第十一届全国机器翻译研讨会（CWMT 2015）评测报告

汪昆[1]，姜文斌[2]，杨海彤[1]，向露[1]，赵红梅[2]

（1．中国科学院自动化研究所 模式识别国家重点实验室，北京 100190；
2．中国科学院计算技术研究所，北京 100190）

摘 要：本文详细介绍了 CWMT 2015 机器翻译评测的总体情况，CWMT 2015 机器翻译评测方案与上届评测（CWMT 2013）相比有较大变化：本次评测取消了上届评测中的灰箱测试、基线系统和人工评测，增加了双盲评测（Double Blind Evaluation）项目。在双盲评测中，评测组织方选择一个语言对进行加密编码，呈现给各个参评单位的是加密后的双语对照数据和句法树。这次的双盲测试最大限度地抛开了词法分析、句法分析、命名实体翻译等预处理和后处理对翻译的影响，最大限度地关注翻译模型本身的问题。此外，本文还详细描述了评测的项目设置、参评单位及评测项目、官方发布的评测数据、评测过程等等。最后给出了所有参评系统的匿名评测结果，并在附件中提供了各参评单位主系统的系统描述。

CWMT 2015 Machine Translation Evaluation Official Report

WANG Kun[1], JIANG Wenbin[2], YANG Haitong[1], XIANG LU[1], ZHAO Hongmei[2]
(1. National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China;
2. Institute of Computing, Chinese Academy of Sciences, Beijing, China)

Abstract: Compared with the past evaluations campaigns, CWMT 2015 MT evaluation is characterized as follows: first, adopting Double-blind Evaluation which can help us offer the utmost attention to translation model itself rather than the pre- and post-processing like lexical analysis, syntactic analysis and named entity translation for the first time, in which the encrypted bilingual data and syntactic tree of one language pair have been presented to the participants; second, cancelling Gray-box evaluation and manual evaluation adopted in the last evaluations campaign; third, the baseline system(s) and the corresponding Gray-box files haven't been provided. This report presents an overall introduction to CWMT 2015 MT evaluation, including a detailed description of the evaluation tasks, the participants and their registered tasks, the official release of the evaluation data, the procedure and most of all the evaluation's results. The detailed system descriptions of all primary systems are also attached.

1 概述

为了全面了解国内机器翻译技术的最新发展水平，促进机器翻译技术的交流和发展。按照惯例，由中国中文信息学会主办的第十一届全国机器翻译研讨会（CWMT 2015）继续组织了统一的机器翻译评测（简称 CWMT 2015 机器翻译评测），本次评测由中国科学院自动化研究所和中国科学院计算技术研究所联合组织。本次机器翻译评测与上届评测相比有较大的变化，主要体现在：

（1）评测取消了上届评测中设置的灰箱测试、基线系统、人工评测以及进展测试（Progress Test）。

（2）评测采用了“双盲评测（Double-Blind Evaluation）”方式，即评测组织方选择一个语言对进行加密编码，呈现给各个参评单位的将是加密后的双语对照数据和句法树。这次的双盲测试将最大限度地抛开词法分析、句法分析、命名实体翻译等预处理和后处理对翻译的影响，将最大限度地关注翻译模型本身的问题。

本次评测的项目设置详见表 1。评测共包含汉英新闻、英汉新闻、蒙汉日常用语、藏汉政府文献、维汉新闻以及双盲评测 6 个评测项目。所有评测项目仅包含一个测试集。

表1 CWMT 2015 评测项目表

评测项目名称	项目代号	语种
汉英新闻	CE	汉语→英语
英汉新闻	EC	英语→汉语
蒙汉日常用语	MC	蒙古语→汉语
藏汉政府文献	TC	藏语→汉语
维汉新闻	UC	维吾尔语→汉语
双盲	DB	******

本文给出了此次评测的组织准备过程、参评单位、评测方法和评测结果等，为了方便大家进行学术交流，本文还在附件中给出了各评测项目参评主系统的系统描述信息。希望本文能为国内外学者和单位研究机器翻译、了解 CWMT 2015 机器翻译评测提供些许帮助。

CWMT 评测是研究而非商业性质的评测，所有评测结果仅供研究使用， 评测成绩可以在研究论文中引用，但不可用于任何出于商业目的的宣传活动。参评单位或个人在研究论文中引用时，如果没有得到其它单位的许可，不得公开其它单位的评测结果。相应地，本评测报告虽然公布所有参评单位的名单，但在评测结果中，不会给出具体的单位名称，而是代之以参评单位的代号。

2 参评单位和系统

本次评测共有 14 家单位参加（含 1 家国外单位）。评测总共提交了 61 个系统翻译结果，包含 27 个主系统翻译结果和 34 个对比系统翻译结果。所谓主系统是指每个单位参加评测的基本系统，其采用的训练数据必须在评测组织方指定的范围内，而对比系统是指参评单位产生对比结果的系统，其采用的训练数据可以不受限制，其中，仅使用了评测组织方指定的训练数据、没有使用外部数据的系统又称为受限系统，使用了外部数据的对比系统称为非受限系统。

表 2 给出了本次评测的参评单位和参评系统的数量。其中，汉英新闻项目参评的单位最多，共有 7 家单位参加；三个民族语言评测项目也得到了大家的普遍关注，其中维汉项目的参评单位最多，有 6 家单位参加； 此次新引入的双盲评测也吸引了 6 家单位参加。表 3 给出了参评单位及参评项目的具体情况，其中“√”表示报名并提交了有效的机器翻译结果，“◇”表示报名但提交的机器翻译结果无效，这主要指：主系统的训练使用了外部数据，“○”表示报名但因故没有提交机器翻译结果。

