文章编号:1003 - 0077(2006)增刊 - 0001 - 06

2005 年度 863 计划中文信息处理与智能人机接口技术评测回顾

钱跃良,林守勋,刘 群,刘 宏

(中国科学院计算技术研究所,北京 100080)

摘要:2005 年度 863 计划中文信息处理与智能人机接口技术评测于 2005 年 9 月 20 - 22 日举行。本次评测涉及机器翻译、语音识别、信息检索三大类技术。本文给出了此次评测的组织过程、参评单位、评测方案、数据准备、结果分析等各方面总体情况,更详细的情况在本论文集收集的各分项评测报告中给出。与往年的 863 评测相比,本次评测的主要特点是全面采用了国际上通行的网上评测的方式,提供了大量的训练数据,并且在评测研讨会上为参评单位提供了更加充分的交流机会。

关键词:技术评测;自然语言处理;人机交互;机器翻译;语音识别;信息检索

中图分类号:TP391 **文献标识码**:A

A Review on 2005 HTRDP (863) Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface

QIAN Yue-liang, LIN Shou-xun, LIU Qun, LIU Hong

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The 2005 HTRDP (863) evaluation on Chinese information processing and intelligent human-machine interface was executed in September 20 - 22, 2005. In this evaluation, technologies including machine translation, acoustics speech recognition and information retrieval are evaluated. This paper gives an overall introduction to various aspects of this evaluation, including organization process, participants, data preparing, result analysis, and etc. More details can be found in the evaluation reports on specific tasks collected in this proceeding. Compared with the previous HTRDP evaluations, the 2005 evaluation has some new distinguishes, including: adopting the internet-based test method instead of the on-the-spot method; providing large scale training data; providing more presentation and discussion time for the participants in the evaluation workshop.

Key words: technology evaluation; natural language processing; human-machine interaction; machine translation; speech recognition; information retrieval

1 引言

众所周知,在自然语言处理、人机交互等研究领域,大规模的数据资源建设与共享和公开的技术评测,已经成为推动学科发展的两个重要的驱动力[2~8]。为了了解国内外在中文信息处理与智能人机接口技术领域的现状,检查 863 计划信息领域计算机软硬件技术主题(以下简称计算机主题)中相关课题的进展情况,促进交流和提高,推动技术进步和成果的应用与产

收稿日期:2005 - 11 - 02 **定稿日期**:2006 - 01 - 16

基金项目:国家 863 计划资助项目(2004AA114010;2003AA111010)

作者简介:钱跃良(1960—),男,正高级工程师,主要研究方向为中文信息处理与人机交互.

业化，并为863计划课题验收和以后的课题评选打下基础，国家863计划计算机软硬件主题（原智能计算机主题）专家组从1991年开始，举办了8次中文信息处理与智能人机接口技术评测，产生了广泛的影响[1,9]。特别是从2003年以来，该项评测活动在停顿5年之后重新恢复，并开始进行了每年一度的评测，与过去的评测相比，组织方式更加规范合理并逐步向国际惯例靠拢，评测的项目和参与的单位都大大增加，863评测的影响日益扩大。2005年度的863评测于9月20日至22日正式举行，本次评测采用通过Internet发布数据和提交结果的方式。评测结束后，于11月28日在北京举行了评测研讨会，会上发布了本次评测的正式结果，同时参评的各单位就本次评测所采用的技术和评测中遇到的各方面问题展开了深入的交流。

本文介绍了本次评测的总体情况。各项评测的详细情况，请参加本次评测的各单项评测报告，这些报告都收录在本论文集中。本文的组织方式如下：第1节是引言，第2节概述了本次评测的全貌，第3节详细介绍本次评测的组织过程，第4节介绍了本次评测的组织机构和对外交流等方面情况，第5节是总结和对今后工作的展望。

2 2005年863评测概述

本次评测共设三个大类的评测，分别为机器翻译、信息检索、语音识别。每个大类的评测又分若干任务，每个任务中包括若干评测项目，如表1所示。

表1 评测任务和项目设置

评测	任 务	项 目	项目代号
机器翻译	机器翻译	汉英翻译	MTCE
		英汉翻译	MTEC
		汉日翻译	MTCJ
		日汉翻译	MTJC
		日英翻译	MTJE
		英日翻译	MTEJ
	词语对齐	汉英词语对齐	WACE
信息检索	相关网页检索	相关网页检索	WEB
语音识别	桌面连续语音识别	桌面连续语音识别（2倍实时）	CSR_PC_2X
		桌面连续语音识别（20倍实时）	CSR_PC_20X
	电话连续语音关键词检测	电话连续语音关键词检测（2倍实时）	KWS_PHONE_2X

（1）机器翻译的评测

设汉－英、英－汉、汉－日、日－汉、日－英、英－日六个方向的机器翻译评测项目，以及汉英词语自动对齐评测项目。

（2）信息检索的评测

只设置一个项目：相关网页检索。

（3）语音识别的评测

评测语种只设汉语，包括桌面连续语音识别和电话连续语音中的关键词检测两个任务，其中桌面连续语音识别分为2倍实时和20倍实时两个项目，电话连续语音关键词检测只设置2倍实时一个项目。

