批准立项年份	2009
通过验收年份	2012
上轮评估年份	
上轮评估结果	

教育部重点实验室评估五年工作总结报告

（2014 年 1 月——2018 年 12 月）

实验室名称：计算语言学教育部重点实验室

实验室主任：穗志方

实验室联系人/联系电话：詹卫东/62765810

实验室联系人 E-mail 地址：zwd@pku.edu.cn

依托单位名称（盖章）：北京大学

依托单位联系人/联系电话：张琰/62752059

依托单位联系人 E-mail 地址：pkuzhangy@pku.edu.cn

2019 年 8 月 18 日填报

简表填写说明

一、总结报告中各项指标只统计 5 年评估期限内的数据（如：2015 年实验室评估材料的起止时间为 2010 年 1 月 1 日至 2014 年 12 月 31 日）。报告中涉及的各项数据统计均需附说明或佐证材料，按要求单独装订。其中，清单列表作为附件一，佐证材料作为附件二。

二、**“研究水平与贡献”**栏中，所有统计数据指评估期内由实验室人员在本实验室完成的重大科研成果，以及通过国内外合作研究取得的重要成果。其中：

1.**“论文与专著”**栏中，成果署名须有实验室。专著指正式出版的学术著作，不包括译著、实验室年报、论文集等。未正式发表的论文、专著不得统计。他引是指论文被除作者及合作者以外其他人的引用。

2. **“奖励”**栏中，取奖项排名最靠前的实验室人员，按照其排名计算系数。系数计算方式为：1/实验室最靠前人员排名。例如：在某奖项的获奖人员中，排名最靠前的实验室人员为第一完成人，则系数为 1；若排名最靠前的为第二完成人，则系数为 1/2=0.5。实验室在评估期内获某项奖励多次的，系数累加计算。部委（省）级奖指部委（省）级对应国家科学技术奖相应系列奖。一个成果若获两级奖励，填报最高级者。未正式批准的奖励不得统计。

3.**“承担任务研究经费”**指评估期实际到账的研究经费、运行补助费和设备更新费。

4.**“发明专利与成果转化”**栏中，某些行业批准的具有知识产权意义的国家级证书（如：新医药、新农药、新软件证书等）视同发明专利填报。国内外同内容专利不得重复统计。

5.**“标准与规范”**指参与制定国家标准、行业/地方标准的数量。

6.**“代表性研究成果”**应是根据科学前沿和国家、行业、区域重大需求所开展的、为促进科学发展或解决关键科技问题以及为国家、行业、区域发展决策提供科技支撑等方面所取得的系列进展，而不是一些关联度不高的研究方向的成果汇总。成果形式包括：论文和专著、标准和规范、发明专利、仪器研发方法创新、政策咨询、基础性工作，等等。

三、**“研究队伍建设”**栏中：

1.统计的范围包括实验室固定人员和流动人员。固定人员指高等学校聘用的聘期 2 年以上的全职人员；流动人员包括访问学者、博士后研究人员等。

2.**“40 岁以下”**是指截至 2014 年 12 月 31 日，不超过 40 周岁。

3.**“科技人才”**和**“国际学术机构任职”**栏，只统计固定人员。

4.**“国际学术机构任职”**指在国际学术组织和学术刊物任职情况。

四、**“学科发展与人才培养”**栏中，与企业/科研院所联合培养和国际联合培养的研究生需培养单位之间签订正式的相关培养协议。

五、**“开放与运行管理”**栏中：

1.**“承办学术会议”**包括国际学术会议和国内学术会议。其中，国内学术会议是指由主管部门或全国性一级学会批准的学术会议。

2.**“国际合作项目”**包括实验室承担的自然科学基金委、科技部、外专局等部门主管的国际科技合作项目，参与的国际重大科技合作计划/工程（如：ITER、CERN 等）项目研究，以及双方单位之间正式签订协议书的国际合作项目。

