中文信息处理技术评测综述

刘 群 钱跃良
中国科学院计算技术研究所

关键词：评测 中文信息处理

引言

中文信息处理技术涵盖的范围很广，大体来说包括语言处理技术、语音处理技术和文字处理技术等几大类。近十几年来，这些领域的研究都取得了很大的进展，并且在发展的过程中呈现出一个重要的共同特点，即公开的、定期的技术评测对这些领域的发展起到了重要的推动作用。甚至可以说，若没有这些技术评测可能就无法取得这样大的进展。

产生这一现象的原因在于这些领域的一些共同特点。众所周知，科学实验中一个重要的原则就是可重复性。任何一篇有学术价值的研究论文，其必要条件是文章中给出了足够的细节，使得其他研究者可以重复其实验结果。不过，在语言、语音和文字等计算机处理技术领域中，实验的可重复性面临着一个严重的问题—由于这个领域的实验需要使用大量的数据，这些数据不可能在一篇论文中给出来，而这些数据的采集通常具有非常大的偶然性，同一个方法在不同的数据条件下得到的结果差异可能会非常大。这样，一篇论文所介绍的研究方法对另一个研究者来说，就很难重复。比如，对于语音识别而言，影响实验结果的因素可能有说话人的性别、年龄、口音、录音的环境、录音时的噪音、话筒的质量、说话的方式（自然方式还是朗读）等。有时，人们通常容易忽略一些看起来似乎微不足道的因素，而这些因素也许恰恰会对实验结果造成重大影响。所以，如果没有共同的数据，一个研究者的实验很难被另一个研究者所重复。而不同的研究者如果采用不同的数据进行实验，其结果几乎不具备可比性。这样整个研究领域的进展就会变得非常困难。为了解决这个问题，一些公开的技术评测应运而生。

在公开的技术评测中，由评测的组织者给出共同的数据集，制定统一的测试方法和评价标准。这样，不同的研究者就可以在相同的条件下进行实验比较，并且可以重复别人的实验结果，从而得到可比的数据。这种评测受到研究人员的普遍欢迎，对研究工作起到了很好的促进作用。这种公开的、周期性的评测为中文信息处理技术领域的研究工作带来了一种新模式，而且该模式正逐渐成为这一研究领域中最主要的模式。

这种研究模式具有如下主要特点：

1. 定期举办公开的技术评测，在评测中，给定公共的训练数据集、测试数据、测试方法和评价标准；

2. 各研究单位根据评测要求自行开发系统，并提交结果；

3. 评测组织者对各单位提交的结果进行评价，并在一定范围内公开评测的结果；

4. 各参加评测的单位提交参加评测的系统报告，详细介绍系统所采用的方法；

5. 评测组织者组织研讨会，在会上参评单位报告各自参加评测的技术细节，并进行学术

交流；

6. 重复上述过程。往往上一届评测中表现出色的理论和方法会被其他研究单位所重复或者模仿，而只有当一种方法被多个研究单位重复并被证明有效时，才能被这个科研共同体所普遍接受，并逐渐淘汰一些无法重复的方法。

7. 通过反复评测，可逐步解决一些原先被认为难点较多的课题，随之而来的是科研共同体的研究兴趣将发生转移，并提出一些新的研究课题这时往往会有新的评测任务取代旧的评测任务。

8. 有时，新的评测任务并非来自研究共同体内部，而是来自外部的需求，比如政府部门或者企业界的需求。这种需求可以转化为某种评测任务，如果定义得当，这种评测可以对研究工作起到很好的促进和引导作用，甚至可能引发一个新的研究方向的诞生。

近年来，中文信息处理的相关领域研究工作取得了很大的进展，技术评测在其中起到了相当大的推动作用。本文将主要介绍与中文信息处理有关的各项评测的基本情况以及国内主要研究单位[1]参评的情况。最后进行总结和展望。

与中文信息处理相关的国际评测

随着我国经济政治影响日益提高，越来越多的与语言语音技术相关的国际评测开始涉及到对中文的处理。下面简要介绍这些与中文处理相关的国际评测。

SIGHAN评测

SIGHAN[2]是国际计算语言学会（ACL[3]）下辖的汉语处理特殊兴趣小组，SIGHAN评测是由SIGHAN举办的系列评测，也是国际评测中唯一专门关注中文处理的评测。第一届（2003年）和第二届（2005年）SIGHAN评测全称为SIGHAN汉语分词竞赛（SIGHAN Chinese Word Segmentation Bakeoff[4]），评测内容仅包括汉语词语切分。从第三届（2006年）起该项评测改名为SIGHAN汉语处理竞赛（SIGHAN Chinese Language Processing Bakeoff），评测内容增加了汉语的命名实体识别。2007年的SIGHAN评测是与中国中文信息学会中文处理评测（CIPS[5]–CLPE[6]）联合举办的，评测内容除了汉语词语切分、命名实体识别以外，还增加了汉语词性标注。

