文章编号：1003-0077(2012)01-0022-09

第七届全国机器翻译研讨会机器翻译评测总结

赵红梅，吕雅娟，贲国生，黄 云，刘 群

（中国科学院 计算技术研究所 中国科学院 智能信息处理重点实验室，北京 100190）

摘 要：该文介绍了第七届全国机器翻译研讨会(CWMT2011)机器翻译评测的具体情况。本次评测重点关注各种语言到汉语的翻译，除了汉英、英汉、日汉三个语言对以外，评测还新增了五种民族语言(藏语、蒙古语、维吾尔语、哈萨克语、柯尔克孜语)到汉语的翻译评测。共有 19 家国内外单位的 165 个系统参加此次评测。除了介绍评测项目的设置、评测数据的准备、评测流程、参评单位等，本文还重点介绍了 CWMT2011 的评测结果，并对评测结果进行了分析，用实例说明了与评测结果相关的几个因素：源语言与目标语言是否相似、评测领域是否集中、测试集与训练及开发集语料是否相似、训练语料的规模、参评系统的技术和成熟度等。

关键词：机器翻译；机器翻译评测；BLEU-SBP；WoodPecker 评测

中图分类号：TP391 **文献标识码**：A

Summary on CWMT2011 MT Translation Evaluation

ZHAO Hongmei, LV Yajuan, BEN Guosheng, HUANG Yun, LIU Qun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract: The 7th China Workshop on Machine Translation(CWMT2011)Evaluation continues the ongoing series of evaluation of machine translation technology in China. This paper presents an overall introduction to CWMT2011 evaluation. This evaluation focuses on the evaluation of MT translation from other languages to Chinese, especially, from ethnic languages (including Mongolian, Tibetan, Uyghur, Kazakh and Kirghiz). 165 systems of 19 participants from home and aboard have taken part in the evaluation. The paper introduces the evaluation tasks, the evaluation data, the evaluation procedure and the participants. We also discuss the evaluation results in details. The examples from this evaluation show that the evaluation result depends on the following factors: the similarity between the source language and the target language, the range of the field which the evaluation task involves, the similarity between the test data and the training/development data, the size of the training data, the technology and the maturity of the participating system, and etc.

Key words: machine translation; machine translation evaluation; BLEU-SBP; WoodPecker evaluation

1 概述

中国中文信息学会主办的第七届全国机器翻译研讨会(CWMT2011)于 2011 年 9 月 23～24 日在厦门召开。为了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，按照惯例，本届机器翻译研讨会继续组织了统一的机器翻译评测，以推进参评单位的实质性交流和机器翻译技术的发展。

本次评测由中国科学院计算技术研究所组织，评测重点关注各种语言(包括我国蒙古族、藏族、维吾尔族、哈萨克族、柯尔克孜族的民族语言)到汉语的翻译，评测共包含 7 个语言对，9 个评测项目和 4 个评测领域(新闻、科技、政府文献和日常用语)。在

收稿日期：2011-11-01 定稿日期：2011-11-11

基金项目：国家自然科学基金项目(60873167)；国家青年科学基金项目(61100082)

作者简介：赵红梅(1968—)，女，学士，工程师，主要研究方向为机器翻译；吕雅娟(1972—)，女，博士，副研究员，主要研究方向为机器翻译和自然语言处理；贲国生(1986—)，男，硕士研究生，主要研究方向为机器翻译和自然语言处理。

汉英—英汉新闻方向的评测中，除了英汉新闻的当前(current)评测外，还设置了英汉和汉英新闻的进展(progress)评测。

本次评测采用以下流程：通过网络，评测组织方在评测前一个月向参评单位提供评测训练语料和开发语料，评测时再统一发放测试语料，参评单位在测试语料发放后约三天之内提交系统翻译结果，组织方对翻译结果进行统一测评后，向所有参评单位公布评测结果。

此次评测主要的自动评测指标为 BLEU-SBP (Chiang et al., 2008)，在汉英方向还采用了 WoodPecker 评测(Zhou et al., 2008)。

本次评测吸引了国内外 19 家教育科研机构和企业单位参加，在 9 个评测项目上共提交了 165 个系统的翻译结果。为了加强技术交流的效果，评测要求每个参评单位撰写一份评测技术报告。技术上有特点的参评单位在 CWMT2011 研讨会上就本单位的评测技术情况进行了口头报告。另外，研讨会还设置了评测的海报展示环节，每个参评单位都以海报的形式展示了自己参评系统的技术情况。评测组织方的总结报告(包括各参评单位主系统的系统描述)、各参评单位的技术报告以及研讨会录用的其它论文都被收录进研讨会的论文集并发放给大家。