目 录

一、简表……1

二、研究水平与贡献……5

1. 科学影响及面向国家需求情况……5
2. 研究成果与贡献……6
3. 承担科学任务……15
4. 发展思路与潜力……18

三、研究队伍建设……19

1. 队伍建设总体情况……19
2. 实验室主任和学术带头人……20
3. 流动人员情况……23

四、学科发展与人才培养……24

1. 学科发展……24
2. 科教融合推动教学发展……25
3. 人才培养……26

五、开发交流与运行管理……28

1. 开放交流……28
2. 运行管理……32
3. 仪器设备……34

六、审核意见……35

一、简表

<table>
<tr><td colspan="2">实验室名称</td><td colspan="6">计算语言学教育部重点实验室</td></tr>
<tr><td colspan="2" rowspan="5">研究方向
(据实增删)</td><td colspan="2">研究方向 1</td><td colspan="4">语言认知机制与计算建模</td></tr>
<tr><td colspan="2">研究方向 2</td><td colspan="4">语言资源与语言知识工程</td></tr>
<tr><td colspan="2">研究方向 3</td><td colspan="4">语言复杂系统处理模型</td></tr>
<tr><td colspan="2">研究方向 4</td><td colspan="4">自然语言深度理解关键技术</td></tr>
<tr><td colspan="2">研究方向 5</td><td colspan="4">语言智能信息处理应用</td></tr>
<tr><td rowspan="2">实验室主任</td><td>姓名</td><td>穗志方</td><td>研究方向</td><td colspan="3">计算语言学</td></tr>
<tr><td>出生日期</td><td>1970 年 6 月</td><td>职称</td><td>教授</td><td>任职时间</td><td>2013 年 1 月</td></tr>
<tr><td rowspan="2">实验室副主任</td><td>姓名</td><td>袁毓林</td><td>研究方向</td><td colspan="3">语言学</td></tr>
<tr><td>出生日期</td><td>1962 年 8 月</td><td>职称</td><td>教授</td><td>任职时间</td><td>2013 年 1 月</td></tr>
<tr><td rowspan="2">实验室副主任</td><td>姓名</td><td>詹卫东</td><td>研究方向</td><td colspan="3">语言学</td></tr>
<tr><td>出生日期</td><td>1972 年 8 月</td><td>职称</td><td>教授</td><td>任职时间</td><td>2013 年 1 月</td></tr>
<tr><td rowspan="2">实验室副主任</td><td>姓名</td><td>常宝宝</td><td>研究方向</td><td colspan="3">计算语言学</td></tr>
<tr><td>出生日期</td><td>1971 年 9 月</td><td>职称</td><td>副教授</td><td>任职时间</td><td>2013 年 1 月</td></tr>
<tr><td rowspan="2">学术委员会主任</td><td>姓名</td><td>李宇明</td><td>研究方向</td><td colspan="3">语言学</td></tr>
<tr><td>出生日期</td><td>1955 年 6 月</td><td>职称</td><td>教授</td><td>任职时间</td><td>2013 年 1 月</td></tr>
<tr><td rowspan="11">研究水平与贡献</td><td rowspan="3">论文与专著</td><td>发表论文</td><td>SCI</td><td>52 篇</td><td>EI</td><td>143 篇</td></tr>
<tr><td colspan="2" rowspan="2">人均论文
(SCI+EI)/实验室人员数</td><td rowspan="2">5 篇/人</td><td>篇均他引</td><td>7.2 次
Google scholar</td></tr>
<tr><td>单篇最高他引次数</td><td>205 次
Google scholar</td></tr>
<tr><td>科技专著</td><td>国内出版</td><td>19 部</td><td>国外出版</td><td>6 部</td></tr>
<tr><td rowspan="4">奖励</td><td>国家自然科学奖</td><td>一等奖</td><td>0 项</td><td>二等奖</td><td>0 项</td></tr>
<tr><td>国家技术发明奖</td><td>一等奖</td><td>0 项</td><td>二等奖</td><td>0 项</td></tr>
<tr><td>国家科学技术进步奖</td><td>一等奖</td><td>0 项</td><td>二等奖</td><td>0 项</td></tr>
<tr><td>省、部级科技奖励</td><td>一等奖</td><td>2 项</td><td>二等奖</td><td>4 项</td></tr>
<tr><td rowspan="2">承担任务研究经费</td><td>5 年项目到账总经费</td><td>7774 万元</td><td colspan="2">前 25 项重点任务</td><td>3795 万</td></tr>
<tr><td>纵向经费 6321 万元</td><td>横向经费 1453 万元</td><td colspan="2">人均经费
(纵向+横向)/实验室人员数</td><td>172 万元/人</td></tr>
<tr><td rowspan="2">发明专利与成果转化</td><td>发明专利</td><td>申请数</td><td>40 项</td><td>授权数</td><td>7 项</td></tr>
<tr><td>成果转化</td><td>转化数</td><td>11 项</td><td>转化总经费</td><td>125 万元</td></tr>
<tr><td>标准与规范</td><td>国家标准</td><td colspan="2">0 项</td><td>行业/地方标准</td><td>0 项</td></tr>
</table>