表2 参评单位和参评系统数量统计表

评测 项目	参评 单位	系统 总数
CE	7	15
EC	6	7
MC	5	9
TC	5	9
UC	6	10
DB	6	11
合计	14	61

表 3 参评单位及参与项目情况

参评单位名称	评测项目					
	汉英	英汉	蒙汉	藏汉	维汉	双盲
北京交通大学计算机学院自然语言处理组	✓	✓				
北京航空航天大学计算机学院智能所	✓	✓	○	○	○	○
ADAPT/CNGL	✓	○				✓
哈尔滨工业大学机器智能与翻译实验室	✓	✓				✓
中国科学院合肥智能机械研究所			✓		✓	
内蒙古师范大学计算机与信息工程学院自然语言信息处理实验室			✓			
中国科学技术信息研究所			✓	✓	✓	
东北大学			◇	◇	◇	
南京大学	○	○				○
澳门大学自然语言处理与中葡机器翻译实验室	✓	✓				✓
西藏大学信息化研究所				✓		
中国科学院新疆理化技术研究所					✓	
新疆大学					✓	
厦门大学	✓			✓		✓

3 评测过程和方法

3.1 评测数据准备

本次机器翻译评测语料包含 5 个语言方向(汉英、英汉、蒙汉、维汉、藏汉） 和 4 个领域（新闻、科技、日常用语和政府文献)。根据国外相关评测及具体情况，我们确定了相应的语料规模。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。 评测的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 4 名经验丰富的专业翻译人员各自独立翻译而成。

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，我们向参评单位开放了在线评测打分网站：http://159.226.21.8/cwmt2015，供参评单位进行机器翻译实验打分使用。

今年我们的评测数据在上届评测数据的基础上，新增了一批训练语料，包括：

- 内蒙古大学汉蒙平行语料库(2015 版)(新增 2 万句对)
- 西北民族大学藏汉平行语料库（2015 版）（新增 2 万句对）
- 中国科学院新疆理化技术研究所维汉双语语料库（2015 版）(新增 3 万句对)
- 中国科学院自动化研究所英汉网络平行语料库（2015 版）(100 万句对)
- 中国科学院计算技术研究所英汉网络平行语料库（2015 全新版）(200 万句对)

具体评测数据情况请见表 4、表 5 和表 6。

3.2 双盲评测

本次双盲测试中，评测组织方选择一个语言对进行加密处理，并为大家提供加密后的双语对照数据、句法树和目标语言的单语语料。这次双盲测试最大限度地抛开了词法分析、句法分析、命名实体翻译等预处理和后处理对翻译的影响，而着重关注翻译模型本身的问题。希望通过这种全新的方式促进大家对机器翻译模型本身的关注，推动机器翻译的研究工作。

表4 CWMT 2015评测开发数据情况

评测项目	双语	双语语料提供单位	单语语料及提供单位
汉英新闻	778万句对	点通公司、 中科院计算所、东北大学、北京大学、中科院自动化所、厦门大学、哈尔滨工业大学	路透社 RCV1 语料， 路透社
英汉新闻			全网新闻语料库 SogouCA（ 2008 版和 2012 版），搜狗实验室
蒙汉日常用语	13.1万句对	内蒙古大学，中科院合肥智能所	
藏汉政府文献	12.7万句对	青海师范大学， 西北民族大学机器翻译研究所、厦门大学人工智能研究所、西藏大学信息化研究所	
维汉新闻领域	14万句对	新疆大学，中科院新疆理化所	
双盲测试	200万句对	中国科学院自动化研究所	中国科学院自动化研究所

表5 CWMT 2015评测开发数据情况

评测项目名称	规模(单位：句对)	提供单位
汉英新闻	1,000	中国科学院计算技术研究所
英汉新闻	1,000	中国科学院计算技术研究所
蒙汉日常用语	1,000	内蒙古大学
藏汉政府文献	1,000	青海师范大学
维汉新闻领域	1,000	新疆大学
双盲测试	1,000	中国科学院自动化研究所

表6 CWMT 2015评测测试数据情况

评测项目名称	规模(单位：句对)	提供单位
汉英新闻	1,100	中国科学院自动化研究所
英汉新闻	1,100	中国科学院自动化研究所
蒙汉日常用语	1,000	内蒙古大学
藏汉政府文献	729	青海师范大学
维汉新闻领域	1,000	中国科学院新疆理化技术研究所
双盲测试	1,000	中国科学院自动化研究所

3.3 评测方法

本次评测全部采用自动评测指标进行评价，即利用自动评价工具对参评单位提交的最终翻译结果文件进行评价。本次评测中的自动评测采用多种自动评价标准，包括：BLEU-SBP、BLEU-NIST、TER、METEOR、NIST 、GTM 、mWER 、mPER 以及 ICT，取消了 WoodPecker 评测指标。

评测组织方进行自动评价时将采用如下设置：

(1) 所有自动评测将采用大小写敏感（case-sensitive）的方式；

(2) BLEU-SBP 作为主要的自动评价指标；

(3) 英汉、藏汉、维汉和蒙汉四个方向将采用基于字符（character-based）的评价方式；

(4) 英汉、藏汉、维汉和蒙汉四个方向中，评测组织方将对 GB2312 编码的 A3 区字符进行全角到半角的转换；

(5) 汉英项目则采用基于词（word-based）的评价方式。

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，测试了所有其它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。