本次评测在语料提供方面得到了新疆大学等多家单位(详细语料提供单位参见表 4.1 和表 4.2)的鼎立支持。多名业内专家在评测准备会上为 CWMT2011 评测提出了很多很好的设想和建议。

本文给出了此次评测的组织准备过程、评测结果和分析。文中将列出所有参评单位的名称，但在评测结果中，不会给出对应的单位名称，而是代之以单位的匿名代号。

2　评测项目

CWMT2011 评测项目的设置如表 2.1 所示。

表 2.1　CWMT2011 评测项目

项目代号	评测项目名称	语种	领域
ZH-EN-NEWS	汉英新闻领域机器翻译	汉语→英语	新闻领域
EN-ZH-NEWS	英汉新闻领域机器翻译	英语→汉语	新闻领域
EN-ZH-SCIE	英汉科技领域机器翻译	英语→汉语	科技领域
JP-ZH-NEWS	日汉新闻领域机器翻译	日语→汉语	新闻领域
MN-ZH-DAIL	蒙汉日常用语机器翻译	蒙古语(简称蒙语)→汉语	日常用语
TI-ZH-GOVE	藏汉政府文献机器翻译	藏语→汉语	政府文献
UY-ZH-NEWS	维汉新闻领域机器翻译	维吾尔语(简称维语)→汉语	新闻领域
KA-ZH-NEWS	哈汉新闻领域机器翻译	哈萨克语(简称哈语)→汉语	新闻领域
KI-ZH-NEWS	柯汉新闻领域机器翻译	柯尔克孜语(简称柯语)→汉语	新闻领域

本次评测共设置了 9 个评测项目，涉及到 7 个语言对，4 个评测领域(新闻、科技、政府文献和日常用语)。与往届评测不同的是，本次评测重点关注了各种语言到汉语的翻译，除了汉英、英汉、日汉三个曾经评测过的语言对以外，评测首次增加了民族语言(藏语、蒙语、维语、哈萨克语、柯尔克孜语)到汉语的翻译评测。在汉英—英汉新闻方向的评测中，除了英汉新闻的当前(current)评测外，还设置了英汉和汉英新闻的进展(progress)评测。

3　参评单位和系统

本次评测共有 19 个单位报名参加，其中国内单位 15 家，国外单位 4 家，教育和科研机构 16 家，企业单位 3 家。参评单位名单如下：

CNGL, School of Computing, Dublin City University

NTT Communication Science Laboratories

SYSTRAN Software, Inc.

北京航空航天大学计算机学院智能所
北京交通大学
东北大学自然语言处理实验室
富士通研究开发中心有限公司
哈尔滨工业大学机器智能与翻译研究室
内蒙古师范大学
南京大学
西安理工大学
厦门大学
新疆大学
中国科学技术信息研究所
中国科学院合肥物质科学研究院智能机械研究所
中国科学院计算技术研究所智能信息重点实验室
中国科学院软件研究所基础软件国家工程研究中心
中国科学院新疆理化技术研究所
中国科学院自动化研究所

19 家单位在 9 个不同的项目和语言方向共提交了 165 个系统的翻译结果。表 3.1 给出了本次评测每个项目的参评单位和系统的数量。

表 3.1 参评单位和系统数量

评测项目	参评单位/主系统	参评系统总数
汉英新闻领域机器翻译	11	28
英汉新闻领域机器翻译	10	24
英汉科技领域机器翻译	13	28
日汉新闻领域机器翻译	8	23
蒙汉日常用语机器翻译	4	10
藏汉政府文献机器翻译	6	20
维汉新闻领域机器翻译	6	16
哈汉新闻领域机器翻译	4	7
柯汉新闻领域机器翻译	4	9
合计	66	165

4 评测组织

4.1 评测方法

评测采用目前国际上普遍采用的评测方式：由评测的组织方提供训练和测试数据，参评单位在给定时间内返回翻译结果，再由评测组织方进行评价。

所有评测项目都是对译文质量进行评测，采用自动评测方法。主要评测指标为 BLEU-SBP (Chiang et al., 2008)，其他自动评测指标包括：BLEU、NIST、GTM、mWER、mPER、ICT，汉英方向还采用了 Woodpecker 评测 (Zhou et al., 2008)。自动评测的算法(包括 WoodPecker)都是大小写敏感的，中文的评测是基于字的，而不是基于词的。

