句子级上各指标的相关度。基本上各个指标之间都保持一种较强的相关关系。其中 NIST 和 GTM、mWER 和 mPER 之间的相关度较高，体现了两种指标的算法之间的相似性。mWER 是单词错误率，mPER 是位置无关的单词错误率，两者之间相差的仅仅是位置关系。NIST 是基于最大 ngram 的算术平均值，GTM 是基于最大匹配长度的，两者的共同点就是它们都是基于最大匹配的长度，只不过是在具体权值和惩罚上有所不同。

(4)人工评测指标和自动评测指标之间的关系

表 22　人工评测指标和自动评测指标之间的关系

		NIST	BLEU	GTM	mWER	mPER	ICT
汉英对话	忠实度	0.9556	0.9092	0.8885	-0.8982	-0.8663	0.8814
	流利度	0.9420	0.9525	0.8787	-0.7725	-0.8346	0.9606
汉英篇章	忠实度	0.8280	0.8548	0.7753	-0.9589	-0.8230	0.6450
	流利度	0.9369	0.9688	0.9380	-0.9458	-0.9593	0.7658
英汉对话	忠实度	0.6477	0.7370	0.4397	-0.6240	-0.4283	0.5501
	流利度	0.6073	0.7021	0.4113	-0.6013	-0.3953	0.5057
英汉篇章	忠实度	0.9072	0.9327	0.8416	-0.9531	-0.8715	0.8665
	流利度	0.9048	0.9330	0.8309	-0.9468	-0.8655	0.8659

和国际上其它评测的情况一样，在汉英方向中，人工评测的两个指标中，NIST 指标和忠实度的相关度最高，BLEU 指标和流利度的相关度最高。而英汉对话、英汉篇章则不同，但 NIST 仍然是第二个同忠实度最相关的。注意到，在英汉对话中 ICT 评测指标和流利度的相关度是 6 个自动评测指标中最高的，这说明 ICT 指标在反映翻译结果的流利程度方面有较好的性能。其他语言方向因为数据量比较小，所以没有分析。从这里也可以看出，目前国际上普遍使用 NIST 和 BLEU 作为评价标准是合适的。需要注意的是，英汉语料的数据相对异常，相关度都比较低。其原因是，其中一家单位在给出翻译结果时采用了以下形式：

我的[虚弱;弱点]肯定是我的[翻转;转弯]。

在结果中列出了所有可能的翻译，这样导致自动评测程序的得分相对偏低，而人工评测的得分较高，影响了总体的相关度分析数据。

(5)人工评测两个指标的关系

和以往的 863 评测不同，考虑到和国际接轨以及更好的判断参评系统的好坏，本次评测采用了忠实度和流利度两个指标。

表 23　忠实度和流利度之间的相关度

	汉英	英汉
对　话	0.8956	0.9922
篇　章	0.9374	0.9982

从表格上看出，忠实度和流利度的相关性是比较高的，当然其原因主要是因为好的句子一般来说忠实度和流利度都是比较高的。相对来说，英汉方向的相关度要高于汉英方向。

表 24　句子级的忠实度和流利度之间的相关度

	汉英	英汉	汉日	日汉	日英	英日
对　话	0.8305	0.8888	0.9603	0.9151	0.9304	0.9393
篇　章	0.8039	0.9554	0.9402	0.9370	0.8816	0.9492

从句子级的相关度来看，六个语言方向的情况大体一致。除了汉英方向以外，其它各方向的流利度和忠实度的相关度都较高，可能的原始是，我们采用的人工评测方法是，每个题目由评分人员同时打出流利度和忠实度得分，这种方法可能造成这两个指标之间的独立性不够。

(6)大小写敏感对得分的影响

从图表上来看，大小写敏感对名次没有影响，各单位在大小写敏感或者不敏感的情况下的排名都完全相同。从 BLEU 得分上来看，大

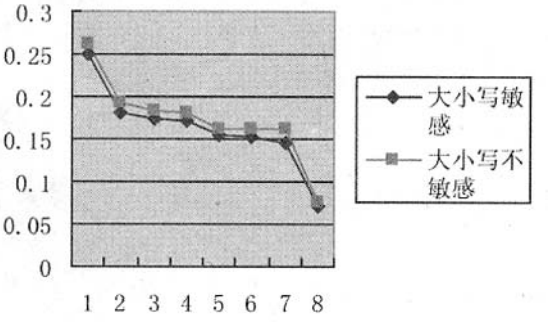

图 5　汉英对话项目大小写与 BLEU 得分关系

小写敏感要比不敏感的得分低0.01左右。

(7)按词和按字评测对得分的影响

考虑到分词可能产生的歧义,所以在863评测英汉、汉日、日汉、英日方向上,我们的评测程序完全是以"字"为单位的,因此,在匹配时匹配上的概率要更高一些。因此从理论上和结果上来看,这些方向的得分都要远高于汉英、日英的得分。为此,我们利用了中国科学院计算技术研究所开发的ICTCLAS对各单位产生的翻译结果和人工翻译的参考答案进行了分词,并根据分词的结果重新进行了打分。

表25 汉英对话和英汉对话BLEU得分的分布

	最低分	最高分	平均分
汉 英	0.071	0.2506	0.1627
英 汉	0.2063	0.3776	0.2979
英汉(分词后)	0.1079	0.2614	0.1903

表26 按字与按词的英汉对话得分的相关度

	NIST	BLEU	GTM	mWER	mPER	ICT
总 体	0.9981	0.9969	0.9952	0.9876	0.9953	0.9964
句子级	0.8665	0.75264	0.8259	0.8965	0.8512	0.9244

从这个结果上来看,经过分词以后,英汉的得分分布已经和汉英的得分分布接近。

无论从总体得分和句子级得分的相关度上来看,按词与按字的得分都具有较高的相关度。句子级BLEU的相关度较低主要是因为许多句子的BLEU得分为0,使得句子级的BLEU得分区分度不高。

6 趋势和建议

评测是推动智能信息处理技术发展的重要手段之一,而对于机器翻译来说,评测手段就更为重要。为了更好的掌握机器翻译的现状,更好的推动机器翻译技术的发展,我们需要做的还很多。

通过前面的分析我们可以看到,通过评测,确实可以达到推动国内外机器翻译技术进步的目的。通过这样的评测,通过吸引国内外单位参加评测一方面可以推动机器翻译技术的进展,另一方面,通过评测,我们可以获得许多往常无法获得的数据,通过这些数据我们可以分析总结机器翻译发展的方向、同时改进评测方法。

参 考 文 献:

[1] http://www.nist.gov/speech/tests/mt[EB].

[2] http://nlp.cs.nyu.edu/GTM/[EB].

[3] F. J. Och, Minimum error rate training in statistical machine translation[C]. In:Proc. of the 41st ACL, Sapporo, Japan, 2003, 160 - 167.

[4] F. J. Och, Statistical Machine Translation: From Single-Word Models to Alignment Templates[D], 38 - 39.

[5] Yasuhiro Akiba, etc. Overview of the IWSLT04 Evaluation Campaign[C]. 2004.