文章编号: 1003-0077(2007)05-0136-03

ACL2007 会议观感

刘群, 刘洋

(中国科学院 计算技术研究所,北京 100080)

每年一度的计算语言学学会年会(Annual Meeting of ACL)是计算语言学界的盛会,也是计算语言学和自然语言处理领域最有影响的学术会议,ACL 每年发表的论文都反映了这一领域的最新研究进展和学术动向,受到研究工作者的广泛重视。今年的 ACL2007 是 ACL 的第 45 届年会,在美丽而又浪漫的东欧古城——捷克首都布拉格召开,同时召开的有 EMNLP-CoNLL 2007 和 IWPT 2007 等 2 个学术会议(Conference)、15 个学术研讨会(Workshop)以及 5 个专题讲座(Tutorial)。会期从 6 月 23 日开始一直持续到 6 月 30 日结束,根据 ACL2007 网站公布的名单,参加会议的人数达到了创纪录的 1 095 人,可谓盛况空前。ACL 主会议于 25～27 日召开,其他各个会议、研讨会和专题讲座都是安排在 ACL 主会议之外的时间段并行进行的。其中,EMNLP 于 28～30 日召开,持续时间最长。其他研讨会或者专题讲座持续时间半天到两天不等。

今年的 ACL 会议一共收到论文 588 篇,录用 131 篇,录用率为 22.3%。各领域论文分布如下:

领 域	收到论文数量	录用论文数量	录用百分比
篇章、对话和语用	58	16	28
句法、句法分析和形式语法	63	15	24
词典、本体和资源建设	63	11	17
词法相关	60	13	22
自然语言处理中的机器学习和算法	67	16	24
信息提取、信息检索和问答	70	15	21
机器翻译	65	15	23
语音、语言模型	41	9	22
语言生成和文摘	48	10	21
语义相关	54	12	22
总计	588	131 ①	22.3

(以上资料来自大会公布的 Newsletter No. 6)

以上的领域划分比较粗,实际上,会议的 Session 划分更细一些,包括:机器翻译、词义消歧、句法分析、语言资源、语法、机器学习方法、语音与形态、语义角色标记、生成、多语言、多模态、情感、复述与文本推理、篇章与对话、语言模型、同指、文摘、语义关系、问答、序列处理、词典和词汇语义、文本挖掘与检索。每个 Session 通常有 4 篇论文。我们可以看到,其中机器翻译是目前最大的研究热点,安排的 Session 达 4.5 个(其中一个 Session 是和 Multilingualty 合并举行的)。而在机器翻译中,统计机器翻译研究又占据了主导地位,纯粹基于规则或者基于实例方法的论文几乎没有。统计机器翻译的文章全面开花,一些传统的基于语言知识的规则方法也开始融入到统计机器翻译研究中,很多论文都试图利用句法结构、句法标记、功能词、词义排歧来改善统计机器翻译的效果。另外,统计机器翻译的一些经典问题,如词语对齐、自动评价、词序调整,依然受到重视。其他一些统计机器翻译研究论文涉及的话题包括解码算法、音译、翻译预处理、系统融合、多语言等。不过,总体上看,这次会议上统计机器翻译论文并没有出现突破性的研究成果。另外,深层次的句法语义分析,也仍然是 ACL 关心的主要话题,论文数量占了相当的比例,而且,今年的 EMNLP-CoNLL 重点讨论了依存分析技术,ACL 主会议上有专门的 Session 讨论语义角色标注,另外有一个 SemEval Workshop 开展了语义分析方面的评

① 大会 Newsletter No. 6 公布的数字中,总录用论文数是 131 篇,但各领域论文数之和是 132 篇,二者不符。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

测。除此之外,其他几个比较热门的研究领域包括语音与形态、篇章与对话、多语言、多模态、语言模型、机器学习方法、情感等,这些领域分别有两个或两个以上的 Session。