对于每个评测项目，参评单位必须提交一个基本结果(Primary Result)，最多可以提交三个对比结果(Contrast Results)。产生基本结果的系统称为参评单位的基本系统或主系统(Primary System)，产生对比结果的系统称为参评单位的对比系统(Contrast System)。基本系统中，对于采用基于实例的机器翻译技术或者统计机器翻译技术实现的模块或系统，所使用的训练数据必须限制在评测组织方指定的数据范围之内，不允许使用任何外部数据；对于采用基于规则的机器翻译技术实现的模块或系统，允许采用通过人工方式构造的翻译知识(例如，规则、模板、词典等)，但是要在系统描述和技术报告中对于所使用的翻译知识的规模、构造和使用方式等进行说明。对比系统则可以使用任何数据进行训练。参评系统也可以采用系统融合技术，但要求在系统描述中进行明确说明，并在技术报告中给出系统融合前单系统的运行结果。评测组织方在发布评测结果时，也会对采用了系统融合技术的系统进行标注。

4.2 测试数据准备

本次机器翻译的评测语料涉及 8 个语言方向(汉英、英汉、日汉、蒙汉、藏汉、维汉、哈汉和柯汉)、4 个领域(新闻、科技、政府文献和日常用语)。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。在评测中输入输出文件均采用 UTF-8 编码(有 BOM)以及严格的 XML 格式。

训练语料中，英文单语语料为路透社的 RCV1 语料，汉语单语语料为搜狗实验室的搜狗全网新闻语料库 SogouCA，双语语料情况见表 4.1；开发和测试语料情况见表 4.2。

测试语料包括真实测试集及干扰集两部分，干扰集的结果在评判时被舍弃。

所有开发集和测试集均为一份原文、四份参考答案。每份参考答案的原始文本均由四名经验丰富的专业翻译人员各自独立翻译而成。

表 4.1　CWMT2011 机器翻译评测双语训练语料情况

评 测 项 目	语料规模	双语语料提供单位
ZH-EN-NEWS	584 万句对/1.05 亿英文单词/1.45 亿汉字	点通公司、中国科学院计算技术研究所、东北大学、北京大学、中国科学院自动化研究所、厦门大学、哈尔滨工业大学
EN-ZH-NEWS		
EN-ZH-SCIE	90 万句/23 728 117 英文单词	中国科学技术信息研究所
JP-ZH-NEWS	28 万句/5 661 592 日文字	南京大学、大连理工大学、北京大学
MN-ZH-DAIL	6.8 万句对/982 135 蒙语词	内蒙古大学
TI-ZH-GOVE	10 万句对/约 1 280 837 藏语词	青海师范大学、厦门大学、西北民族大学
UY-ZH-NEWS	5 万句对/1 091 903 维语词	新疆大学
KA-ZH-NEWS	5 万句对/965 570 哈语词	新疆大学
KI-ZH-NEWS	5 万句对/1 175 823 柯语词	新疆大学

表 4.2　CWMT2011 机器翻译评测开发集和测试集情况

评 测 项 目	开发集	真实测试集	语料提供单位
ZH-EN-NEWS	1 006 句对/42 562 汉字	progress：1 003 句对/39 828 汉字	中国科学院计算技术研究所
EN-ZH-NEWS	1 000 句对/21 790 英文单词	progress：1 002 句对/21 537 英文单词 current：1 001 句对/22 394 英文单词	中国科学院计算技术研究所
EN-ZH-SCIE	1 008 句对/21 044 英文单词	1 497 句对/56 182 英文单词	中国科学技术信息研究所
JP-ZH-NEWS	500 句对/28 316 日文字	500 句对/28 546 日文字	中国科学院自动化研究所
MN-ZH-DAIL	1 000 句对/8 590 蒙语词	500 句对/3 444 蒙语词	内蒙古大学、中国科学院计算技术研究所
TI-ZH-GOVE	650 句对/约 11 937 藏语词	658 句对/约 11 879 藏语词	青海师范大学（中国科学院计算技术研究所提供了藏语的切分版本）
UY-ZH-NEWS	700 句对/15 688 维语词	700 句对/15 350 维语词	新疆大学
KA-ZH-NEWS	700 句对/14 348 哈语词	700 句对/14 356 哈语词	新疆大学
KI-ZH-NEWS	700 句对/16 566 柯语词	700 句对/16 960 柯语词	新疆大学