本次评测吸引了来自国内外的二十多个研究单位参加，具体参评情况如下：参加机器翻译

评测的有12个研究单位,提交了23个参评系统;参加语音识别的有8个研究单位,提交了17个参评系统;参加信息检索的有5个研究单位,提交了5个参评系统,另有3个报名参评的单位没有提交最终结果。

3 评测的过程

本次评测的进度安排如下:

表2 评测时间表

日 期	事 项
3月-4月	评测大纲征求意见
4月29日	评测大纲(中、英文)发布
7月29日	参评报名截止
8月1日	训练集数据发布
8月22日	开发集数据发布
8月22日	信息检索源数据发布
9月20日	测试数据发布
9月22日	测试结果提交截止
10月21日	评测结果通知
11月28-29日	评测研讨会

下面我们详细介绍本次评测的具体组织过程。

3.1 确定评测项目、制订评测大纲

这个阶段的工作主要是根据2004年度评测的情况,结合目前863计划相关课题的设置和技术发展趋势,确定2005年度的评测内容,并制定了评测大纲(主要包括评测内容、评测方法和评测指标等)。

在评测大纲的制订过程中,评测组通过开会讨论、电子邮件等多种方式充分征求了往年的参评单位和一些著名专家的意见,形成了比较规范合理、被参评单位所普遍认可的评测大纲。

3.2 公告和受理报名

我们将评测大纲和报名表以中英文两种形式在评测网站(http://www.863data.org.cn)以及一些著名的邮件列表上公布,并受理报名。参评单位根据评测大纲选择所要参加的评测类别和评测项,分类填写报名表,并在规定的报名时间内提交。在报名表上,参评单位要给出单位和系统的基本信息,并按照评测的要求做出相应的承诺,然后将报名表以传真和电子邮件两种方式发送到评测组。

3.3 组织评测数据

根据评测大纲,设计和组织评测数据。评测数据包括训练集、开发集和测试集数据。

训练集数据供参评单位训练系统用。本次评测与往年评测相比的一个重要特点是提供了大量的训练语料库。例如,汉英机器翻译训练语料库达87万句子对,对于电话语音关键词识别评测,我们提供了5个小时的完全自然的电话语音训练语料库,这些都是非常宝贵的资源,对于参评单位来说,可以免费得到并用于参加评测。

开发集数据模仿测试集数据的模式,供参评单位调试参评系统。本次评测对于所有的评测项目都提供了开发数据集。

测试数据用于正式评测。所有测试数据都有评测组开发或者由合作单位提供。

这些数据都通过评测网站在网上发布,参评单位可以得到评测组提供的一个账号和密码。

评测单位在网站上下载数据后,在自己的软硬件环境中运行参评系统,在规定的期限前,登录网站并通过网络提交结果。

对于测试数据,评测组(和合作单位)还制作了参考答案,用于对参评系统提交的结果进行评价。同时,评测组还为每项评测提供了评测的软件工具。

在评测结束后,评测组会将所有的数据(含参考答案和评测软件)打包整理,并通过 ChineseLDC 向社会公开发布,以促进研究工作。

3.4 统计和公布评测结果

结果评价的方法依具体评测项目有所不同,总体上可以分为主观评价和客观评价两大类。

对于客观评价的评测项目,由评测软件自动统计各系统提交的运行结果,形成评测结果。

对于主观评判的评测项目,由评测单位聘请有关专家,按照评测大纲的规定对参评系统提交的结果进行人工评判,最后进行统计,形成评测结果。

语音识别结果的评价采用客观的评价方法,机器翻译结果的评价采用主观评价与客观评价相结合的方法,信息检索的评价采用客观评价方法,但参考答案的制作采用 Pooling 方法,这又有一定的主观因素,因此可以认为是一种主观评价与客观评价相结合的方法。

由评测单位对评测结果进行核对,上报 863 专家组。同时,将各单位自己的结果和相关项目的最好结果通过电子邮件通知各参评单位,完整的评测结果在评测研讨会上公布。

3.5 组织评测技术研讨会

评测研讨会于 11 月 28 日至 29 日上午在北京翠宫饭店举行。会议邀请了 863 计划计算机软硬件主题专家组组长怀进鹏教授致开幕词,并邀请了欧盟 TC-STAR 项目的协调人、意大利 ITC-IRST 研究所的 Gianni Lazzari 教授和复旦大学的黄萱菁副教授做特邀报告。Lazzari 教授介绍了 TC-STAR 项目评测的情况,黄萱菁副教授介绍了复旦大学信息检索课题组参加多次国际评测的体会和经验。这些报告都让参会者收益匪浅。在后面的议程中,参评单位按照评测的技术类别分为三个分会场,进行了分组报告。每个技术报告最后都安排了充分的时间,供参评单位进行系统演示和自由讨论。