	代表性研究成果(不超过 5 项)	**序号**	**成果名称**	**成果形式**
		第 1 项	面向语言理解的汉语意合语法理论	论文、专著
		第 2 项	中文语义知识库	论文、发明专利、软件系统
		第 3 项	面向大规模复杂结构的语言理解模型	论文、发明专利、软件系统
		第 4 项	自然语言多层次理解技术	论文、发明专利、软件系统
		第 5 项	基于文本生成的机器写作应用	论文、发明专利、软件系统

研究队伍建设	科技人才	实验室固定人员	45 人	实验室流动人员	16 人
		院士	0 人	千人计划	0 个
		长江学者	特聘 1 人	国家杰出青年基金	0 人
		青年长江	2 人	国家优秀青年基金	1 人
		青年千人计划	1 人	新世纪人才	1 人
		其他国家、省部级人才计划	4 人	国家自然科学基金委创新群体	0 个
		科技部创新团队	0 个	教育部创新团队	0 个

	国际学术机构任职(据实增删)	**姓名**	**任职机构或组织**	**职务**
		常宝宝	• CCL 2017 程序委员会 • ACL、EMNLP 等学术会议	副主席/PC Member
		董秀芳	• 国际中国语言学学会（IACL） • Language, Communication,and Culture（俄罗斯）杂志	理事/编委
		范晓蕾	• Linguistics: An Interdisciplinary Journal of the Language Sciences（De Gruyter Mouton 出版） • Current Research in Chinese Linguistics(香港中文大学出版)	匿名审稿人
		冯岩松	• ACL	会员
		胡旭辉	• University of Cambridge-Chinese University of Hong Kong Joint Laboratory	Bilingualism 成员
		黄居仁	• the International Committee on Computational Linguistics • ACL Special Interest Group for Annotation (SIGANN) • Cambridge Studies in Natural Language Processing • Frontiers in Chinese Linguistics • Studies in East Asian Linguistics	Permanent Member/ Committee Member / Chief Editor/ Editor in Chief/ Editor in Chief
		孔江平	• 美国嗓音学会 • JCL（Journal of Chinese Linguistics)	会员/特约编辑
		李素建	• ACL 2017 • CCL 2017	领域主席

		林幼菁	• Himalayan Linguistics		编辑部成员
		陆俭明	• 国际中国语言学学会		会长
		苏祺	• 剑桥大学出版社 SNLP 系列 • CLSW2017		副主编/主编
		穗志方	• 第 16 届汉语词汇语义学国际研讨会		主席
		孙栩	• EMNLP 2015 • Transactions of Association of Computational Linguistics (TACL Journal) • NLPCC 2018 (Machine Learning) • IJCAI-ECAI 2018		Area Chair/Standing Review Committee/Area Chair/Senior Program Committee (SPC)
		万小军	• NAACL2016 • ACL2015		领域主席
		王彦晶	• IJCAI16(25th International Joint Conference on Artificial Intelligence) • 7th Indian Conference on Logic and its Applications (ICLA) • the 15th International Conference on Principles of Knowledge Representation and Reasoning • LOFT2016 (The Twelfth Conference on Logic and the Foundations of Game and Decision Theory) • The 25th Workshop on Logic, Language, Information and Computation (WoLLIC 2018)		程序委员会委员
		吴云芳	• 第 18 届汉语词汇语义学国际研讨会（CLSW 2017）		程序委员会联合主席
		俞士汶	• 汉语词汇语义学国际研讨会		学术指导委员会荣誉指导委员
		袁毓林	• 日本《现代中国语研究》		特约编委
		詹卫东	• 《科技与中文教学》（美 Journal of Technology and Chinese Language Teaching，ISSN: 1949-260X）		编委
	访问学者	国内	8 人	国外	0 人
	博士后研究人员	进站博士后	8 人	出站博士后	6 人
	40 岁以下实验室人员代表性成果(不超过 3 项，可	**序号**	**成果名称**		**成果类型**
		第 1 项	面向大规模复杂结构的语言理解模型		论文、发明专利、软件系统