为了了解各参评单位的系统进步情况，今年汉英新闻和英汉新闻方向均设置了进展（progress）项目（使用的是 CWMT2009 的评测语料），另外，英汉新闻方向还设置了当前（current）项目（使用的是 2011 年新制作的语料），我们在评测结果中对比了进展项目中两年评测的系统变化情况。另外，在汉英新闻方向，我们继续进行了 WoodPecker 评测，该评测全部采用 CWMT2009 中 WoodPecker 评测的测试数据和相关参数。有关 WoodPecker 评测的详细情况，请参见 CWMT2009 机器翻译评测报告，该评测报告以及计算所组织的历届全国机器翻译评测的相关资料可参考以下评测网页及相关链接：

http://nlp.ict.ac.cn/new/CWMT/index.php

本次评测中所有项目的参考译文均不提供给参评单位，而是留到下次评测时继续使用，以便了解各参评单位在这一段时间间隔内的技术进步。在参评单位提交评测结果之后、研讨会开始之前这段时间，我们向各参评单位开放了在线评测打分网站，供参评单位进行机器翻译实验时打分使用。

4.3　评测流程

本次 CWMT 评测采用了网上评测的方式，表 4.3 给出了此次评测的流程。

表 4.3　CWMT2011 评测流程

1	2011 年 6 月 26 日	报名截止日期
2	2011 年 7 月 1 日	评测组织方发放训练数据和开发数据

续表

3	2011年8月1日上午10:00	评测组织方发放汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译3个项目的测试数据
4	2011年8月2日上午10:00	评测组织方发放日汉新闻领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译3个项目的测试数据
5	2011年8月3日上午10:00	评测组织方发放维汉新闻领域机器翻译、哈汉新闻领域机器翻译、柯汉新闻领域机器翻译3个项目的测试数据
6	2011年8月4日下午5:30	参评单位提交汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译3个项目的翻译结果
7	2011年8月5日下午5:30	参评单位提交日汉新闻领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译3个项目的翻译结果
8	2011年8月6日下午5:30	参评单位提交维汉新闻领域机器翻译、哈汉新闻领域机器翻译、柯汉新闻领域机器翻译3个项目的翻译结果
9	2011年8月10日	评测组织方向汉英新闻领域机器翻译、英汉新闻领域机器翻译、英汉科技领域机器翻译3个项目的参评单位通知评测结果
10	2011年8月11日	评测组织方向日汉新闻领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译3个项目的参评单位通知评测结果
11	2011年8月12日	评测组织方向维汉新闻领域机器翻译、哈汉新闻领域机器翻译、柯汉新闻领域机器翻译3个项目的参评单位通知评测结果
12	2011年8月12日	在线评测平台开放
13	2011年8月19日	参评单位提交评测技术报告
14	2011年8月24日	评测组织方返回评测技术报告(供参评单位修改)
15	2011年9月1日	评测技术报告终稿
16	2011年9月23～24日	在研讨会上进行研讨 在线评测平台关闭

5 评测结果与分析

本节给出各评测项目主系统在主要评测指标BLEU-SBP上的评测结果，并对评测结果进行了分析。我们在BLEU-SBP的基础上，针对各主系统的翻译结果，进行了结果之间差异的显著性检验——符号检验(Collins et al.，2005)，总的做法是：分别以每个主系统为基准系统，测试了所有其他主系统与基准系统结果差异的显著性程度，以此构造了所有主系统翻译结果的差异显著性矩阵，因篇幅有限，本文仅显示汉英新闻进展(progress)评测的差异显著性结果(图1)。详细的评测结果参见CWMT2011机器翻译评测报告(http://nlp.ict.ac.cn/new/CWMT/CWMT2011.php)。

在下面评测结果的图表中，横坐标是该项目各参评单位提交的主系统(用各单位代号来表示)，纵坐标是主要评测指标的得分。横坐标中出现的“◆”代表其左边的系统采用了系统融合技术。

5.1 汉英新闻(progress)

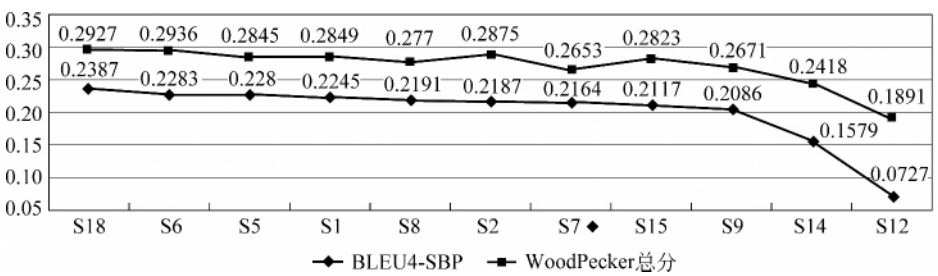

图1 CWMT2011汉英新闻(progress)评测结果

表 5.1.1　汉英新闻(progress)各主系统 2009 年与 2011 年评测结果对比

Site	BLEU4-SBP (2009)	BLEU4-SBP (2011)	WoodPecker (2009)	WoodPecker (2011)
S18	0.226 0	0.238 7 ↑	0.298 1	0.292 7 ↓
S6	0.222 3	0.228 3 ↑	0.293 4	0.293 6 ↑
S5	0.205 4	0.228 0 ↑	0.267 2	0.284 5 ↑
S1	0.226 1	0.224 5 ↓	0.277 0	0.284 9 ↑
S8	0.208 8	0.219 1 ↑	0.286 2	0.277 0 ↓
S7◆	0.240 7	0.216 4 ↓	0.277 3	0.265 3 ↓
S14	0.195 6	0.157 9 ↓	0.269 6	0.241 8 ↓
S12	0.154 1	0.072 7 ↓	0.225 6	0.189 1 ↓