本次评测特别强化了评测研讨会的交流作用。评测组通过各种方式鼓励参评单位参加评测研讨会,这些措施包括:将完整的评测结果放到评测研讨会上公布;要求参评单位提交完整详细的技术报告,以便于其他单位能完整了解各系统的技术细节;规定只有提交了完整的技术报告并参加评测研讨会的单位才能得到测试数据的参考答案和评测工具;研讨会免费注册;研讨会上给参评单位提供充分的报告时间(每个报告 30 分钟),并提供了专门的自由讨论时间;研讨会的所有报告将以《中文信息学报》专刊的形式发表;等等。

4 评测的组织机构和对外交流

本年度的评测由 863 计划计算机主题专家组委托中国科学院计算技术研究所主办,由中国科学院软件研究所和日本情报通信研究机构(NICT)协办。今年,中国科学院计算技术研究所专门成立了“多语言交互技术评测实验室”,全面负责与评测相关的组织工作。实验室主任由钱跃良研究员担任。评测实验室的成立,从各方面为评测工作开展提供了更加有力的保障。

在具体的评测实施过程中,我们在一些具体的工作中与以下单位开展了合作:

- 日本情报通信研究机构 Keihanna 情报通讯融合研究中心(机器翻译)
- 中国科学院软件研究所开放系统与中文信息处理中心(信息检索)
- 北京大学计算机网络与分布式系统实验室(信息检索)

• 微软亚洲研究院(信息检索)

具体合作内容包括:在机器翻译评测中,与日本情报通信研究机构(NICT)进行了密切的合作,评测中部分有关日语的数据采集和加工都是由 NICT 完成的,一些有关日语的人工评测工作也是由 NICT 完成的;在信息检索评测中,与北京大学计算机网络与分布式系统实验室组织的"SEWM2005 中文 Web 信息检索评测"进行了协调与合作,采用了北京大学提供 CWT100g 数据集;信息检索评测的测试题构造和测试结果的评价(含 Pooling)是由中国科学院软件研究所完成的;微软亚洲研究院为信息检索评测大纲的制定和大纲的发布提供了帮助。

另外,我们今年还进行的广泛的国际学术交流,具体包括:

• 今年 1 月,林守勋、刘群、陶建华一行访问了美国著名的评测机构 - 国家标准技术局 NIST 和 CMU、MIT、UIUC 等著名大学,期间与 NIST 相关评测的负责人进行了有效深入的交流。

• 今年 4 月,在香港理工大学做了一次学术报告,详细介绍了 863 评测的情况。

• 今年 4 月,访问了意大利 ITC-IRST 研究所,参加了 TC-STAR 的评测研讨会,并在会上做报告介绍了 863 评测。

• 今年 9 月,应邀参加国际机器翻译峰会 MT Summit X 并做特邀报告,详细介绍了 863 机器翻译评测的情况。

• 今年 11 月,在中日自然语言处理促进会议 CJNLP05 和中日通信与计算机技术交流会 ICT2005 上做报告,宣传与介绍 863 评测工作。

• 本次评测研讨会邀请了来自意大利 ITC-IRST 研究所的欧盟 TC-STAR 项目协调人 Lazzari 博士做专题报告。

通过这些学术交流,我们一方面对国际上一些主要的评测活动有了更全面的了解,跟这些评测的组织者建立了沟通的渠道,同时也广泛宣传了 863 评测,扩大了 863 评测的国际影响。这些都为今后 863 评测的进一步国际化打下了很好的基础。

5 总结

本年度的评测与往年的评测相比,最主要的特点就是全面与国际主流的评测方式接轨,评测更加科学化规范化。具体来说,有以下一些特点:

(1)全面采用网上评测方式;

(2)提供了大规模的训练数据集;

(3)各项评测方法尽量向国际惯例靠拢;

(4)评测大纲制定过程中加强与参评单位和有关专家的沟通,评测大纲更加科学合理;

(5)更加重视评测研讨会的交流作用。

本次评测在国家科技部的领导下,在 863 专家组的直接指导下,在各协作单位的支持下,在各参评单位的大力配合下,取得了圆满的成功。我们希望在今后的工作中,能够继续得到各级领导、合作单位和参评单位的大力支持,使 863 评测更加科学、更加严谨、更有吸引力,使之能够对我国的各项相关研究工作产生持续长远的推动作用,并成为真正具有国际影响力的评测。

参 考 文 献:

[1] 863 评测网站[EB]:http://www.863data.org.cn,英文版:http://www.863data.org.cn/english.

[2] NIST 语音类评测网站[EB]:http://www.nist.gov/speech/tests/index.htm.
[3] NIST 机器翻译评测网站[EB]:http://www.nist.gov/speech/tests/mt/index.htm.
[4] TREC 网站[EB]:http://trec.nist.gov/.
[5] CLEF 评测网站[EB]:http://www.clef-campaign.org/.
[6] NTCIR 评测网站[EB]:http://research.nii.ac.jp/ntcir/workshop/.
[7] MUC7[EB]:http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.
[8] SIGHAN 网站[EB]:http://www.sighan.org/.
[9] 黄昌宁.统计语言模型能做什么?[J].语言文字应用,2002 年,(1):77-84.