<table>
<tr><td></td><td>与代表性成果重复)</td><td>第 2 项</td><td colspan="5">基于文本生成的机器写作应用</td><td>论文、发明专利、软件系统</td></tr>
<tr><td></td><td></td><td>第 3 项</td><td colspan="5"></td><td></td></tr>
<tr><td rowspan="9">学科发展与人才培养</td><td>依托学科(据实增删)</td><td>学科 1</td><td colspan="2">计算机应用技术</td><td>学科 2</td><td colspan="3">语言学及应用语言学</td></tr>
<tr><td>博士研究生</td><td colspan="3">毕业学生数</td><td>38 人</td><td colspan="2">在读学生数</td><td>56 人</td></tr>
<tr><td>硕士研究生</td><td colspan="3">毕业学生数</td><td>110 人</td><td colspan="2">在读学生数</td><td>115 人</td></tr>
<tr><td>联合培养研究生</td><td>校内跨院系</td><td colspan="2">0 人</td><td>与企业/科研院所</td><td>0 人</td><td>国际联合培养</td><td>0 人</td></tr>
<tr><td>承担本科课程</td><td colspan="4">10294 学时</td><td colspan="2">承担研究生课程</td><td>12212 学时</td></tr>
<tr><td>大专院校教科书</td><td colspan="4">2 部</td><td colspan="2">高等学校教学名师奖</td><td>0 人</td></tr>
<tr><td>国家级教学成果奖</td><td colspan="4">1 项</td><td colspan="2">省部级教学成果奖</td><td>1 项</td></tr>
<tr><td>国家精品课程</td><td colspan="4">0 项</td><td colspan="2">省部级精品课程</td><td>0 门</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="8">开放与运行管理</td><td>承办学术会议</td><td>国际</td><td colspan="2">2 次</td><td>国内（含港澳台）</td><td colspan="2">1 次</td></tr>
<tr><td colspan="2">国际合作计划</td><td colspan="2">3 项</td><td>国际合作经费</td><td colspan="2">60 万元</td></tr>
<tr><td colspan="2">实验室面积</td><td colspan="2">1300M^2</td><td>实验室网址</td><td colspan="2">klcl.pku.edu.cn</td></tr>
<tr><td colspan="2">主管部门五年经费投入</td><td colspan="2">（直属高校不填）万元</td><td>依托单位五年经费投入</td><td colspan="2">810 万元</td></tr>
<tr><td>学术委员会人数</td><td>15 人</td><td colspan="2">其中外籍委员</td><td>0 人</td><td colspan="2">五年共计召开实验室学术委员会议 5 次</td></tr>
<tr><td colspan="3">五年内是否出现学术不端行为：否</td><td colspan="4">五年内是否按期进行年度考核：是</td></tr>
<tr><td colspan="2">实验室科普工作形式</td><td colspan="5">开放日，五年累计向社会开放共计 50 天；
科普宣讲，五年累计参与公众 5500 人次；
发表科普类文章 2 篇；
2016 年江苏高校语言能力协同创新中心等单位举办第三届全国优秀大学生夏令营，俞士汶老师应邀于 7 月 19 日做了“语言、人脑与电脑”的科普讲座；
2018 年第八届汉语言文字学高级研讨班暨青年学者论坛于 7 月底和 8 月初在吉林大学举办。俞士汶老师应邀于 8 月 3 日做了“计算语言学介绍”的科普讲座。</td></tr>
</table>

二、研究水平与贡献

1、科学影响及面向国家需求情况

简述实验室总体定位。结合研究方向，客观评价实验室在国内外相关学科领域中的地位和影响，在国家科技发展、社会经济发展、国家安全中的作用等。（800字以内）

1) 总体定位

语言是人类交流思想、表达情感最主要的方式，使用自然语言来表达和交流思想是人类高度智能的表现，语言的理解已成为实现人工智能的一个重要支撑。但是，自然语言具有歧义性、非规范性和个性化表达等特点，同时语言还承载着丰富的知识积累以及在此基础上的思维推理过程，这些特点和挑战成为了阻碍自然语言处理取得更大突破的拦路石。2018年10月31日，中央政治局会议，习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。国务院颁布的《新一代人工智能发展规划》将自然语言处理技术列为人工智能领域亟需突破的共性关键技术。本实验室的总体定位是：瞄准语言智能理解的核心瓶颈，充分利用计算机科学和语言学等多学科交叉融合的优势，面向国际学术前沿、国民经济建设主战场与国家重大战略需求，从语言学本体理论基础、语言资源与语言知识工程、语言复杂系统处理模型、语言深度理解关键技术以及语言智能信息处理与应用系统等多方面开展系统性的深入研究。建立以中文为核心的自然语言理解理论与方法体系。实验室的研究成果将对提升国家语言智能化处理与服务水平具有重要的战略意义，也将极大地提升我国语言信息产业的核心竞争力。