注：↑表示得分提高，↓表示得分下降。

表 5.1.2　汉英新闻(progress)各主系统 BLEU4-SBP 差异显著性检验结果表

(显著标志●，不显著标志○，$p<0.05$)

	S18	S6	S5	S1	S8	S2	S7◆	S15	S9	S14	S12
S18	—	●	●	●	●	●	●	●	●	●	●
S6	●	—	○	●	●	●	●	●	●	●	●
S5	●	○	—	○	●	●	●	●	●	●	●
S1	●	●	○	—	○	○	○	●	●	●	●
S8	●	●	●	○	—	○	○	○	●	●	●
S2	●	●	●	○	○	—	○	●	●	●	●
S7◆	●	●	●	○	○	○	—	○	●	●	●
S15	●	●	●	●	○	●	○	—	○	●	●
S9	●	●	●	●	●	●	●	○	—	●	●
S14	●	●	●	●	●	●	●	●	●	—	●
S12	●	●	●	●	●	●	●	●	●	●	—

分析：从表 5.1.1 中可以看出，既参加了 CWMT2009 又参加了 CWMT2011 汉英新闻评测的单位中，有一半单位的成绩有所提高，例如 S5 的 BLEU-SBP 值提升了 2.3 个百分点，有一半单位因各种缘故成绩有所下降。总的来说，参加评测的汉英新闻评测系统的差异性不大，排在第 1 名和第 9 名的系统的 BLEU-SBP 值的差异只有 3 个百分点，很多系统间的差异性不显著(表 5.1.2)。从评测报告和评测结果来看，绝大多数参评单位采用的汉英机器翻译技术差异不大，各单位系统之间的差距在逐渐缩小。

5.2　英汉新闻

表 5.2.1　英汉新闻(progress)各主系统 2009 年与 2011 年评测结果对比

Site	BLEU5-SBP (2009)	BLEU5-SBP (2011)
S6	0.335 2	0.352 2 ↑
S7◆	0.356 3	0.342 9 ↓
S8	0.321 7	0.331 9 ↑
S18	0.313 8	0.327 0 ↑
S14	0.278 1	0.303 3 ↑

注：↑表示得分提高，↓表示得分下降。

分析：从表 5.2.1 可以看出，在进展(progress)项目中，与 2009 年相比，2011 年英汉新闻领域的 BLEU 值除个别单位(S7)略有下降外，其他四个单位均有提高。当前(current)项目的评测结果中，横坐标上，除了前两个单位和后两个单位外，中间单位的 BLEU 值差异基本上不明显(详情可参考 CWMT2011 评测报告中的显著性检验结果，如图 2、图 3 所示)。

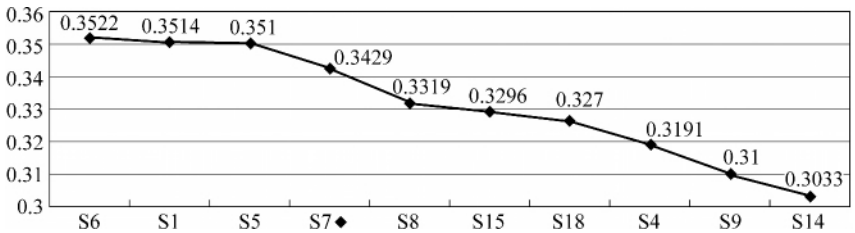

图 2　CWMT2011 英汉新闻(progress)评测结果(BLEU5-SBP)

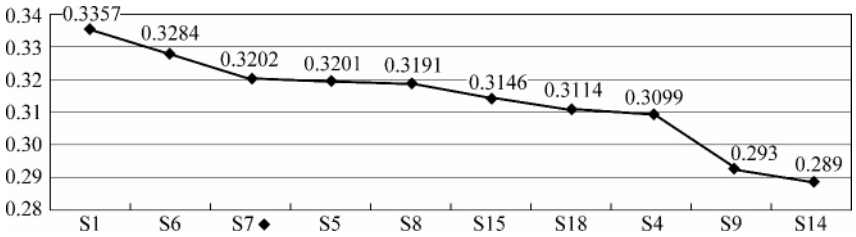

图 3　WMT2011 英汉新闻(current)评测结果(BLEU5-SBP)