这次会议收到的论文来自 45 个国家,其中亚洲太平洋地区和欧洲各占 35%,北美洲 27%,中东 2%,南美和非洲都不到 1%。从被录用论文所占比例来看,北美最高,为 46.6%,欧洲是 31.3%,亚洲和太平洋地区为 16.8%。一些主要的国家和地区论文录用情况如下:

国家和地区	投稿论文数	录用论文数
澳大利亚	18	4
中国大陆	74	6
香港	10	0
印度	10	0
日本	53	3
新加坡	10	8
中国台湾	20	1
法国	21	4
爱尔兰	10	5
德国	32	6
意大利	10	3
荷兰	11	3
西班牙	20	0
英国	49	14
加拿大	11	4
美国	146	57
以色列	10	6

(以上资料来自大会程序委员会主席的报告,仅列出投稿论文 10 篇以上国家和地区)

从中我们可以看到,研究实力最强的国家仍然是美国。不仅投稿数量最多,而且录用率也相当高。纯粹从录用率看,新加坡达到 80%,是最高的,欧美国家的录用率普遍较高。中国大陆和日本投稿论文数量都很多,但录用率偏低。以中国大陆学者为第一作者的论文一共有 6 篇,作者单位分别是:微软亚洲研究院(与哈尔滨工业大学合作)、东芝(中国)研究开发中心、北京大学、哈尔滨工业大学(与清华大学、微软亚洲研究院、新加坡国立大学合作)、中国科学院计算技术研究所、重庆大学(与微软亚洲研究院合作)。虽然来自中国大陆本土的论文并不多,不过,会上的中国学者并不少,很多来自世界各地的中国学者以海外研究者的身份发表了论文。从会议录用的论文列表上也可以看到,其中中文姓名的作者占据了相当的比例。在参会的人员中,根据我们自己的统计,来自中国大陆的学者有 20 多位,而根据会议 Newsletter 公布的资料,来自中国大陆的学者有 31 位。海外华人学者人数更多,我们在会上就见到很多来自世界各大著名高校的中国学者。根据大会 Newsletter 公布的资料,所有参会人员中中文姓名的人数大约在 110 人左右。记得黄昌宁教授在 1998 年参加 COLING-ACL 会议回国后写的一篇介绍文章中还感慨会上很难见到中国学者的身影,如今这种状况已经完全改变。特别可喜的是,一些年轻的中国学者脱颖而出,相信假以时日,中国学者在这一领域的影响必将越来越大。

按照惯例,每次 ACL 会议都要颁发一个终身成就奖和一个最佳论文奖。本次会议的终身成就奖颁发给了来自 Palo Alto Research Center (PARC) 的学者 Lauri Karttunen,而最佳论文奖颁发给了来自美国德克萨斯大学的一篇论文 Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus,其第一作者 Yuk Wah Wong 也是一个华人学者。另外,会议还举办了丰富多彩的各种活动,包括一次招待餐会、一次正式宴会、一次音乐会和多次参观旅游,这些活动都是大家自愿参加,有些是收费的。布拉格是个非常美丽的历史名城,在古罗马时期曾经是欧洲第二大城市,位于波西米亚平原上,盛产水晶。会议召开的时候正值布拉格在举行音乐节活动,各种专场音乐会应接不暇,各种公开场合也有很多露天音乐会,以及一些中小学生的公开表演,让人感觉到浓浓的音乐氛围,体会到音乐已经深深扎根在这个民族的文化传统之中。

另外,除主会议外,同样值得关注的还有同期举行的卫星会议,其中比较重要的是 EMNLP-CoNLL 2007。EMNLP (Conference on Empirical Methods in Natural Language Processing) 每年召开一次,由 ACL 关于语言学数据和基于语料库方法的自然语言处理特别兴趣小组 SIGDAT 资助并主办。本次会议与 CoNLL (Conference on Computational Natural Language Learning) 联合召开,程序委员会主席是约翰霍普金斯大学的 Jason Eisner。EMNLP-CoNLL 2007 今年共收到论文 398 篇,录用 109 篇(录用率为 27%,其中 66 篇为大会报告论文,43 篇为张贴论文),覆盖了机器翻译、问答、文本生成、文本摘要、语义、文档分析等多个研究领域。其中第一作者来自中国大陆学者的论文一共有 9 篇①,作者单位分别是:东北大学(与微软亚洲研究院合作),东北大学(与美国南加州大学合作)、东芝(中国研究开发中心、南开大学(与微软亚洲研究院合作)、中国科学计算技术研究所、中国科学院自动化研究所、微软亚洲研究院、苏州大学(与新加坡 I2R 研究所合作)、清华大学。