SIGHAN是国际上最有影响的汉语词法分析评测，吸引了世界上很多著名的研究机构参加，对汉语分词技术的进展起到了极大的推动作用。

在SIGHAN出现以前的汉语分词评测（主要是国内的863评测和973评测）中，分词标准一直是困扰汉语分词评测的一个重要因素，以至于实际的评测中，通常采用半自动方法进行评判，即采用人工校对与标准答案不一致的分词结果，只要没有“硬伤”，分词结果都算正确。这样得到的结论有时存在很多争议。而SIGHAN的汉语分词评测一开始就回避了这个问题。在SIGHAN评测中，评测组织方认为汉语存在多种分词标准是正常的，评测组织者并不关心分词标准问题，而只关心汉语分

[1] 本文中提到的中国的研究机构仅涉及中国大陆地区，不含港澳台地区。

[2] Special Interesting Group on Chinese Language Processing，中文语言处理特别兴趣小组。http://www.sighan.org/

[3] Association for Computational Linguistics。http://www.aclweb.org

[4] Bakeoff的本意是面包烘烤厨艺比赛，这里用来指具有竞赛性质的技术评测。

[5] Chinese Information Processing Society of China，中国中文信息处理学会

[6] Chinese Language Processing Bakeoff，汉语处理评测

词算法的改进。因此，在SIGHAN评测中，评测组织方通常同时提供多套汉语简体和繁体的训练集和测试集，并不直接提供相关的分词规范。参评单位只能根据训练语料库来自动训练词语切分算法，而不能根据规范去针对性地修改算法。好的算法应该可以自动适应不同的分词规范，只要给定相应的训练数据，就能得到一个符合该规范的词语切分软件。这种做法把研究者的注意力从对分词规范的关注中解脱出来，可以集中精力解决分词算法的自动学习问题。实践证明，这种做法发挥了很好的作用。SIGHAN评测的成功可看作是统计方法在汉语分词领域的成功。通过SIGHAN的实践证明，与传统的基于规则的方法相比，用统计方法解决汉语分词问题更加简单有效。在SIGHAN评测中，近年来大部分取得优异成绩的系统都采取了一种非常简单的思想，也就是薛念文（Nianwen Xue）等人在第一届SIGHAN评测中提出的基于汉字进行构词角色标注的方法[2]。尽管他们在第一届SIGHAN评测中并没有取得很突出的成果，但是在第二届SIGHAN评测中，采用其思想的新加坡国立大学差不多在所有项目的开放语料评测中获得了第一名，在第三届SIGHAN评测中微软亚洲研究院采用其思想也取得了很好的成绩，其他一些名列前茅的系统也几乎都采用了这种思想，只是各家采用的机器学习算法都有所不同。毫不夸张地说，通过几次SIGHAN评测，汉语词语切分技术迈上了一个新台阶。这也是研究者最希望通过评测看到的结果。在历次SIGHAN评测中，很多国内的科研机构都积极参与，其中一些科研机构如中科院计算所、哈尔滨工业大学、北京邮电大学、南京大学等还取得了很好的成绩。

NIST[7]评测中与中文信息处理相关的部分评测

美国国家标准技术研究院（NIST）在美国国防部高级研究计划署（DARPA[8]）等部门支持下，开展了一系列周期性的技术评测工作，在全球范围内吸引了大量的研究工作者参加，产生了巨大的影响。这也是到目前为止国际上影响力最大的系列评测。

美国国家标准技术研究院评测涵盖的范围很广，涉及的领域主要包括信息检索、语音识别、机器翻译、信息提取和自动文摘等，曾经组织的评测大大小小有数十个。美国国家标准技术研究院评测中很多评测都与中文信息处理有关。在美国国家标准技术研究院评测中与中文处理相关的部分主要包括以下几种：