5.3 英汉科技

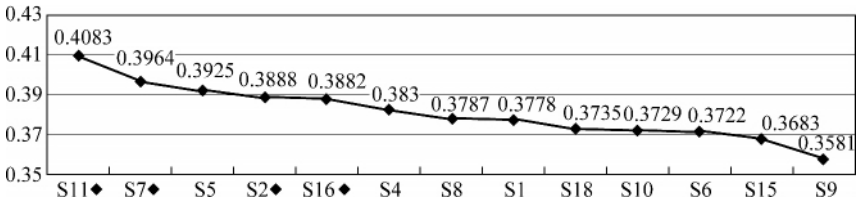

图 4 CWMT2011 英汉科技评测结果(BLEU5-SBP)

分析：从图 4 可以看出，英汉科技领域的 BLEU 值比较高，这可能与领域比较集中(主要集中在计算机和通讯两个领域)，测试集与开发集、训练集的语料内容比较一致，训练语料规模比较大(表 4.1)有关。

5.4 日汉新闻

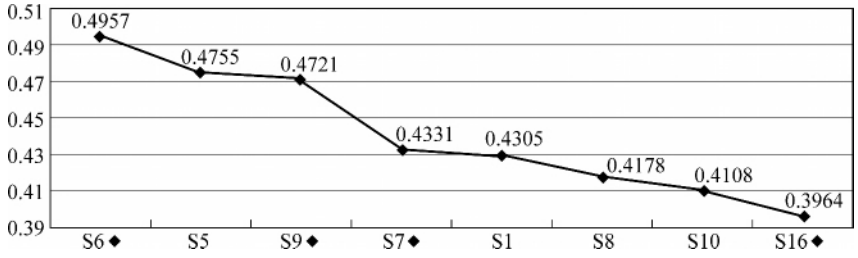

图 5 CWMT2011 日汉新闻评测结果(BLEU5-SBP)

分析：从上图可以看出，日汉新闻领域的 BLEU 值比较高，各系统间的差异比较显著。此次日汉新闻项目提供的训练语料规模比较大，但内容比较庞杂，与开发集和测试集语料(内容都集中在新闻领域)的相似度并不高，然而，笔者通过对比原文和参考译文发现：日文和中文这两种语言的相似程度非常高，这可能是该项目 BLEU 值较高的主要原因。

例如，原文：国家開発銀行が今回香港で発行したCDは、主に機関投資家を対象としたもので、個人投資家は購入できない。

参考译文之一：中国开发银行这次在香港发行的 CD，主要以集团投资家为对象，个人投资家不得购入。

5.5 藏汉政府文献和蒙汉日常用语

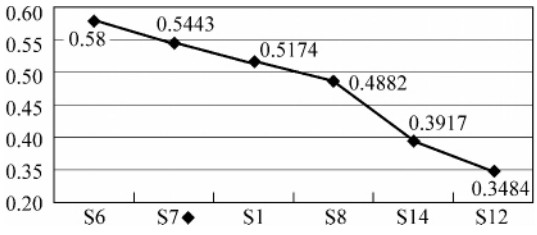

图 6 CWMT2011 藏汉政府文献评测结果(BLEU5-SBP)

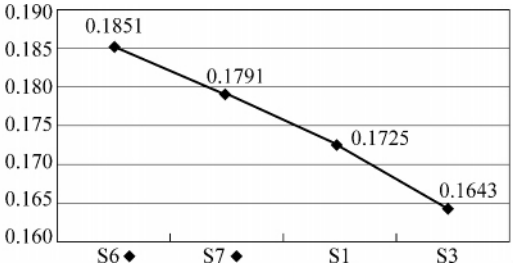

图 7 CWMT2011 蒙汉日常语评测结果(BLEU5-SBP)

分析：

从图 6 可以看出，藏汉政府文献领域的 BLEU 值很高，各系统间的差异很显著。我们分析 BLEU 值偏高的原因，发现：1)评测语料主要来源于政府文献，领域相对集中，固定表达多，且使用频率高；2)相对于其它民语来说，训练语料较多(表 4.1)。这两个原因有可能是藏汉政府文献 BLEU 值偏高的主要原因。

从图 7 可以看出，蒙汉日常用语领域的 BLEU 值很低，各系统间的差异不太显著。我们考察了一下各参评单位提交的翻译结果，从翻译质量上来看，各个系统还很不成熟，译文中漏译现象比较严重，命名实体普遍没有翻译出来，译文长度偏短。

