ACL-IJCNLP 2009 评述

王海峰

东芝（中国）研究开发中心　北京　100738

wanghaifeng@rdc.toshiba.com.cn

1 概况

计算语言学学会（The Association for Computational Linguistics）旗下的 ACL 年会是国际上最有影响的计算语言学及自然语言处理学术会议，而 IJCNLP 则是亚洲自然语言处理联合会（Asian Federation of Natural Language Processing）的旗舰会议。2009 年，第 47 届 ACL 和第 4 届 IJCNLP 的联合会议（ACL-IJCNLP 2009）于 8 月 2 ~ 7 日在新加坡成功召开，这是 ACL 历史上第三次在亚洲召开（前两次分别为 2000 年在中国香港、2003 年在日本札幌）。除 ACL-IJCNLP 主会议外，还有一个会议（Conference）、12 个 专题研讨会（Workshop）和 6 个专题讲座（Tutorial）等同期举行（详细列表见附录）。

令会议组织者及所有参会学者欣慰的是，虽然有国际金融危机和甲型 H1N1 流感全球大流行等因素影响，今年的 ACL-IJCNLP 仍然备受关注，有来自 44 个国家和地区的 800 余人参加了本次会议，参会人数仅次于在东欧旅游胜地布拉格举行的 ACL 2007 列历史第二位。其中，来自中国大陆的学者 65 人，在美国和东道主新加坡之后列第 3 位，这一参会人数超过之前中国大陆参会学者最多的 COLING/ACL 2006 一倍有余。

令国内同行欣喜的是，除参加会议和发表论文外，中国大陆学者还对会议有了更为全面的参与。例如：中科院自动化所宗成庆担任了专题讲座主席（Tutorial Chair），中科院计算所刘群和微软亚洲研究院林钦佑分别担任了分会场主席（Session Chair），笔者担任了领域主席（Area Chair）、分会场主席和最佳论文委员会（Best Paper Committee）成员，来自北京大学、清华大学、哈尔滨工业大学、中科院计算所和自动化所、微软亚洲研究院及东芝（中国）研发中心等单位的多位学者担任了程序委员会成员（PC Member）等。

2 论文

每年的 ACL 论文，尤其是主会议长文（Full Paper），基本反映了自然语言处理领域最新研究进展、代表了本领域最高研究水平。今年的主会议论文投稿数达到了创纪录的 925 篇，其中长文 569 篇、短文 356 篇。被录用的长文有 120 篇，录用率 21.1%；短文被录用 93 篇，录用率 26.1%。以下着重介绍最受关注的主会议长文的情况（各项统计数据按第一作者计算）。从地域分布来看，北美仍以 56 篇论文居于领先地位，亚太地区以 46 篇居第二位，其它 18 篇来自欧洲，而南美等地区虽也有少量投稿，但未能被录用。从国家分布来看，美国的 50 篇占有绝对优势，而中国大陆则以 15 篇在世界上居于第二，其它发表论文较多的国家有日本、新加坡、德国、加拿大等。中国大陆的研究机构中，近年发表 ACL 论文较多的微软亚洲研究院、东芝（中国）研发中心、中科院计算所及哈尔滨工业大学继续保持良好势头，再次分别发表多篇论文。同时，北京大学、清华大学、南开大学、苏州大学、中科院自动化所等单位也有斩获。个人方面，继去年哈工大博士生赵世奇以第一作者身份发表两篇 ACL 主会议长文后，今年计算所刘洋博士也以第一作者身份独中两元，充分展示了我国年轻学者的实力。回顾中国大陆学者过去几年参与 ACL 的情况，记得 ACL 2005 时，来自中

国大陆的主会议论文还只有 3 篇、参会学者才仅有 4 位。今天的繁荣与几年前的冷清形成了鲜明对比，这些年的长足进步实在值得欣喜。

今年的 ACL 论文分为 14 个大领域，各领域的主会议长文的分布情况如下表。

领域	投稿数	录用数	录用比例
Phonology, Morphology, Segmentation, POS, Chunking	31	8	25.8%
Syntax and Parsing	49	14	28.6%
Semantics	67	14	20.9%
Discourse, Dialogue and Pragmatics	43	9	20.9%
Summarization and Generation	44	8	18.2%
Statistical and Machine Learning Methods	40	6	15.0%
Machine Translation	82	23	28.0%
Information Retrieval	28	4	14.3%
Information Extraction	49	10	20.4%
Sentiment Analysis, Opinion Mining, Classification	45	7	15.6%
Spoken Language Processing	19	6	31.6%
Question Answering	25	3	12.0%
Text Mining and NLP Applications	21	4	19.0%
Language Resource	26	4	15.4%
总计	569	120	21.1%

从上表可以看出，近年来广受重视的机器翻译（Machine Translation）研究仍然保持大热，投稿量和论文录用数量都大幅超过其它领域。从机器翻译领域研究内容来看，解码（Decoding）受到了比以往更多的重视，从提高解码的性能、速度到在解码器上做系统融合，都有一些很好的结果产生，而本领域推荐的唯一最佳论文候选（Variational Decoding for Statistical Machine Translation）也是关于解码的。其它研究较多的内容还包括基于语法的翻译、对齐、评价等。值得一提的是，华人学者贡献了 23 篇机器翻译长文中的 11 篇（其中 5 篇来自中国大陆），几乎占据了半壁江山。而笔者受邀担任机器翻译领域的两主席之一（另一位是 The University of Edinburgh 的 Philipp Koehn）、计算所刘群与笔者分别担任机器翻译分会场主席等，相信也都与华人学者在该领域的整体实力及贡献的提高有关。