3.3 评测流程

本次评测的流程如下（表 7 给出了此次评测的具体流程）：

(1) 在训练阶段，评测组织方为参评单位发放训练数据、开发数据；

(2) 在正式评测开始时，评测组织方发放测试数据，参评单位在规定时间内提交最终翻译结果文件；

(3) 评测组织方对参评单位提交的最终翻译结果文件进行评测，并为参评单位提供各参评系统的评测结果；

表 7 CWMT 2015 评测流程

时间	事项
4月30日	评测报名截止
6月5日	评测组织方发放训练集、开发集数据，以及BLUE-SBP打分程序、格式检查程序（通过ftp方式发放）
7月13日上午10:00	评测组织方发放汉英、英汉、维汉新闻领域机器翻译三个项目的测试数据
7月17日下午17:30	参评单位提交汉英、英汉、维汉新闻领域机器翻译三个项目的翻译结果
7月20日上午10:00	评测组织方发放蒙汉日常用语、藏汉政府文献机器翻译、双盲测试三个项目的测试数据
7月24日下午17:30	参评单位提交蒙汉日常用语、藏汉政府文献机器翻译、双盲测试三个项目的翻译结果
8月11日	评测组织方向参评单位通知初步评测结果
8月11日	在线评测平台开放（9月25日关闭）
8月20日	参评单位提交评测技术报告
8月26日	评测组织方返回评测技术报告，供参评单位修改
8月30日	评测技术报告终稿
9月23日-9月25日	研讨会召开，会上正式报告评测结果并进行研讨；在线评测平台关闭

4 评测结果

4.1 汉英新闻评测结果

表 8 CWMT 2015 汉英新闻评测主系统评测结果

PARTICIPANTS	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
ce-2015-S14-primary-a.result	0.2338	0.2485	7.5610	0.7170	0.7447	0.5174	0.2907	0.1963	0.7128
ce-2015-S10-primary-a.result	0.2044	0.2160	7.0829	0.7009	0.7660	0.5368	0.2634	0.1845	0.7297
ce-2015-S3-primary-a.result	0.2025	0.2156	7.1771	0.6838	0.7609	0.5357	0.2675	0.1966	0.7209
ce-2015-S4-primary-a.result	0.1953	0.2058	6.8739	0.6738	0.7559	0.5539	0.2510	0.1854	0.7360
ce-2015-S2-primary-a.result	0.1725	0.1818	6.6324	0.6636	0.7842	0.5579	0.2399	0.2392	0.6922
ce-2015-S1-primary-a.result	0.1659	0.1819	5.9812	0.6525	0.7689	0.5693	0.2705	0.1555	0.7392

表 9 CWMT 2015 汉英新闻评测所有系统评测结果

受限系统									
PARTICIPANTS	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
ce-2015-S14-primary-a.result	0.2338	0.2485	7.5610	0.7170	0.7447	0.5174	0.2907	0.1963	0.7128
ce-2015-S14-contrast-b.result	0.2266	0.2465	7.3206	0.7085	0.7360	0.5187	0.2975	0.1859	0.7067
ce-2015-S14-contrast-c.result	0.2264	0.2456	7.3490	0.7018	0.7426	0.5192	0.2955	0.1907	0.7050
ce-2015-S14-contrast-a.result	0.2231	0.2387	7.3950	0.7114	0.7560	0.5205	0.2852	0.1910	0.7153
ce-2015-S3-contrast-c.result	0.2122	0.2217	7.2453	0.7091	0.7427	0.5233	0.2670	0.2041	0.7207
ce-2015-S3-contrast-b.result	0.2103	0.2165	7.1417	0.7177	0.7324	0.5152	0.2509	0.2104	0.7357
ce-2015-S10-primary-a.result	0.2044	0.2160	7.0829	0.7009	0.7660	0.5368	0.2634	0.1845	0.7297
ce-2015-S10-contrast-c.result	0.2035	0.2170	7.0761	0.6947	0.7714	0.5346	0.2699	0.1832	0.7254
ce-2015-S3-primary-a.result	0.2025	0.2156	7.1771	0.6838	0.7609	0.5357	0.2675	0.1966	0.7209
ce-2015-S4-primary-a.result	0.1953	0.2058	6.8739	0.6738	0.7559	0.5539	0.2510	0.1854	0.7360

ce-2015-S4-contrast-a.result	0.1904	0.2005	6.7534	0.6656	0.7597	0.5663	0.2452	0.1831	0.7474
ce-2015-S10-contrast-b.result	0.1831	0.1962	6.2118	0.6477	0.7522	0.5552	0.2800	0.1677	0.7156
ce-2015-S2-primary-a.result	0.1725	0.1818	6.6324	0.6636	0.7842	0.5579	0.2399	0.2392	0.6922
ce-2015-S1-primary-a.result	0.1659	0.1819	5.9812	0.6525	0.7689	0.5693	0.2705	0.1555	0.7392
ce-2015-S1-contrast-c.result	0.1619	0.1772	5.8119	0.6472	0.7695	0.5760	0.2698	0.1519	0.7417
非受限系统									
无									

表 10 CWMT 2015 汉英新闻评测各主系统 BLEU-4SBP 差异显著性检验结果表

（显著标志●,不显著标志○， $p<0.05$）

	ce-2015-S14-primary-a.result	ce-2015-S10-primary-a.result	ce-2015-S3-primary-a.result	ce-2015-S4-primary-a.result	ce-2015-S2-primary-a.result	ce-2015-S1-primary-a.result
ce-2015-S14-primary-a.result	—	●	●	●	●	●
ce-2015-S10-primary-a.result	●	—	●	●	●	●
ce-2015-S3-primary-a.result	●	●	—	●	●	●
ce-2015-S4-primary-a.result	●	●	●	—	●	●
ce-2015-S2-primary-a.result	●	●	●	●	—	●
ce-2015-S1-primary-a.result	●	●	●	●	●	—