5.6　维汉新闻

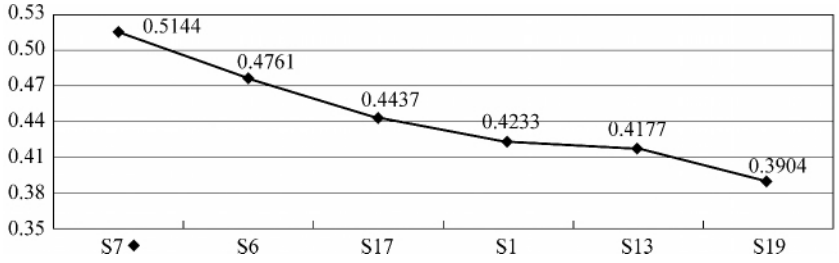

图 8　CWMT2011 维汉新闻评测结果(BLEU5-SBP)

分析：从图 8 可以看出，维汉新闻领域的 BLEU 值较高，各系统间的差异比较显著。我们考察了评测语料及翻译结果，发现和藏汉政府文献翻译的情形类似，维汉新闻的测试语料和训练语料主要来源于中国政府发布的官方新闻，领域比较集中，固定表达多且使用频率高，而且测试语料与训练语料相似程度高，从而导致系统译文的质量比较好，BLEU 值比较高。

5.7　哈汉和柯汉新闻

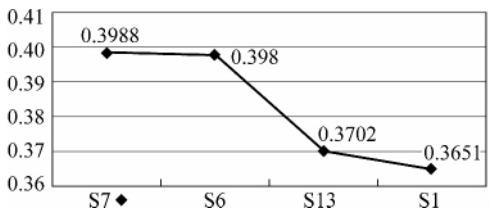

图 9　CWMT2011 哈汉新闻评测结果(BLEU5-SBP)

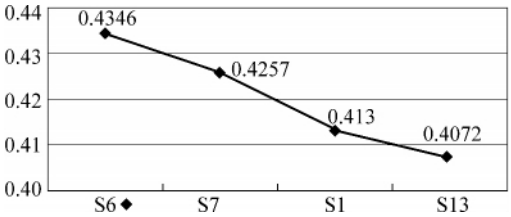

图 10　CWMT2011 柯汉新闻评测结果 BLEU5-SBP

分析：从图 9 和图 10 可以看出，哈汉和柯汉新闻领域的 BLEU 值也比较高，各系统间的差异不太显著。与维汉新闻翻译的情形类似，哈汉新闻和柯汉新闻的测试语料和训练语料主要来源于中国政府发布的官方新闻，而且测试语料与训练语料相似程度比较高，所以得分较高，而且柯汉新闻更有相当一部分(超过 1/3)的测试语料和训练语料来源于政府颁布的一些法规和条例，领域更加集中，这也许可以说明为什么柯汉新闻的翻译效果更好。

5.8　各评测项目情况对比及总体分析

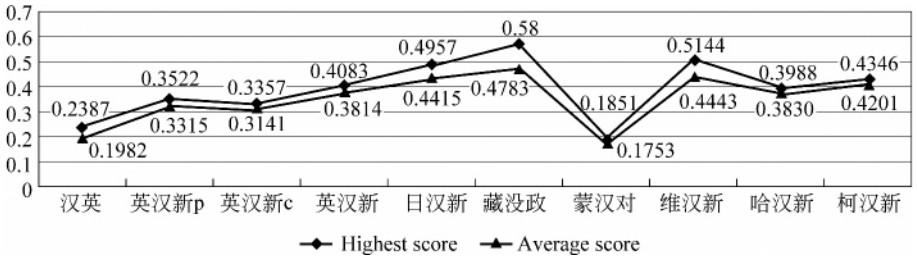

图 11　CWMT2011 各评测项目对比

图 10 横坐标中从左至右分别对应着表 2.1 中从上到下九个评测项目，其中英汉新闻包括两个子项目：英汉新 p 代表进展(progress)项目，英汉新 c 代表当前(current)项目。纵坐标为 BLEU-SBP 的数值，其中汉英的评测指标是 BLEU4-SBP，其他是 BLEU5-SBP。

总的来看，评测 BLEU-SBP 得分比较高的项目有：藏汉政府文献、维汉新闻、日汉新闻、柯汉新闻、英汉科技和哈汉新闻；评测得分比较低的项目有：汉英新闻和蒙汉新闻。