语音识别系列评测[9] 美国国家标准技术研究院举办过多个系列各种形式的语音识别评测，其影响较大。评测内容包括广播语音、自然语音识别、对话语音识别、会议识别、说话人识别和语种识别等。不过总体上与汉语相关的项目比较少，国内参加的单位也不多，主要有声学所、北京大学、清华大学和科大讯飞等单位。

文本检索评测（TREC[10]） TREC是信息检索领域最有影响力的评测，截止到2007年已举办了16届，每次参评的课题组都超过100个（每个课题组可以有多个系统参评）。评测内容几乎覆盖了信息检索的所有领域，包括文本过滤、文本检索和问答系统等。TREC会议虽然不审稿，但由于会议论文都有具体的评测系统支持，因此其会议论文的影响也非常大。虽然在早期的TREC评测中包含关于汉语的评测，但后来因为另外两个专门的跨语言检索评测NTCIR[11]和CLEF[12]的出现，TREC决定不

[7] National Institute of Standard and Technology

[8] Defense Advanced Research Projects Agency

[9] http://www.nist.gov/speech/tests/index.htm

[10] Text REtrieval Conference, http://trec.nist.gov

再举办有关多语种的评测，目前基本上只关注英文。国内一些研究单位对TREC评测热情较高，一些单位也取得过很好的成绩。国内参加TREC评测较活跃的单位有复旦大学、清华大学、中科院计算所、中科院软件所和中科院自动化所等。

机器翻译评测（MT[13]） 美国国家标准技术研究院机器翻译评测是国际上最具指标意义、影响最大的评测。2002年，该评测在美国国防部高级研究计划署（DARPA[14]）支持的TIDES[15]项目下开始实施，每年举办一次。TIDES项目已于2006年结束，但鉴于美国国家标准技术研究院机器翻译评测影响巨大，该院已经向美国政府申请了专门的资助继续主办此项评测。正是在美国国家标准技术研究院机器翻译评测中，统计机器翻译确定了其优势地位。在这一评测的推动下，出现了一批新的模型和算法，如对数线性模型、基于短语的统计翻译模型、基于句法的统计翻译模型等。随着机器翻译水平逐年提高，沉寂已久的机器翻译研究出现了一个新的高潮。美国国家标准技术研究院机器翻译评测主要关注阿拉伯语到英语和汉语到英语的新闻领域的翻译评测，目前也开始延伸到一些其他领域和语种。美国国家标准技术研究院评测为参评单位提供了一个规模非常大的数据集，对汉英翻译而言，该数据集规模超过了500万句子对。我国的统计机器翻译研究起步较晚，但进步很快。我国的研究者主要关注汉语到英语的翻译。近年来，参加美国国家标准技术研究院评测的国内单位有中科院计算所、自动化所、软件所、厦门大学和哈尔滨工业大学等单位。其中中科院计算所在2006年的评测中第一次取得了在24个参评单位中名列第5的较好成绩，这也标志着我国的统计机器翻译研究水平已经进入国际先进行列。

信息提取评测（MUC[16]、ACE[17]） 信息提取的目的是将文本、网页等非结构化或者半结构化信息转换成数据库或者表格之类的结构化信息，这种转换可以认为需要一定程度的理解，对自然语言处理技术有较高的要求。美国国防部高级研究计划署在1986年至1998年间在TIDES项目的支持下连续7次举办了MUC会议及相关评测。该项评测将美国情报部门的需求以评测的形式进行了合理的抽象和严格的定义，也催生了信息提取这一新兴的研究领域。在MUC评测之后，美国国家标准技术研究院又开始着手推动另一项新的信息提取评测—ACE。ACE在评测任务的设置上，比MUC评测更加细致和具体。目前，ACE已经成为信息提取领域最重要的评测。ACE评测也包含了中文信息提取的评测，我国不少研究机构都参加了这项评测，如中科院软件所、中科院自动化所、北京大学和哈尔滨工业大学等，他们在其中一些项目上也取得过较好的成绩。

话题检测与跟踪评测（TDT[18]） 话题检测与跟踪的目的是从连续的文本流（如广播语音识别的结果）中，发现并跟踪话题，并将文本按照话题进行归类。这也是美国国防部高级

[11] National Center for Science Information Systems Test Collections for IR，日本国家科学资讯系统中心信息检索测试集合

[12] Cross-Language Evaluation Forum，交叉语言评测论坛

[13] NIST Machine Translation Open Evaluation, http://www.nist.gov/speech/tests/mt