4.2 英汉新闻评测结果

表 11 CWMT 2015 英汉新闻评测主系统评测结果

PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
ec-2015-S4-primary-a.result	0.3773	0.3801	0.3148	10.4170	10.4322	0.8565	0.6625	0.3373	0.3505	0.5646	0.5549
ec-2015-S10-primary-a.result	0.3522	0.3600	0.2941	10.1868	10.1990	0.8316	0.7043	0.3599	0.3566	0.5381	0.5505
ec-2015-S2-primary-a.result	0.3403	0.3450	0.2801	10.0238	10.0343	0.8341	0.7002	0.3547	0.3420	0.5327	0.5631
ec-2015-S1-primary-a.result	0.2884	0.3087	0.2468	9.2912	9.3019	0.7850	0.7333	0.4104	0.3335	0.4762	0.5853

表 12 CWMT 2015 英汉新闻评测所有系统评测结果

受限系统											
PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
ec-2015-S4-primary-a.result	0.3773	0.3801	0.3148	10.4170	10.4322	0.8565	0.6625	0.3373	0.3505	0.5646	0.5549
ec-2015-S4-contrast-a.result	0.3706	0.3734	0.3076	10.3363	10.3505	0.8554	0.6673	0.3400	0.3452	0.5623	0.5600
ec-2015-S10-primary-a.result	0.3522	0.3600	0.2941	10.1868	10.1990	0.8316	0.7043	0.3599	0.3566	0.5381	0.5505
ec-2015-S10-contrast-b.result	0.3522	0.3600	0.2941	10.1868	10.1990	0.8316	0.7043	0.3599	0.3566	0.5381	0.5505
ec-2015-S2-primary-a.result	0.3403	0.3450	0.2801	10.0238	10.0343	0.8341	0.7002	0.3547	0.3420	0.5327	0.5631
ec-2015-S1-primary-a.result	0.2884	0.3087	0.2468	9.2912	9.3019	0.7850	0.7333	0.4104	0.3335	0.4762	0.5853
ec-2015-S1-contrast-c.result	0.2855	0.3103	0.2482	9.1263	9.1369	0.7783	0.7309	0.4183	0.3378	0.4684	0.5842
非受限系统											
无											

表 13 CWMT 2015 汉英新闻评测各主系统 BLEU-5SBP 差异显著性检验结果表

（显著标志●,不显著标志○， p<0.05）

	ec-2015-S4-primary-a.result	ec-2015-nlp2ct-primary-a.result	ec-2015-S2-primary-a.result	ec-2015-S1-primary-a.result
ec-2015-S4-primary-a.result	—	●	●	●
ec-2015-S10-primary-a.result	●	—	●	●
ec-2015-S2-primary-a.result	●	●	—	●
ec-2015-S1-primary-a.result	●	●	●	—

4.3 蒙汉日常用语评测结果

表 14 CWMT 2015 蒙汉日常用语评测主系统评测结果

PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
mc-2015-S5-primary-a.result	0.6559	0.7077	0.6835	10.9016	10.9628	0.8336	0.2558	0.2139	0.7930	0.7680	0.2350
mc-2015-S6-primary-a.result	0.5704	0.6279	0.6013	10.1126	10.1583	0.7931	0.3289	0.2697	0.7324	0.7083	0.3013
mc-2015-S7-primary.result	0.3642	0.4061	0.3749	7.1364	7.1589	0.6060	0.5429	0.4831	0.4895	0.5038	0.5066

表 15 CWMT 2015 蒙汉日常用语评测所有系统评测结果

受限系统											
PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
mc-2015-S5-primary-a.result	0.6559	0.7077	0.6835	10.9016	10.9628	0.8336	0.2558	0.2139	0.7930	0.7680	0.2350
mc-2015-S5-contrast-b.result	0.6390	0.6869	0.6627	10.5573	10.6141	0.8281	0.2660	0.2304	0.7960	0.7497	0.2475
mc-2015-S5-contrast-c.result	0.6091	0.6542	0.6252	10.4710	10.5209	0.8216	0.2893	0.2306	0.7694	0.7379	0.2582

mc-2015-S6-contrast-b.result	0.5778	0.6291	0.6032	10.0884	10.1357	0.7971	0.3151	0.2630	0.7438	0.7117	0.2887
mc-2015-S6-primary-a.result	0.5704	0.6279	0.6013	10.1126	10.1583	0.7931	0.3289	0.2697	0.7324	0.7083	0.3013
mc-2015-S6-contrast-c.result	0.5673	0.6234	0.5978	10.0207	10.0666	0.7921	0.3303	0.2720	0.7324	0.7046	0.3020
mc-2015-S7-primary.result	0.3642	0.4061	0.3749	7.1364	7.1589	0.6060	0.5429	0.4831	0.4895	0.5038	0.5066
mc-2015-S7-contrast.result	0.3618	0.3933	0.3642	6.6981	6.7233	0.6007	0.5431	0.4830	0.5034	0.4916	0.5044
非受限系统											
无											

表 16 CWMT 2015 蒙汉日常用语评测各主系统 BLEU-5SBP 差异显著性检验结果表

（显著标志●,不显著标志○， $p<0.05$）

	mc-2015-S5-primary-a.result	mc-2015-S6-primary-a.result	mc-2015-S7-primary.result
mc-2015-S5-primary-a.result	—	●	●
mc-2015-S6-primary-a.result	●	—	●
mc-2015-S7-primary.result	●	●	—