2) 主要研究方向

基于上述定位，实验室在以下五个方向开展创新性、系统性和前瞻性研究：

- 理论层面：充分利用语言学、计算机科学、认知科学等多学科融合的优势，探索语言深度理解的内涵，构建语言理解的多元认知理论基础。
- 资源层面：将语言学理论与计算机工程相结合，基于中文语言特点，构建面向中文深度理解的大规模语言知识资源基础设施。
- 模型层面：融合结构化学习与深度学习方法，解决自然语言理解的大规模复杂结构学习问题，提升语言深度理解的效果。
- 技术层面：将深度学习与复杂结构建模相结合，研发自然语言多层次理解及海量文本挖掘核心技术，研发海量文本内容分析系列关键技术。

➢ 应用层面：研制自然语言处理应用系统，服务社会，推动中文信息处理相关产业的技术变革

3) 国内外学术地位和影响

北京大学的计算语言学是北大计算机学科和语言学科深度融合、凸显文理交叉特色的研究方向。五年来，实验室的发展也带动了北京大学相关学科的提升。在 2017 年教育部组织的全国高校学科评估中，北京大学的计算机科学与技术、中国语言文学、外国语言文学均被评为 A+学科。2014-2018 年期间，在全球院校计算机科学领域实力排名 CS ranking 中，北京大学的自然语言处理名列全球第 1。在 2018 QS 世界大学学科排名中，北京大学的语言学排名全球第 10，全国第 1。

4) 在国家需求和社会发展中的作用

实验室在面向国家需求，引领国家科技发展中发挥了重要作用。2018年10月31日，中央政治局会议，习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量,加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。在此次中央政治局会议上，北京大学计算语言学教育部重点实验室学术委员会副主任高文院士对发展新一代人工智能的重大意义、任务和规划等问题作了讲解，并就促进人工智能同社会经济发展深度融合、引领新一轮科技革命和产业变革发表意见和建议。

实验室利用自然语言处理技术,成功研制多款写作机器人,应用于今日头条、南方都市报、光明网、腾讯等单位，已自动撰写与发布新闻资讯十万多篇。显著提高了写作效率与覆盖率，在业界取得良好反响，被国内外上百家国内外媒体广泛报道，推动了新闻出版行业的技术变革。

2、研究成果与贡献

结合研究方向，简要概述取得的重要研究成果与进展，包括论文和专著、标准和规范、发明专利、仪器研发方法创新、政策咨询、基础性工作等。总结实验室对国家战略需求、地方经济社会发展、行业产业科技创新的贡献，以及产生的社会影响和效益。（1000 字以内）

北京大学的计算语言学是北大计算机学科和语言学科深度融合、凸显文理交叉特色的研究方向。五年来，实验室充分利用语言学、计算机科学、认知科学等多学科融合的优势，探索语言深度理解的内涵，构建语言理解的多元认知理论基础。将语言学理论与计算机工程相结合，基于自然语言的特点，构建语言深度计算知识资源基础设施。将深度学习与复杂结构建模相结合，研发自然语言深度计

算及海量文本挖掘核心技术以及语言智能信息处理与应用系统系统，开展了创新性、系统性研究，取得了一系列具有重要影响力的成果：

(1) 理论——面向语言理解的汉语意合语法理论

在北京大学语言学、计算机科学、认知科学、心理学、逻辑学、哲学等多学科的传统积累基础上，充分利用学科交叉与融合的优势，探索语言深度理解的内涵，构建面向语言理解的汉语意合语法描写体系。出版语言学基础理论研究著作25 部，发表论文 75 篇，在过去五年（2014-2018）的 QS 全球大学排名中，北京大学的语言学和现代语言两个单项学科排名稳步提升，且都已进入全球前 10。北京大学中国语言学研究中心在教育部人文社会科学重点研究基地“十二五”评估（其中语言、文学、文献基地共 16 个）中，是评估结果获得“优秀”评级的唯一一家单位。2018 年，中文系陈保亚教授领衔的“语言学理论教学团队”获得国家级教学成果一等奖。袁毓林教授获得长江学者称号，入选中组部“万人计划”哲学社会科学领军人才，詹卫东教授、董秀芳教授获得青年长江学者称号。