[14] The Defense Advanced Research Projects Agency，美国国防部高级研究计划署

[15] Transligual Information Detection Extraction and Summarization，机器翻译语言信息探测、抽取和总结

[16] Message Understanding Conference，http://www-nlpir.nist.gov/related_projects/muc/。

[17] Automatic Content Extraction，http://www.nist.gov/speech/tests/ace/index.htm。

[18] Topic Detection and Tracking，http://www.nist.gov/speech/tests/tdt/。

研究计划署根据情报部门的实际需求抽象并提炼出来的评测任务。从1998年到2004年TDT共举办了7次，其中也涉及到汉语的话题检测任务。国内的中科院计算所等研究机构参加过TDT评测，并取得了不错的成绩。

多文档文摘评测（DUC[19]） 多文档文摘的目的是从一组描述同一话题的文档中获取一个总的文摘。尽管其方法比单文档文摘复杂得多，但它的应用范围很广泛。自2000年开始，DUC会议每年举办一次，从2001年开始每年都伴随有相应的评测。该评测已经引起了很多研究者的关注。其中，关于多文档文摘评测方法本身的研究也取得了较大进展，并且已在评测中广泛使用。因为文摘的评测与机器翻译的评测一样，都是非常困难的工作，所以不存在一种简单的方法可以直接对自动生成的文摘进行评估。DUC评测也涉及到中文，国内的一些研究单位如复旦大学、哈尔滨工业大学、北京大学和中科院计算所等参加过这些评测，并取得了不错的成绩。

其他与中文信息处理相关的国际评测

美国国家标准技术研究院评测属于官方组织的评测。此外，还有一些非官方组织的评测，同样具有非常大的影响力。下面略举几个。

在信息检索领域，除了文本检索评测以外，NTCIR[20]和CLEF[21]两个评测影响也比较大，分别由日本和欧洲举办，主要专注于跨语言检索的评测。中科院软件所、北京大学和华中科技大学等都参加过该评测，其中一些单位取得了比较好的成绩。

在机器翻译领域，除了美国国家标准技术研究院评测外，还有IWSLT[22]评测和TC-STAR[23]评测。IWSLT评测是由国际语音翻译先进研究联盟（C-STAR[24]）组织的，主要关注语音翻译。TC-STAR是欧盟的一个项目评测，已经连续举办了四届，主要关注欧洲各国语言之间的语音翻译评测，随着该项目的结束，TC-STAR评测也完成了其历史使命。中科院计算所和中科院自动化所在这两项评测中分别取得过很好的成绩。

在语言分析领域，比较著名的评测是CoNLL会议[25]的共享任务（Shared Task）。CoNLL会议每年都要定义一些共享任务，虽然这些任务不一定与往年的相同，但具有一定的连续性。从1999年开始，CoNLL会议陆续开展了名词短语划分、组块划分、命名实体识别、语义角色标注和依存分析等评测任务。其中大部分项目的评测都涉及包括汉语等在内的多种语言。中科院自动化所和哈尔滨工业大学分别在依存分析和语义角色标注评测中取得过好成绩。另外，由SemEval[26]会议主办的系列语义技术评测（前三次名称是Senseval会议）也受到了人们的关注。SemEval评测包含了对中文语义标注的评测，哈尔滨工业大学和北京大学分别参加过Senseval-3（2004年）和SemEval-2007的汉语评测项目的组织工作，其中哈尔滨工业大学在SemEval-2007评测中取得了很好的成绩。

---

[19] Document Understanding Conference，http://duc.nist.gov/。

[20] NII Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/

[21] Cross-Language Evaluation Forum, http://www.clef-campaign.org/

[22] International Workshop on Spoken Language Translation, http://iwslt07.itc.it/ (IWSLT-2007)

[23] Technology and Corpora for Speech to Speech Translation, http://www.tc-star.org/

[24] Consortium for Speech Translation Advanced Research

[25] Conference on Computational Natural Language Learning, http://www.cnts.ua.ac.be/conll/ (CoNLL-7)

[26] International Workshop on Semantic Evaluations, http://nlp.cs.swarthmore.edu/semeval/ (SemEval-2007)

国内的中文信息处理评测

国家863中文信息处理与智能人机接口评测

我国最有影响的系列中文信息处理技术评测是国家863计划组织的中文信息处理与智能人机接口评测，简称863评测[27]。该项评测涵盖了中文信息处理和人机交互技术的大部分研究领域。