4.4 藏汉政府文献评测结果

表 17 CWMT 2015 藏汉政府文献评测主系统评测结果

PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
tc-2015-S7-primary.result	0.8809	0.8973	0.8773	12.9015	13.0378	0.9442	0.1384	0.0839	0.9218	0.9052	0.1014
tc-2015-S14-primary-a.result	0.8801	0.8951	0.8755	12.8554	12.9906	0.9443	0.1381	0.0921	0.9154	0.8962	0.1092
tc-2015-S11-primary-a.result	0.8061	0.8340	0.8075	12.3585	12.4819	0.9247	0.2013	0.1141	0.8801	0.8690	0.1432

表 18 CWMT 2015 藏汉政府文献评测所有系统评测结果

受限系统											
PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
tc-2015-S7-primary.result	0.8809	0.8973	0.8773	12.9015	13.0378	0.9442	0.1384	0.0839	0.9218	0.9052	0.1014
tc-2015-S14-primary-a.result	0.8801	0.8951	0.8755	12.8554	12.9906	0.9443	0.1381	0.0921	0.9154	0.8962	0.1092
tc-2015-S7-contrast-a.result	0.8751	0.8895	0.8684	12.8625	12.9973	0.9438	0.1441	0.0885	0.9137	0.9001	0.1069
tc-2015-S14-contrast-c.result	0.8748	0.8895	0.8691	12.8186	12.9524	0.9429	0.1442	0.0949	0.9113	0.8939	0.1127
tc-2015-S14-contrast-a.result	0.8714	0.9013	0.8812	12.6624	12.7969	0.9370	0.1510	0.0901	0.9311	0.8932	0.1085
tc-2015-S7-contrast-b.result	0.8152	0.8228	0.7924	12.5526	12.6592	0.9490	0.2102	0.0932	0.8717	0.8871	0.1333
tc-2015-S11-primary-a.result	0.8061	0.8340	0.8075	12.3585	12.4819	0.9247	0.2013	0.1141	0.8801	0.8690	0.1432
非受限系统											
tc-2015-S14-contrast-b.result	0.8727	0.9025	0.8834	12.6330	12.7651	0.9341	0.1499	0.0917	0.9310	0.8929	0.1089

表 19 CWMT 2015 藏汉政府文献评测各主系统 BLEU-5SBP 差异显著性检验结果表

（显著标志●,不显著标志○， $p<0.05$）

	tc-2015-S7-primary.result	tc-2015-S14-primary-a.result	tc-2015-S11-primary-a.result
tc-2015-S7-primary.result	—	○	●
tc-2015-S14-primary-a.result	○	—	●
tc-2015-S11-primary-a.result	●	●	—

4.5 维汉新闻评测结果

表 20 CWMT 2015 维汉新闻评测主系统评测结果

PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
uc-2015-S7-primary.result	0.4190	0.4390	0.3956	10.0733	10.0906	0.7645	0.5009	0.3504	0.4473	0.5782	0.4415
uc-2015-S5-primary-a.result	0.4108	0.4244	0.3764	9.9105	9.9273	0.7847	0.4723	0.3471	0.4963	0.5811	0.4199
uc-2015-S12-primary-a.result	0.4105	0.4262	0.3826	9.9230	9.9404	0.7652	0.4723	0.3504	0.4640	0.5777	0.4258
uc-2015-S13-primary-a.result	0.3733	0.3866	0.3390	9.3740	9.3874	0.7682	0.5013	0.3697	0.4590	0.5507	0.4505

表 21 CWMT 2015 维汉新闻评测所有系统评测结果

受限系统											
PARTICIPANTS	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
uc-2015-S7-primary.result	0.4190	0.4390	0.3956	10.0733	10.0906	0.7645	0.5009	0.3504	0.4473	0.5782	0.4415
uc-2015-S5-primary-a.result	0.4108	0.4244	0.3764	9.9105	9.9273	0.7847	0.4723	0.3471	0.4963	0.5811	0.4199
uc-2015-S12-primary-a.result	0.4105	0.4262	0.3826	9.9230	9.9404	0.7652	0.4723	0.3504	0.4640	0.5777	0.4258
uc-2015-S5-contrast-b.result	0.3918	0.4041	0.3565	9.6633	9.6781	0.7674	0.4900	0.3567	0.4685	0.5662	0.4347
uc-2015-S7-contrast.result	0.3839	0.4004	0.3565	9.4760	9.4907	0.7481	0.5016	0.3767	0.4286	0.5519	0.4537
uc-2015-S13-primary-a.result	0.3733	0.3866	0.3390	9.3740	9.3874	0.7682	0.5013	0.3697	0.4590	0.5507	0.4505
非受限系统											
uc-2015-S12-contrast-c.result	0.4530	0.4693	0.4269	10.5995	10.6189	0.7873	0.4396	0.3166	0.4978	0.6140	0.3894
uc-2015-S12-contrast-b.result	0.4514	0.4658	0.4240	10.5193	10.5383	0.7841	0.4421	0.3168	0.5021	0.6127	0.3902
uc-2015-S13-contrast-a.result	0.2886	0.3034	0.2581	8.6228	8.6309	0.7126	0.5848	0.4109	0.3712	0.4940	0.5167