除了机器翻译，其它一些传统方向，如语义、句法与分析等也占有较大比重。而句法与分析领域更是由于最高录用比例（28.6%）和 2 篇最佳论文而受到关注。

总体看来，自然语言处理领域的各个方向都处于平稳进步的阶段。其中一个值得注意的现象是，算法的时间效率问题受到了广泛关注。记得一位本领域的国际著名学者曾说过："我对超过多项式时间复杂性的算法毫无兴趣。"姑且不讨论这句话的是或非，但算法效率的确非常重要。ACL-IJCNLP 2009 上，Stanford 的 Michel Galley、Berkeley 的 John DeNero、Google 的 Shankar Kumar 等很多学者都在提高算法效率方面做了很好的工作，包括分析领域的 2 篇最佳论文在内的更多论文也都讨论了效率问题。事实上，在真实应用环境下，算法的效率何止重要，它甚至是决定性的。相信在很多研究致力于提高诸如翻译质量、分析准确率、生成正确率等性能指标的同时，提高算法效率方面也一定大有可为。

3 奖项

每年的 ACL 都会颁发两个极受关注的奖项：终身成就奖（Lifetime Achievement Award）

和最佳论文奖（Best Paper Awards）。

终身成就奖是对曾对本领域做出过卓越贡献的学者的终极承认与褒奖，是本领域颁发的最高个人荣誉。约翰霍普金斯大学（Johns Hopkins University）的 Frederick Jelinek 因在语音识别及机器翻译等领域的卓越贡献而获得今年的终身成就奖，其获奖演讲的题目是：The Dawn of ASR and MT。

最佳论文则基本上代表着当年本领域最好的研究成果，也一定程度上预示着未来的发展动向。今年，各个领域主席共推选出 6 篇最佳论文候选，最后，最佳论文委员会从中选择了 3 篇作为最佳论文。其中，有 2 篇来自传统的句法与分析领域：CMU 的 Andre Martins 等人的“Concise Integer Linear Programming Formulations for Dependency Parsing”和 Berkeley 的 Adam Pauls 等人的“K-Best A* Parsing”，另一篇是 MIT 的 S.R.K. Branavan 等人的“Reinforcement Learning for Mapping Instructions to Actions”。三篇最佳论文全都来自美国的一流名校，也从一个侧面反映出这些名校的确名不虚传。

有趣的是，终生成就奖得主 Frederick Jelinek 和最佳论文得主之一 S.R.K. Branavan 不约而同地提到了投稿被拒的往事。Frederick Jelinek 在回顾上世纪八十年代后期他所在的 IBM 团队创立统计机器翻译方法的历程时提到，他们的第一篇统计机器翻译论文曾被 COLING 拒稿。无独有偶，S.R.K. Branavan 在获奖感言中也提到她的获奖论文曾被先前召开的 NAACL HLT 2009 拒稿，引得现场一片哄堂大笑。可见，投稿被拒并不可怕，只要坚定信心继续努力，是金子总会发光的。

4 结束语

除了上述中国大陆学者的情况外，来自港台、新加坡、日本及欧美各国的华人学者也在 ACL-IJCNLP 2009 上展示了强大实力，本届大会主席（General Conference Chair）苏克毅、程序委员会主席（Program Chair）之一苏俭、组委会主席（Local Organizing Chair）李海洲等就是其中的杰出代表。笔者相信，中国及全世界华人学者的水平仍将继续提高、贡献势将持续增加、影响也必将不断扩大。

致谢

感谢孙乐、宗成庆、刘群等老师为本文提出修改意见。

附录　同期举行的会议、专题研讨会及专题讲座列表

[会议]
Conference on Empirical Methods in Natural Language Processing (EMNLP) 2009
[专题研讨会]
WS1: Applied Textual Inference (TextInfer)
WS2: Grammar Engineering across Frameworks
WS3: Knowledge and Reasoning for Answering Questions (KRAQ'09)
WS4: Language Generation and Summarisation
WS5: Multiword Expressions: Identification, Interpretation, Disambiguation and Applications
WS6: The Third Linguistic Annotation Workshop (The LAW III)
WS7: 2nd Workshop on Building and Using Comparable Corpora: from parallel to non-parallel corpora
WS8: TextGraphs-4: Graph-based Methods for Natural Language Processing
WS9: The People's Web meets NLP: Collaboratively Constructed Semantic Resources

WS10: The 7th Workshop on Asian Language Resources (ALR7)
WS11: Named Entities Workshop - Shared Task on Transliteration (NEWS on Transliteration)
WS12: Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL)
[专题讲座]
T1: Fundamentals of Chinese Language Processing (Chu-Ren Huang and Qin Lu)
T2: Topics in Statistical Machine Translation (Kevin Knight and Philipp Koehn)
T3: Semantic Role Labeling: Past, Present and Future (Lluís Màrquez)
T4: Computational Modeling of Human Language Acquisition (Afra Alishahi)
T5: Learning to Rank (Hang Li)
T6: State-of-the-art NLP Approaches to Coreference Resolution: Theory and Practical Recipes (Simone Paolo Ponzetto and Massimo Poesio)