863评测起步很早。在1990年，国家863计划智能机主题（现为软硬件主题）专家组就开始酝酿通过公开的评测活动，对相关的研究进行评价并促进研究的发展。1990年依托中科院计算所进行了一次预演性质的语音识别评测，共有5个系统参加。1991年进行了第一次正式评测，有语音识别和汉字识别两个类别，16个系统参加。截止到1998年，共进行了5次正式的863评测。这些评测由863计划智能机主题专家组出面组织，全国很多高校和研究机构共同参与。在863专家组的支持下，该项评测在停顿5年之后于2003～2005年恢复举行。其主办方为中科院计算所，国内外一些高校和研究机构参与协办。到2005年为止，863评测共举办了8届。其中，2003年和2004年的863评测是和北京市政府支持的“多语言奥运信息服务系统”项目评测联合举办的。

历届863评测曾经设置的类别包括：

- 语音识别（Automatic Speech Recognition，ASR）；
- 语音合成（Machine Translation，TTS）；
- 机器翻译（Machine Translation，MT）；
- 汉语分词（Chinese Word SEGmentation，SEG）（含词性标注和命名实体识别）；
- 信息检索（Information Retrieval，IR）；
- 文本分类（Text Categorization，TC）；
- 文本摘要（Text Summary，TS）；
- 文字识别（Character Recognition，CR）；
- 人脸检测与识别（Face Recognition，FR）。

863评测对我国中文信息处理的研究起到了有力的推动作用。863评测受到863计划的支持，其参评单位几乎包括了我国相关研究领域

表1 历次863评测的类别数和参评系统数

年份	1990	1991	1992	1994	1995	1998	2003	2004	2005
届别	预备	1st	2nd	3rd	4th	5th	6th	7th	8th
评测类别数	1	2	2	4	6	6	8	8	3
参评系统数	5	16	17	39	65	43	46	113	45

表2 历届863评测类别设置详情

	1990	1991	1992	1994	1995	1998	2003	2004	2005
语音识别	√	√	√	√	√	√	√	√	√
语音合成				√	√	√	√	√	
机器翻译				√	√	√	√	√	√
汉语分词					√	√	√	√	
信息检索							√	√	√
文本分类							√	√	
文本摘要					√	√	√	√	
文字识别		√	√	√	√	√	√		
人脸检测与识别								√	

[27] http://www.863data.org.cn

所有的著名研究单位，结果也反映了我国在该领域内的最高水平。在国家863计划的持续支持下，我国中文信息处理领域涌现了一大批具有世界先进水平的研究成果，如中科院自动化所的“汉王”和清华大学“文通”的汉字识别技术、中科院“华建”和中软公司“译星”的机器翻译技术、中国科技大“讯飞”的语音合成技术、中科院自动化所和声学所的语音识别技术、北京信息工程学院TRS的信息检索技术等，这些成果绝大部分都经历过863评测的洗礼和考验。其中很多技术获得过国家级的科技奖励，还有一些成功转化成产品，成为国内甚至国际同类产品中的佼佼者。

著名的自然语言处理专家黄昌宁教授曾经这样说过：“国家863计划智能计算机专家组曾对语音识别、汉字（印刷体和手写体）识别、文本自动分词、词性自动标注、自动文摘和机器翻译译文质量等课题多次进行过统一测试数据和统一计分方法的全国性评测，对这些领域的技术进步起到了积极的推动作用。但是，这期间也遇到了一些阻力，有些人试图用各种理由来抵制这样的统一评测，千方百计用‘自评’来取代统评。其实，废除统一的评测就等于丧失了可比的基础。这种损失将让任何理由都变得异常苍白。”[1]

其他国内中文信息处理评测

除了863评测外，国内有影响的中文信息处理评测还比较少。

在2001～2003年期间，受973 “图像、语音、自然语言理解与知识发掘” 项目委托，东北大学姚天顺教授组织了3次评测。第一次评测只包括汉语词法分析，第二次评测增加了汉语句法分析，第三次评测又增加了汉英机器翻译。由于这几次评测仅局限于该973项目的参加单位，并没有对外公开征集报名，因此评测的语料和结果也是不公开的。

另一个有一定影响的评测是由北京大学网络与分布式系统实验室主办的全国搜索引擎和网上信息挖掘学术研讨会系列评测。该项评测从2003年开始举办，具体又分为中文万维网信息检索评测、中文网页分类评测和垃圾邮件过滤评测等项目。该评测每年举办一次，其评测结果在研讨会上报告并交流，吸引了国内的一些相关研究单位参加。全国搜索引擎和网上信息挖掘学术研讨会评测为参评单位提供了一个100G（后来扩充到200G）规模的网页文档数据集，这是评测的重要基础。该数据集也被2004年和2005年的国家863信息检索评测所采用。