表 22 CWMT 2015 维汉新闻评测各主系统 BLEU-5SBP 差异显著性检验结果表

（显著标志●,不显著标志○， p<0.05）

	uc-2015-S7-primary.result	uc-2015-S5-primary-a.result	uc-2015-S12-primary-a.result	uc-2015-S13-primary-a.result
uc-2015-S7-primary.result	—	●	●	●
uc-2015-S5-primary-a.result	●	—	○	●
uc-2015-S12-primary-a.result	●	○	—	●
uc-2015-S13-primary-a.result	●	●	●	—

4.6 双盲评测结果

表 23 CWMT 2015 双盲评测主系统评测结果

PARTICIPANTS	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
db-2015-S14-primary-a.result	0.1973	0.2093	6.7074	0.6585	0.7769	0.5641	0.2381	0.2238	0.6925
db-2015-S4-primary-a.result	0.1934	0.2029	6.6491	0.6611	0.7648	0.5699	0.2286	0.2251	0.7042
db-2015-S3-primary-a.result	0.1870	0.1969	6.5662	0.6537	0.7840	0.5700	0.2311	0.2205	0.7009
db-2015-S10-primary-a.result	0.1768	0.1888	6.3917	0.6422	0.7990	0.5799	0.2296	0.2150	0.7030

表 24 CWMT 2015 双盲评测所有系统评测结果

受限系统									
PARTICIPANTS	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
db-2015-S4-contrast-b.result	0.1977	0.2051	6.6193	0.6656	0.7549	0.5604	0.2214	0.2281	0.7150
db-2015-S14-contrast-b.result	0.1975	0.2094	6.7102	0.6584	0.7769	0.5646	0.2382	0.2240	0.6931
db-2015-S14-primary-a.result	0.1973	0.2093	6.7074	0.6585	0.7769	0.5641	0.2381	0.2238	0.6925
db-2015-S14-contrast-a.result	0.1969	0.2079	6.7370	0.6554	0.7752	0.5625	0.2366	0.2248	0.6947
db-2015-S4-contrast-a.result	0.1965	0.2035	6.5375	0.6675	0.7539	0.5621	0.2183	0.2276	0.7210
db-2015-S4-contrast-c.result	0.1943	0.2049	6.6926	0.6634	0.7699	0.5693	0.2331	0.2252	0.7010
db-2015-S4-primary-a.result	0.1934	0.2029	6.6491	0.6611	0.7648	0.5699	0.2286	0.2251	0.7042
db-2015-S3-primary-a.result	0.1870	0.1969	6.5662	0.6537	0.7840	0.5700	0.2311	0.2205	0.7009
db-2015-S10-primary-a.result	0.1768	0.1888	6.3917	0.6422	0.7990	0.5799	0.2296	0.2150	0.7030
db-2015-S10-contrast-c.result	0.1752	0.1880	6.3165	0.6369	0.7965	0.5799	0.2298	0.2128	0.7028
db-2015-S10-contrast-b.result	0.1745	0.1855	6.3361	0.6426	0.7992	0.5807	0.2257	0.2141	0.7103
非受限系统									
无									

表 25 CWMT 2015 双盲评测各主系统 BLEU-4SBP 差异显著性检验结果表

（显著标志●,不显著标志○， $p<0.05$）

	db-2015-S14-primary-a.result	db-2015-S4-primary-a.result	db-2015-S3-primary-a.result	db-2015-S10-primary-a.result
db-2015-S14-primary-a.result	—	●	●	●
db-2015-S4-primary-a.result	●	—	○	●
db-2015-S3-primary-a.result	●	○	—	○
db-2015-S10-primary-a.result	●	●	○	—

致谢

特别感谢 CWMT 2015 的合作单位和评测资源提供单位：北京大学、点通数据有限公司、东北大学、东芝（中国）研究开发中心、哈尔滨工业大学、南京大学、内蒙古大学、青海师范大学、西北民族大学、西藏大学、厦门大学、新疆大学、中国科学院合肥智能机械研究所、中国科学院计算技术研究所、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。

参考文献

[1] 赵红梅，谢军，吕雅娟，于惠，张昊亮，刘群，第九届全国机器翻译研讨会（ CWMT 2013）评测报告，第九届全国机器翻译研讨会（ CWMT2013），2013 年 10 月 31-11 月 1 日， 昆明

[2] David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In Proc. EMNLP 2008, pages 610-619.

[3] Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statS12al machine translation. In Proc. ACL 2005, pages 531–540.

[4] LDC. 2005. LinguS12 data annotation specification: Assessment of fluency and adequacy in translation. Revision 1.5.

[5] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.

[6] Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational LinguS12s (ACL-2005), Ann Arbor, Michigan, June 2005

[7] Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang and Tiejun Zhao. 2008. Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed LinguS12 CheckPoints. In Proc. Coling 2008, pages 1121–1128.