(2) 资源——中文语义知识库

语言知识库是支撑语言信息处理的基础设施。将语言学理论与计算机工程相结合，构建支撑中文深度计算的语言知识资源基础设施。充分借鉴语言科学、认知科学的研究成果。针对中文“意合”的语言特点，建立了一套涵盖多层次语义信息的中文深层语义描述体系。发表语言知识资源构建相关规范 3 项。建立了中文语言知识资源构建工具集 CNLPware，覆盖从词法分析、句法分析到语义分析的核心构建工具软件，其中基于深度学习的汉语分词构件是国际上最早的中文深度分词模型，结合词典知识和标注数据的词义构件是目前国际最好的词义消歧基线方法，图解码深度依存分析构件，国际上最早发表的深度图解码依存分析，首次实现端到端的依存分析建模。构建了基于群体智慧的语言知识资源构建平台，实现规范化和(半)自动化的语言知识工程构建方法，建立了多层次大规模中文语义知识库。在国内外顶级会议和期刊发表论文 40 余篇，获得多项国际会议最佳论文奖，组织多项国际评测。

(3) 技术——面向大规模复杂结构的语言理解模型

在融合结构化学习方法与深度学习方法，解决自然语言理解的大规模复杂结构学习问题方面，取得了一系列具有国际影响力的研究成果，包括复杂语言结构降解理论和算法 StructReg、深层神经网络优化理论和算法 AdaBound、稀疏化语言学习算法 meProp 等，相关论文发表在自然语言处理和机器学习的国际顶级会议 ACL、ICML、NIPS、ICLR、COLING 等。提出的方法和理论在多个语言理解和生成任务刷新本领域准确度，被广泛应用于学术界和产业界，在 Github 的总 Star 数超过了 6000。孙栩研究员于 2014 年入选中组部“青年千人计划”; 2015 年获得香港求是基金会“求是杰出青年学者奖”，为该年度计算机领域唯一获奖

学者；2016 年在自然语言处理顶级国际会议之一 EMNLP 开设三小时的特邀 Tutorial 报告向国际学术界介绍结构化 NLP 技术，并以 119 人注册成为最受欢迎的 2 个 Tutorial 之一。2018 年获得“中国计算机学会自然语言处理与中文计算（NLPCC）青年新锐奖”； 2018 年获自然语言处理顶级会议之一 COLING 最佳论文奖（Best Paper Award）、为该年度中国唯一获奖论文。

(4) 技术——自然语言多层次理解技术

基于语言学的研究成果，从词汇、语句、篇章多层次进行自然语言的深层理解，最早提出了利用依存结构表示文本单元之间的关系，可以表示出文本单元之间的非投射关系，降低了分析的难度，并在新闻、科技领域构建了篇章依存结构语料库，在计算语言学理论和技术上进行了积极有益的创新性探索。同时结合心理认知学模拟人类重复阅读行为，提出多阶段多任务神经网络模型解析篇章结构，提高了篇章分析的性能。技术成果荣获顶尖国际学术会议 ACL 2017 两项杰出论文奖（全球唯一获两个奖项的实验室）。和百度合作构建了最大规模的中文阅读理解语料 Dureader，已有 130 多个团队参加该语料的评测；开发了阅读理解模型，在微软阅读理解数据集 MS MARCO 和斯坦福大学阅读理解数据集 SQuAD 上取得过当时的第一名。科研工作被世界知名学者广泛引用，其中包括美国宾州大学教授 Mitchell Marcus(AAAI Fellow)、美国斯坦福大学教授 Chris Manning（ACM/AAAI Fellow）、美国伊利诺伊大学芝香槟分校教授 Jiawei Han（IEEE/ACM Fellow）、美国卡耐基梅隆大学教授 Jaime Carbonell(AAAI Fellow)等。指导过的研究生获得过 Google 奖研金，微软学者称号等。

(5) 应用——基于文本生成的机器写作应用

在文本生成研究及机器写作应用方面取得了一系列国际一流成果和智能应用技术。在语义分析(Parsing)、智能问答、文本语义推理、微博检索等国际权威评测中连续多年获得第一名；研制了 PKUBase 知识图谱构建与问答平台和 gStore 图数据库系统、基于人机对话的智能投研与量化投资系统等智能应用技术。提出了一系列新颖的自动文摘与文本生成方法，包括图注意力神经网络生成模型、混合生成对抗网络模型以及 SentiGAN 等，以原创与二次创作两种方式实现高质量、长短可控、风格多样的文本稿件（包括新闻、文摘、评论、诗歌等）的智能创作，研究论文获得 IJCAI 2018 杰出论文奖，技术成果荣获吴文俊人工智能技术发明奖，所研制的机器写作系统应用于今日头条、南方都市报、日本三菱等多家单位，累计生产稿件十多万篇，大大提高了写稿效率与覆盖率，受到上百家国内外媒体的广泛报道，实现了人工智能技术在新闻出版行业的落地应用，推动了新闻出版行业的技术变革。