中文信息学会系列评测

进入“十一五”以后，863计划暂时没有对中文信息处理方面的技术评测提供直接支持。为了保持国内中文信息处理评测的连续性，中文信息学会语言资源建设和管理工作委员会决定举办“中文信息学会系列评测”。中文信息学会语言资源建设和管理工作委员会于2007年开始筹备成立评测工作组，并协调有关“中文信息学会系列评测”的组织工作。

2007年，中文信息学会系列评测已经举办了两次评测。一次是中文信息学会SSMT[28] 2007机器翻译评测，该评测由“第三届统计机器翻译研讨会（SSMT2007）”负责组织，已于2007年7月顺利举行，相关的SSMT2007研讨会于8月初在哈尔滨举行，总共有11个参评单位参加了5个项目的评测，其中有一家参评单位来自日本。另一次是第一届中国中文信息学会汉语处理评测（CIPS-CLPE2007），本次评测与第四届SIGHAN汉语处理评测联合举办，评测已于2007年10月顺利举行，相关的研讨会将于2008年1月与SIGHAN一起在印度海德拉巴举行。

与以往的863评测不同，“中文信息学会

[28] Symposium on Statistical Machine Translation

系列评测”不是官方组织的评测，而是由学术界自身倡导并组织的评测。我们希望这一系列评测将采用比以往更加开放、更加民主的做法，能够真正团结国内的中文信息处理研究人员，办成自己主导的具有国际影响力的中文信息处理技术评测。

总结与展望

从上面的介绍可以看到，目前国内外都开展了大量的与中文信息处理相关的技术评测。应该承认的是，在某些领域，无论从规模还是深度上说，一些国际评测的影响都远远超过了国内的相关评测。国际评测在项目设置、评测组织以及评测后的学术交流等方面，都有很多值得我们学习的地方。

早期的国内评测比较重视应用，很多863评测项目的设置都有很好的应用背景，从而大大促进了很多研究成果走向应用市场，这是一个很好的传统。与国际评测相比，在某些领域，国内评测更占有优势并独具特色。比如863评测中的语音识别评测，在汉语语音识别方面，比美国国家标准技术研究院评测的内容更加丰富，而且在噪声场景的语音识别评测方面做得很有特色。863的汉字识别评测和汉语语音合成评测都在国际上占有很强的优势。此外，国内的机器翻译评测虽然影响和规模不如美国国家标准技术研究院的评测，但在数据的多样性和语种的多样性方面都具有自己的特色，也吸引了一些国外的研究机构参与。相对而言，在科学问题的提炼方面国内评测总体上做得不够。还有，在通过评测促进学术思想交流和学术研究进步方面，与国际评测相比相对不足，这些都是值得改进的地方。

虽然国际上对中文信息处理越来越重视，但对国际评测来说，中文信息处理评测只是其中一部分内容。而对我国的研究者来说，中文信息处理则几乎是他们研究内容的全部。因此，搞好中文信息处理的评测，对我们来说是义不容辞和当仁不让的。我们期望，国内的评测能越办越好，能够在中文信息处理评测方面建立自己的评价体系和标准（Benchmark），并能够得到国际研究界普遍采用和承认。未来，希望有更多的中国学者能够在各种国际评测中发挥其作用，并在国际上的中文信息处理评测中取得更好的成绩。

致谢

本文在写作过程中得到了中科院自动化所丁鹏博士、赵军博士、中科院软件所孙乐博士、清华大学周强博士和北京大学吴云芳博士的帮助，负责组织本期中文信息处理特辑的刘挺博士也给出了很多有益的建议，特此表示感谢。

刘群
中国计算机学会高级会员、名词工作委员会委员。中国科学院计算技术研究所研究员。研究兴趣：自然语言处理、机器翻译。

钱跃良
中国计算机学会高级会员。中国科学院计算技术研究所正高级工程师，多语言交互技术实验室主任。研究兴趣：中文信息处理、智能人机交互。

参考文献

[1] 黄昌宁. 统计语言模型能做什么. 语言文字应用, 2002,1: 77 ~ 84
[2] Nianwen Xue and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging, In Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03. Sapporo, Japan