**初学者如何查阅自然语言处理学术资料(2016修订)**

**说明**：这篇文章大约是我在2013年给组内新生写的一篇查阅文献的入门文章。几年时间过去，发生了很多变化。所以这里做一次系统的修订，再次分享给大家，希望能够帮助有兴趣的同学更快地进入NLP领域前沿。

最近实验室刚进组的一位同学发邮件来问我如何查阅学术论文，这让我想起自己刚读研究生时茫然四顾的情形：看着学长们高谈阔论领域动态，却不知如何入门。经过研究生几年的耳濡目染，现在终于能自信地知道如何了解最新科研动态了。我想，这可能是初学者们都会有的困惑，与其只告诉一个人知道，不如将这些经验写下来，来造福更多同学吧。当然，这个总结不过是一家之谈，只盼有人能从中获得一点点益处。受个人认知所限，难免挂一漏万，望大家海涵指正。

1. 国际学术组织、学术会议与学术论文

自然语言处理（natural language processing，NLP）在很大程度上与计算语言学（computational linguistics，CL）重合，是计算机科学与语言学的交叉学科，也是人工智能的重要方向。与其他很多计算机分支方向类似，由于技术发展迅速，NLP/CL重视学术会议胜过学术期刊。由于发表周期短，并可以通过会议进行交流，绝大多数最新的重要科研进展都首先发表在学术会议上，而学术期刊更多用来发表会议论文的扩展整理版本，起到归档作用。

NLP/CL领域重要的国际学术会议均由国际专业学会发起和举办。NLP/CL最权威的国际专业学会是The Association for Computational Linguistics（ACL，URL：http://aclweb.org/），这个学会主办了NLP/CL历史最悠久也最权威的国际会议，即ACL年会。ACL学会还会在北美和欧洲召开分年会，分别称为NAACL和EACL。此外，ACL学会下设多个特殊兴趣小组（special interest groups，SIGs），性质类似大学校园的兴趣社团，聚集了NLP/CL不同子领域的学者。其中比较有名的包括SIGDAT（Linguistic data and corpus-based approaches to NLP）、SIGNLL（Natural Language Learning）等。这些SIGs也会组织一些国际学术会议，其中比较有名的就是SIGDAT组织的EMNLP（Conference on Empirical Methods on Natural Language Processing）和SIGNLL组织的CoNLL（Conference on Natural Language Learning）。此外，还有一个全称International Committee on Computational Linguistics的老牌NLP/CL学术组织，每两年组织International Conference on Computational Linguistics（COLING）国际会议，也是NLP/CL的重要学术会议。NLP/CL领域的绝大部分重要进展都发表在这些学术会议上。

从论文水平和关注度来看，一般公认ACL、NAACL、EMNLP、COLING为NLP/CL最重要的四大顶级学术会议。其中ACL是公认最权威的NLP/CL会议，参会人数也最多。ACL 2015在北京举办，参会人数为1200；ACL 2016在德国举办，参会人数达到1600多人。美国是NLP/CL重镇，因此NAACL的论文质量也非常高。EMNLP是相对年轻的学术会议，创立于1995年前后。创立EMNLP的目的就是关注统计机器学习方法在NLP/CL中的应用。因此，随着最近10余年互联网高速发展，基于大规模数据的机器学习方法特别是近年来深度学习大行其道，最近几届EMNLP发展很快，例如EMNLP 2016有800余人参加。COLING是老牌的NLP/CL的学术会议，历史非常悠久。也许由于每两年召开一次的传统，似乎不太适应最近NLP/CL技术高速发展的特点，近年来COLING论文关注度有较为明显的下降，不过仍然是NLP/CL值得关注的重要会议。

作为NLP/CL领域的学者最大的幸福在于，ACL学会网站建立了称作ACL Anthology的页面（URL：http://aclanthology.info/），为其主办的所有国际学术会议论文提供免费下载，甚至也囊括了其他组织主办的学术会议，例如COLING、IJCNLP等。该平台还支持基于Google的全文检索功能，可谓一站在手论文我有。

NLP/CL的旗舰学术期刊是Computational Linguistics（http://www.mitpressjournals.org/loi/coli），发表过很多经典学术论文。该期刊每期只有几篇文章，平均质量高于会议论文，而且一般内容比较全面详细，值得认真研习。特别是其中的综述文章，是快速全面了解相关NLP/CL任务的最好文献。此外，ACL学会为了提高学术影响力，也创办了论文要求篇幅相对较短的Transactions of ACL（TACL，URL：http://www.transacl.org/），值得关注。值得一提的是，这两份期刊也都是开放获取的。此外也有很多与NLP/CL有关的期刊，如ACM Transactions on Information Systems，ACM Transactions on Speech and Language Processing，ACM Transactions on Asian Language Information Processing，Journal of Quantitative Linguistics等等。

根据Google Scholar Metrics 2016年对NLP/CL学术期刊和会议的评价，ACL、EMNLP、NAACL、LREC、Computational Linguistics位于前5位，COLING排在第9位，基本上反映了本领域学者的关注程度。

NLP/CL作为交叉学科，其相关领域也值得关注。主要包括以下几个方面：（1）信息检索和数据挖掘领域。相关学术会议主要由美国计算机学会（ACM）主办，包括SIGIR、WWW、KDD、WSDM等；（2）人工智能领域。相关学术会议主要包括AAAI和IJCAI等，相关学术期刊主要包括Artificial Intelligence和Journal of AI Research；（3）机器学习领域，相关学术会议主要包括ICML，NIPS，AISTATS，UAI等，相关学术期刊主要包括Journal of Machine Learning Research（JMLR）和Machine Learning（ML）等。例如最近兴起的knowledge graph研究论文，就有相当一部分发表在人工智能和信息检索领域的会议和期刊上。我国计算机学会（CCF）制定了一份“中国计算机学会推荐国际学术会议和期刊目录”（http://www.ccf.org.cn/sites/ccf/paiming.jsp），已经经过若干次重要修订，是迅速了解每个领域主要期刊与学术会议的方式。