研究成果充分显示了实验室的原始创新能力，在国内外同领域具有重要地位和影响。

评估期内，实验室成员发表 SCI 论文 52 篇；出版专著 25 部，获专利授权 7 项。在 2017 年教育部组织的全国高校学科评估中，北京大学的计算机科学与技术、中国语言文学、外国语言文学均被评为 A+学科。2014-2018 年期间，在全球院校计算机科学领域实力排名 CS ranking 中，北京大学的自然语言处理名列全球第 1。在 2018 QS 世界大学学科排名中，北京大学的语言学排名全球第 10，全国第 1。

上述研究得到了国内外同行的高度评价。实验室成员在国际学术大会中作特邀报告 36 次，获得教育部自然科学奖二等奖 (2018)、中国人工智能学会 2017 年技术发明奖二等奖（2017）、中国计算机学会自然科学二等奖（2014） 等奖项。

实验室在面向国家需求，引领国家科技发展中也发挥了重要作用。2018 年 10 月 31 日，中央政治局会议，习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。在此次中央政治局会议上，计算语言学教育部重点实验室学术委员会副主任高文院士对发展新一代人工智能的重大意义、任务和规划等问题作了讲解，并就促进人工智能同社会经济发展深度融合、引领新一轮科技革命和产业变革发表意见和建议。

代表性研究成果简介

序号	成果名称	成果形式	第一完成单位	实验室参加人员姓名(排名)	成果产生年度
1	面向语言理解的汉语意合语法理论	专著、论文	北京大学	袁毓林、詹卫东、孔江平、陈保亚、郭锐、董秀芳、朱彦	2014.1-2018.12

与以英语为代表的形态丰富语言相比，汉语缺乏形态标记，语法机制以意合为主，这对计算机进行中文的深度理解而言，是公认的一大难题。在过去五年，实验室充分利用北京大学在语言学、计算机科学、认知科学、逻辑学等基础学科的综合优势，积极探索自然语言语义深度理解的内涵，围绕“汉语意合语法的可计算知识表征框架”“普通话语音的多模态研究与中华虚拟发音人生理建模”“中国口传文化的数字化保护方法与基础理论”“中国境内民族语言及方言接触类型与纵向横向传递途径和机制研究”“一般会话含义与特殊会话含义理解途径的脑成像实验对比研究”等五个基础研究课题，通过文理跨学科知识深度融合、团队人员互动协同的创新研究机制，走出了汉语意合语法长期停留在观念层面而无具体可操作描写方法的困境，解决了普通话语音模型从声学建模向生理建模扩充过程中的参数获取难题，形成了一批高质量的研究成果，共计出版语言学基础理论研究著作 25 部，在权威刊物发表高水平学术论文 75 篇。

在借鉴生成词库论、论元结构理论、事件语义学、框架语义学、构式语法理论等多种语言学理论的基础上，袁毓林教授领导研究团队提出了基于语言大数据驱动的汉语意合语法理论描写模型，遵循“小语法、大词库”的语言知识表征路线，面向汉语计算处理，首次论述了“词库-构式”互动的汉语意合语法描写模型，在动态的语义组合和语义推理平面上，用语义的扩散性激活和缺省推理等动力学机制来说明比较特殊的词语组合、句子形式的语义解释，并基于这一模型构建了汉语常用名词、动词、形容词和副词等的物性结构、论元结构、句式构造、构式系统等完整的知识描述，构建了可服务于文本蕴含推理的概念——事件网络模型。孔江平教授领导研究团队将前沿的言语科技与中华传统有声文化的整理传承，以及语音的多模态研究有机结合在一起，提出了集“音律、格律、曲律、乐律”为一体的“四律”语音学理论模型。上述成果的启示是：从认知心理学和计算机处理自然语言的角度，以汉语为核心，研究句法、语义以及语音问题，要重视语言理论模型与语言工程的结合，汉语意合语法理论的宏观高度，离不开语言资源的大数据基础及其落地实现。具有可操作性的高度形式化和数据化的汉语意合语法理论体系，深深植根于北京大学语言学、实验语音学、心理学、逻辑学和计算机科学的交叉融合的学术土壤之中。

序号	成果名称	成果形式	第一完成单位	实验室参加人员姓名(排名)	成果产生年度
2	中文语义知识库	论文、发明专利、软件系统	北京大学	穗志方、俞士汶、常宝宝、詹卫东、刘扬、段慧明	2014.1-2018.12