通过对参评系统所采用的技术以及翻译结果的分析，我们发现：

1）统计机器翻译技术在本次参评的系统中占主流地位。参评系统绝大多数采用了统计机器翻译技术（66 个主系统中有 62 个采用的是纯统计机器翻译技术）。与以往相比，本次评测更多单位采用了基于句法的统计机器翻译模型（包括基于形式句法的层次短语模型），这表明更多的单位掌握了这项技术。

2）规则和统计相结合的系统在评测中表现出一定的优势。参评系统中只有少量系统（4 个主系统）结合了规则式方法和统计式方法，但是均取得了不错的效果。例如，S18 采用的是在比较成熟的规则式系统的翻译结果上运用统计式方法进行后编辑，在汉英新闻的评测项目中其 BLEU 值排名第一；S11 采用的方法是，在基于统计和基于规则这两类机器翻译多引擎的翻译输出的基础上，进行系统融合，其在英汉科技领域提交的翻译结果 BLEU 值排名第一。（另外还有一个单纯的规则系统作为对比系统，在其所在的评测项目组中成绩不太理想。）

3）系统的翻译质量取决于多种因素。总体来说，源语言与目标语言相似程度越高（如日汉新闻），评测的领域越集中，测试语料与训练语料/开发语料的相似程度越高（如藏汉政府文献、维汉新闻、柯汉新闻、英汉科技等），训练语料规模越大（如英汉科技、藏汉政府文献），参评系统采用的技术越先进，参评系统的成熟度越好（包括对一些细节问题的处理，如：对评测语料的前期处理、对翻译结果的译后处理、对命名实体的处理，以及系统开发者的技术熟练程度等），系统表现越好。

6 总结

CWMT2011 评测主要侧重于其他语种到汉语的评测，共设立了 9 个评测项目，其中包括汉英双向的进展性评测，新增了五个语言对的评测，新增的评测主要是民族语言到汉语的评测。此次评测的评测项目和参评单位的数量都位居历届全国机器翻译评测之首。

从评测结果来看，虽然一些项目的语种和领域是参评单位从未接触过的，但不论是从自动评测的结果（BLEU 值等）还是从笔者人工考察的译文质量来看，机器翻译的效果都超过了我们的预期，这充分证明了统计式机器翻译技术强大的适应性。但是也存在着一些问题，例如，汉英新闻的翻译，通过进展性评测，我们发现结果喜忧参半（大约有一半单位成绩有所提高，还有大约一半的单位在后退或者止步不前），希望这个问题引起大家的重视。

评测的成绩取决于多种因素，包括源语言与目标语言的相似程度、评测领域的集中程度、测试语料与训练/开发语料的相似程度、训练语料的规模以及参评系统采用的技术和成熟度等。但是评测的结果不是我们评测的真正目的，我们的目的是通过评测这个手段，给大家提供一个技术交流的平台，让大家及时发现问题，跟踪最新的机器翻译技术，互帮互学，共同前进，推动我国机器翻译事业稳步向前发展，最终达到利益大众的目的。

CWMT2011 机器翻译评测能得以顺利进行，得益于各同行单位和同仁的大力支持，没有大家在评测语料上的无私奉献和在评测组织上的宝贵意见，这么大规模评测的实施是无法想象的。我们在此对所有为 CWMT2011 评测提供评测语料、参加 CWMT2011 评测、关心和支持 CWMT2011 评测的单位和同仁表示最诚挚的感谢！

CWMT2011 研讨会已初步确定了以后 CWMT 机器翻译评测每隔两年举行一次，下一次机器翻译评测的时间定在 2013 年，我们热忱地欢迎各同行单位届时踊跃报名参加评测，也期待着各参评单位在 CWMT2013 中再创佳绩，将我国机器翻译的研究和开发应用推向一个新的高潮！

参考文献

[1] 刘群，赵红梅. 第五届全国机器翻译研讨会（CWMT2009）评测报告[R]. 第五届全国机器翻译研讨会（CWMT2009），2009 年 10 月 16～17 日，南京.

[2] 赵红梅，吕雅娟，贲国生，等. 第七届全国机器翻译研讨会（CWMT2011）评测报告[R]. 第七届全国机器翻译研讨会（CWMT2011），2011 年 9 月 23～24 日，厦门.

[3] David Chiang, Steve DeNeefe, Yee Seng Chan, et al. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms[C]// Proc. EMNLP 2008, pages 610-619.

[4] Michael Collins, Philipp Koehn, Ivona Kučerová. 2005. Clause restructuring for statistical machine translation[C]//Proc. ACL 2005, pages 531-540.

[5] Ming Zhou, Bo Wang, Shujie Liu, et al. 2008. Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points[C]//Proc. Coling 2008, pages 1121-1128.