最后，ACL学会维护了一个wiki页面（http://aclweb.org/aclwiki/），包含了大量NLP/CL的相关信息，如著名研究机构、历届会议录用率，等等，都是居家必备之良品，值得深挖。例如，我们可以从相应wiki页面（https://www.aclweb.org/adminwiki/index.php?title=2016Q3\_Reports:\_Program\_Chairs ）看到ACL 2016 PC主席的工作总结，了解投稿主题分布、审稿流程、录用决策方式、录用情况等信息，这对研究选题、改进论文写作和提高投稿命中率很有帮助。

值得一提的是，随着深度学习技术的兴起，NLP/CL的新模型层出不穷，这两年来有越来越多的国际顶级学者选择在预印本文库arxiv.org上首先发布自己的最新成果，已经成为了解NLP/CL深度学习最新技术的重要信息源。由于arxiv.org上的论文发布并不包括同行评审，发布速度比学术会议更快；但相应的，arxiv.org上的论文良莠不齐，不乏内容粗劣不堪的占坑之作。因此，建议大家在关注arxiv.org论文时，特别注意那些来自著名研究机构（如Google DeepMind、Facebook AI Research等）和著名学者（如Yoshua Bengio等）的论文，而对来源不太知名的论文则应多些审慎态度，避免浪费时间和精力。

1. 国内学术组织、学术会议与学术论文

国内NLP/CL领域最权威的专业学会是中国中文信息学会（http://www.cipsc.org.cn/）。通过学会理事名单（http://www.cipsc.org.cn/lingdao.php），基本能够了解国内从事NLP/CL的主要单位和学者。学会每年组织很多学术会议，例如全国计算语言学学术会议（CCL）、全国信息检索学术会议（CCIR）、全国社会媒体处理大会（SMP）、全国知识图谱大会（CCKS）、全国机器翻译研讨会（CWMT），等等，是国内NLP/CL学者进行学术交流的重要平台。近年来国内NLP/CL学者在国际会议上日益活跃，因此ACL、COLING等国际顶级会议也选择在国内举办。2010年在北京召开的COLING，以及2015年在北京召开的ACL，中文信息学会都是主要承办者，这一定程度上反映了学会在国内NLP/CL领域的重要地位。此外，中国计算机学会中文信息技术专委会组织的自然语言处理与中文计算会议（NLP&CC）也是最近崛起的重要国内学术会议。

中文信息学会下设一个青年工作委员会，基本上囊括了国内大部分NLP/CL的一线青年学者（http://www.cipsc.org.cn/qngw/?page\_id=62），在全国各地不定期地组织NLP/CL专题研讨会，例如知识图谱研讨会等。从今年开始，青工委还组织很多顶级会议的国内论文报告会，如IJCAI 2016、ACL 2016、EMNLP 2016、COLING 2016等等，基本反映了国内NLP/CL的最新重要进展，是很好的学习交流机会。这些活动信息均会在青工委微信公众号上发布，大家可以在微信中搜索“中文信息学会青年工作委员会”关注最新动态。

中文信息学会旗下的《中文信息学报》是国内该领域的重要学术期刊，发表过很多篇重量级论文，目前由清华大学孙茂松教授担任主编。国内著名的《计算机学报》、《软件学报》等期刊上也常有NLP/CL论文发表。

1. 社会媒体

过去国内主要的线上交流平台是水木社区BBS的AI、NLP等版面，如今早已成为明日黄花。随着社会媒体的发展，越来越多学者使用新浪微博、微信群和知乎，可以进行便捷地学术交流。很多在国外任教的老师和求学的同学也活跃在新浪微博和知乎上，例如王威廉（http://weibo.com/u/1657470871）、熊辰炎（https://www.zhihu.com/people/xiong-chen-yan）等，经常爆料国际上的最新动态，非常值得关注。总之，学术研究既需要苦练内功，也要重视交流。“言者无心、听者有意”，也许其他人的一句话就能点醒你苦思良久的疑惑。

1. 如何快速了解某个领域研究进展

最后简单说一下快速了解某领域研究进展的经验。你会发现，搜索引擎是查阅文献的重要工具，尤其是谷歌提供的Google Scholar，由于其庞大的索引量，将是我们披荆斩棘的利器。

当需要了解某个领域，如果能找到一篇该领域的最新研究综述，就省劲多了。最方便的方法还是在Google Scholar中搜索“领域名称 + survey / review / tutorial / 综述”来查找。也有一些出版社专门出版各领域的综述文章，例如NOW Publisher出版的Foundations and Trends系列，Morgan & Claypool Publisher出版的Synthesis Lectures on Human Language Technologies系列等。它们发表了很多热门方向的综述，如文档摘要、情感分析和意见挖掘、学习排序、语言模型等。Google Scholar上也有越来越多学者建立了个人的论文主页，可以方便地看到相关领域重要学者的研究脉络。最近AI2也建立了一个叫做Semantic Scholar的学术网站，提供更精细的作者和论文引用分析。

如果方向太新还没有相关综述，一般还可以查找该方向发表的最新论文，阅读它们的“相关工作”章节，顺着列出的参考文献，就基本能够了解相关研究脉络了。当然，还有很多其他办法，例如去videolectures.net上看著名学者在各大学术会议或暑期学校上做的tutorial报告，去直接咨询这个领域的研究者，等等。