附录：参评主系统描述

系统名称	系统描述
ce-2015-S1-primary-a	Hardware and OS: Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-24-generic x86_64) train set： 6493765 sentence-pairs, dev set：1000 sentence-pairs External Tools: NiuTrans open source statistical machine translation system Technique: Hierarchy phase-based model, reodering model
ce-2015-S2-primary-a	软硬件环境：Ubuntu 14.04 CPU：Intel(R) Core(TM) i7-3770 3.4GHz 内存：24G 基于对数线性模型的短语统计机器训练，增加预调序 训练集：6021778 句对 开发集：1006 句对 外部技术：Moses 开源统计机器翻译系
ce-2015-S3-primary-a	Hardware and software environment: Linux, 24 CPUS, 256GB memory. Excution Time: 200 hours. Technology outline: domain adaptation. Training Data: selected. 2013 testing, 2015 tuning. External Technology: moses.
ce-2015-S4-primary-a	Hardware and OS: Linux Centos 6.0 Cpu_core_number=8 Intel Xeon E5-2670 2.60GHz RAM_size=62G Technique:Phrase-based translation model, hier lex reordering model, operation sequence model, phrase scoring with KneserNey smooth, k-batch mira, rule-based spelling of name. Used data and corpus: tranning set data from CWMT2015 chinese-to-english, tuning dev data from CWMT2015 chinese-to-english, language model from the target of CWMT2015 chinese-to-english and sougou mono language. External Tools: Srilm; Fast_align; Giza++; NPLM; RNN.

ce-2015-S10-primary-a	Hardware and OS: Linux Centos 5.5 CPU_core_number=24 24 Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz RAM_SIZE=189G. Technique: phrase-based translation model, lexical reordering model, language model, data selection with RAE,MERT. Used data and corpus: CE: constrained, all training data and test data from CWMT 2015, develop data from the target language of training data plus Reuters-Corpus-English. External Tools: SRILM, fast-align, RAE,MERT
ce-2015-S14-primary-a	Ubuntu 12.04 10CPU 2.6GHz 100GB 13h hpb + 5srilm + 5blm + 15nnjm + 15nnltm 主办方发放的训练数据和开发集 srilm, moses
ec-2015-S1-primary-a	软硬件环境： Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-24-generic x86_64) 训练集： 6493765 句对 开发集：1000 句对 外部技术：NiuTrans 开源统计机器翻译系统 技术概要：层次短语模型 重排序模型
ec-2015-S2-primary-a	软硬件环境：Ubuntu 14.04 CPU：Intel(R) Core(TM) i7-3770 3.4GHz 内存：24G 基于对数线性模型的短语统计机器训练，增加预调序 训练集：6021778 句对 开发集：1000 句对 外部技术：Moses 开源统计机器翻译系统
ec-2015-S4-primary-a	Hardware and OS: Linux Centos 6.0 Cpu_core_number=8 Intel Xeon E5-2670 2.60GHz RAM_size=62G Technique:Phrase-based translation model, hier lex reordering model, operation sequence model, phrase scoring with KneserNey smooth, k-batch mira, rule-based spelling of name.

	Used data and corpus: tranning set data from CWMT2015 chinese-to-english, tuning dev data from CWMT2015 chinese-to-english, language model from the target of CWMT2015 chinese-to-english and sougou mono language. External Tools: Srilm; Fast_align; Giza++; NPLM; RNN.
ec-2015-S10-primary-a	Hardware and OS: Linux Centos 5.5 CPU_core_number=24 24 Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz RAM_SIZE=189G. Technique: phrase-based translation model, lexical reordering model, language model, data selection with RAE,MERT. Used data and corpus: EC: constrained, all training data and test data from CWMT 2015, develop data from the target language of training data plus Sogou-corpus-chn. External Tools: SRILM, fast-align, RAE,MERT
mc-2015-S5-primary-a	(1)软硬件环境：64 位的操作系统 ubuntu12.04；cpu 为 2*16 核；cpu 类型及频率为 Inter(R) Xeon(R) CPU E5-2687W v2 @ 3.40GHz；系统内存大小为：31.4GB； (2)运行时间：约 12 分钟； (3)技术概要：该系统为主系统，是一个基于短语的词干级统计机器翻译系统；首先对蒙古语进行形态切分，提取出蒙古语词干，使用两种词对齐（词干级 berkeley 对齐、词干级 GIZA++词对齐）、两种语言模型(训练集单语语言模型、搜狗语言模型)，最大短语长度为 7，该系统使用对数线性模型对各种参数特征进行融合；使用命名实体、数字时间词识别模块；对解码出的 100best 译文进行 ter 融合，再通过 PageRank 对全排矩阵进行排序；对译文进行了去除未登录词处理； (4)训练数据说明：训练数据受限，仅采用由评测方提供的语料以及搜狗新闻语料，但对语料进行了筛选； (5)外部技术说明：使用 moses 进行短语的解码
mc-2015-S6-primary-a	ubuntu 12.04 GNU/Linux 3.2.0-29-generic x86_64 CPU 4

	Intel(R)Xeon（R）CPU E7-4830@2.13GHz Memory 32G 系统运行时间约 6 分钟 训练语料和开发集均为 CWMT2015 组织方提供的语料； 语言模型训练语料由《搜狗全网新闻语料库》08 版和训练语料中汉语部分构成； 本系统中的蒙古文语料全部采用了内大拉丁转写形式，采用基于短语的统计机器翻译模型实现蒙古文的词性标注和切分，词性标注的未登录词用我们实验室开发的基于隐马尔科夫模型的词性标注系统进行标注，采用我们实验室开发的基于词缀词典的切分系统对未登录词中的动词和分写附加成分进行切分，采用 ICTCLAS2015 对汉语语料进行分词和词性标注； 本系统采用基于短语的因子化模型，融入蒙古文词干，词缀，词性和汉语词性等因子，实现蒙古文到汉语的翻译； 采用的开源软件有：ICTCLAS2015，语言模型训练工具 SRILM，词语对齐工具 GIZA++和解码器 moses 等；
mc-2015-S7-primary	软硬件环境：LINUX 操作系统，版本为 Red Hat Enterprise Linux Server release 6.3，8 CPU、Intel(R) Xeon(R) CPU E7520@1.87GHz、系统内存 132GB； 运行时间：大约 5 个小时。 技术概要：使用了 Berkeley 词对齐，采用了基于短语的统计机器翻译。 训练数据说明：统计系统采用了评测组织方发布的训练数据和开发数据； 外部技术说明：Niutrans 系统的部分工具和 Berkeley 词对齐工具。
tc-2015-S7-primary	软硬件环境：LINUX 操作系统，版本为 Red Hat Enterprise Linux Server release 6.3，8 CPU、Intel(R) Xeon(R) CPU E7520@1.87GHz、系统内存 132GB； 运行时间：大约 5 个小时。 技术概要：使用了 sogou 数据训练大的语言模型，采用了多翻译系统融合的翻译方法。 训练数据说明：统计系统采用了评测组织方发布的训练数据和开发数据，以及 sogou 数据； 外部技术说明：Moses、Niutrans 系统的部分工具。