总体上,这次会议非常成功。本届 ACL 执行委员会主席 Mark Steedman 在会议组织的宴会上做了一个精彩报告,对计算语言学近年来的进展和发展趋势做了一个很好的总结。他说,在 2006/2007 年间,计算语言学团体变大变强了,最有力的证明就是我们已经有了高可读性、具有实际用途的

① 由于 EMNLP-CoNLL2007 的论文集没有区分口头报告论文,我们这里就一并计算了。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

基于统计方法的阿拉伯语到英语的机器翻译系统,其他一些语言的翻译和其他方面的应用也取得了很大的进展。计算语言学一个好的趋势是和语音研究团体的联系更加紧密了。但也有一个令人担心的现象,就是我们和语言学家变得疏远了。他表示,希望计算语言学能够像物理学一样赢得人们的尊敬。

物理学之所以重要的原因是:

Everything is made of particles, so Physics is very important.

现在我们同样可以说:

The World-Wide-Web is made of Language, so Computational Linguistics is very important.

致谢:

感谢孙茂松、王海峰、赵军、车万翔给本文提出的修改意见。感谢王海峰给作者转发了 ACL Newsletter No. 6,使作者得以修正了原文中的一些数字。感谢《中文信息学报》编辑部将本文初稿转发给上述老师修改补充,使本文更加完美。

(上接第 100 页)

每个词元的标注句子链接(这些句子就是各种语义结构模式的具体实现,都标注了语义信息)、框架元素的句法实现方式汇总表、词元的语义搭配模式汇总表、框架 - 框架关系图示报告,也可以进行框架元素、短语类型和句法功能的分别提取以及演示等。

4 结语

CFN 构建工程量大,难度很高,目前课题组只能在有限领域下开展工作,但是,已经探索了一条可行的技术路线,取得了阶段性成果,为实现语义 Web 中的语义知识共享以及智能化、个性化的 Web 服务提供了基础资源。近期我们计划扩展到其他领域,继续构建一批框架,并将其应用到阅读理解式问答系统和智能搜索系统等应用研究中,以检验 CFN 的实践效果。研究小组近期正在进行基于 CFN 的句法语义角色自动标注软件的设计与开发,目的是开发高性能的汉语句法语义分析器,为进行大规模真实文本的语义信息标注提供有力支持。

参考文献:

[1] 刘开瑛,由丽萍. 汉语框架语义知识库构建工程[A]. 中文信息处理前沿进展,中国中文信息学会成立二十五周年学术会议论文集[C]. 2006, 11: 64-71.

[2] Charles J. Fillmore. Frame semantics and the nature of language [A]. In: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech[C]. 1976, 280: 20-32.

[3] Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical data bank which provides deep semantics[A]. In: Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation[C]. Hong Kong: 2001, 3-26.

[4] Charles J. Fillmore, Collin F. Baker et al. The Berkeley FrameNet project [A]. In: Proceedings of COLING/ACL[C], Montreal, Canada: 1998. 86-90.

[5] Collin F. Baker, Charles J. Fillmore and Beau Cronin. The Structure of the Framenet Database [J]. International Journal of Lexicography, 2003, 16(3): 281-296.

[6] Charles J. Fillmore, Collin F. Baker and H. Sato. The FrameNet Database and Software Tools[A]. In: Proceedings of the Third International Conference on Lan- guage Resources and Evaluation [C]. Las Palmas, Spain: 2002.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net