tc-2015-S11-primary-a	本提交结果由“阳光”统计翻译系统自动生成。仅采用提供的数据。
tc-2015-S14-primary-a	Ubuntu 12.04, 12 cores, 2.6GHz, 100G 2h hpb + 5sirlm + 5blm + 15nnjm + 15nnltm 主办方提供的训练数据和开发集 srilm, moses
uc-2015-S5-primary-a	(1)软硬件环境：64 位的操作系统 ubuntu12.04；cpu 为 2*16 核；cpu 类型及频率为 Inter(R) Xeon(R) CPU E5-2687W v2 @ 3.40GHz；系统内存大小为：31.4GB； (2)运行时间：约 20 分钟； (3)技术概要：该系统为主系统，是一个基于层次短语的统计机器翻译系统；使用三种词对齐（词级 GIZA++、词级 berkeley 对齐、词干级 GIZA++词对齐）、两种语言模型(训练集单语语言模型、搜狗语言模型)，最大短语长度为 7，该系统使用对数线性模型对各种参数特征进行融合；使用自底向上的 CKY 算法进行解码；对解码出的 100best 译文进行 ter 融合，再通过 PageRank 对全排矩阵进行排序；对译文进行了去除未登录词处理； (4)训练数据说明：训练数据受限，仅采用由评测方提供的维汉双语对齐语料库以及搜狗新闻语料； (5)外部技术说明：使用 moses 进行层次短语的解码；使用 IlgharPad 进行传统维文到拉丁维文的转换
uc-2015-S7-primary	软硬件环境：LINUX 操作系统，版本为 Red Hat Enterprise Linux Server release 6.3，8 CPU、Intel(R) Xeon(R) CPU E7520@1.87GHz、系统内存 132GB； 运行时间：大约 5 个小时。 技术概要：使用了 sogou 数据训练大的语言模型，采用了基于短语的统计机器翻译。 训练数据说明：统计系统采用了评测组织方发布的训练数据和开发数据，以及 sogou 数据； 外部技术说明：Niutrans 系统的部分工具。
uc-2015-S12-primary-a	OS : Ubuntu server 12.04.2 CPU: Intel(R) Xeon(R) CPU E5-2690 0 @ 2.90GHz, 16 cores

	Mem: 128G HDD: 1.5T 技术概要: 使用 giza++进行词对齐，使用短语模型，使用 moses 进行解码。对语料进行 token 以及 stem 处理 训练数据: 从 14W 训练集中抽取 139792 句对作为训练语料，开发集使用提供的语料，并进行拼写检查 外部数据: 主要使用 giza++，moses，使用 CRF 进行中文分词
uc-2015-S13-primary-a	新疆大学机器翻译系统 软硬件环境：操作系统：ubuntu、版本：14.04.1、CPU 数量：8、CPU 类型： Intel(R) Xeon(R) CPU E5-2609 v2、CPU 频率：2.50GHz 、系统内存大小：132G; 运行时间：参评系统从接受输入到产生全部输出所花费的时间:约 186 分钟; 技术概要：使用了 Moses 开源工具，进行了基于短语的机器翻译技术。对源语言采用了新疆大学 tokenizer 工具和新疆大学词干提取工具，只保留了词干。 训练数据说明：使用了评测单位提供的的训练数据和开发数据,对应语言模型使用了搜狗语料+训练语料; 外部技术说明：汉语分词使用了中科院的词法分析工具
db-2015-S3-primary-a	Hardware and software environment: Linux, 24 CPUS, 256GB memory. Excution Time: 200 hours. Technology outline: domain adaptation. Training Data: all. External Technology: moses.
db-2015-S4-primary-a	Hardware and OS: Linux Centos 6.0 Cpu_core_number=8 Intel Xeon E5-2670 2.60GHz RAM_size=62G Technique:Phrase-based translation model, lex reordering model, operation sequence model, phrase scoring with KneserNey smooth, k-batch mira. Used data and corpus: tranning set data from CWMT2015 cwmtdb_2015, tuning dev data from CWMT2015 cwmtdb_2015, language model from the target of cwmtdb_2015.

	External Tools: Srilm; Fast_align; Giza++; NPLM; RNN.
db-2015-S10-primary-a	Hardware and OS: Linux Centos 5.5 CPU_core_number=24 24 Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz RAM_SIZE=189G. Technique: phrase-based translation model, lexical reordering model, language model,MERT. Used data and corpus: DB: constrained, all training data and test data from CWMT 2015, develop data from the target language of training data plus monolingual corpus for target language. External Tools: SRILM, fast-align
db-2015-S14-primary-a	Ubuntu 12.04 10CPU 2.6GHz 100GB hpb + 5srilm + 5blm + 10nnjm + 6nnltm 主办方发放的训练数据和开发集 srilm, moses