自然语言的语义分析和内容信息的理解，离不开大规模多层次的语言资源知识库的支持，它是帮助计算机“理解”人类语言的一个媒介和桥梁，也是让计算机逐渐“智能”起来的一个物质前提，是自然语言处理领域不可或缺的一项基础资源。对语言的深度理解涉及语言与认知的交互，目前盛行的统计方法很难突破。关于中文深度理解的描述体系、资源构建，国内外还没有系统深入的研究。北京大学研制成功的综合型语言知识库（简称 CLKB）是以汉语为核心的多语言知识库建设中最全面、最重要的研究成果，为以汉语为核心的多语言信息处理技术的发展提供了重要的基础设施和技术支持。在 CLKB 的基础上，我们借鉴论元结构理论、事件语义学、生成词库论、构式语法理论,突破语义角色标注等浅层语义分析的瓶颈，制订中文深度理解的描述规范。从计算机深度计算和语言工程的角度，对文本蕴含的语义信息进行分层次细粒度的深入挖掘。最终建立了多层次细粒度的大规模中文语义知识库，为支撑文本内容深度理解奠定了资源基础。

在自然语言处理核心技术方面，研究基于深度学习的中文语义计算技术，赋予机器理解中文语义内容的能力。先后在中文词法、句法和句义分析方面取得了一系列重要进展。在中文词法分析层面，于 2014 年提出的基于张量神经网络的中文分词模型，是目前国际国内引用最多的深度分词模型研究之一。在中文依存句法分析方面，于 2016 年提出的基于 LSTM 的深度依存分析模型，采用双向长短记忆网络模型进行特征自动提取，大幅简化了特征工程，并取得依存分析精度的显著进展。所提出的 LSTM-Minus 语段表示模型也被国际同行应用于成分句法分析语段表示建模并取得显著效果。在中文语义角色标记方面，先后提出基于语义组块的汉语语义角色标记方法，基于深度学习模型，有效融合异语、异质语义标注资源，取得了国际上领先的中文语义角色标注效果。针对标注资源贫乏的领域，也先后开展了无指导非参贝叶斯模型的研究工作，先后提出了无指导词义归纳模型和无指导分词模型，将层次狄利克雷模型、Zipf 意义律等应用于词义归纳和汉语分词任务，取得领先的无指导词义归纳和无指导分词效果。基于这些研究，在相关国际会议上组织了中文语义计算系列评测，吸引了近百家研究机构的参与。有多项研究成果发表在计算语言学领域国际著名会议 ACL、EMNLP、COLING 和 NAACL 上，获得 CCL and NLP-NABD 2015 Best Paper Award、NLPCC 2016 Best Student Paper 、NLPCC 2017 Best Student Paper 等多项论文奖项。

序号	成果名称	成果形式	第一完成单位	实验室参加人员姓名(排名)	成果产生年度
3	面向大规模复杂结构的语言理解模型	论文、发明专利、软件系统	北京大学	孙栩、苏琪、吴云芳、李素建	2014.1-2018.12

在融合结构化学习方法与深度学习方法，解决自然语言理解的大规模复杂结构学习问题方面，取得了一系列具有国际影响力的研究成果。提出基于序列生成的文本理解模型，在新闻分类、科研论文分析等任务获得显著效果提升；提出稀疏化神经网络反向传播算法(meProp)，加速深度学习反向传播速度 20-60 倍，在不损失准确率的前提下模型压缩达 9x；提出图像话题融合神经元网络模型 解决多模态注意力融合的问题；提出基于 GAN 的多样化对抗生成神经网络模型，捕捉深层隐含结构依赖，改善端到端自然语言生成理解问题。

相关论文发表在自然语言处理和机器学习的国际顶级会议 ACL、ICML、NIPS、ICLR、COLING 等。提出的方法和理论在多个语言理解和生成任务刷新本领域准确度，被广泛应用于学术界和产业界，在 Github 的总 Star 数超过了 6000。孙栩研究员于 2014 年入选中组部“青年千人计划”；2015 年获得香港求是基金会“求是杰出青年学者奖”，为该年度计算机领域唯一获奖学者；2016 年在自然语言处理顶级国际会议之一 EMNLP 开设三小时的特邀 Tutorial 报告向国际学术界介绍结构化 NLP 技术，并以 119 人注册成为最受欢迎的 2 个 Tutorial 之一。2018 年获得“中国计算机学会自然语言处理与中文计算（NLPCC）青年新锐奖”；2018 年获自然语言处理顶级会议之一 COLING 最佳论文奖（Best Paper Award）、为该年度中国唯一获奖论文。

序号	成果名称	成果形式	第一完成单位	实验室参加人员姓名(排名)	成果产生年度
4	自然语言多层次理解技术	论文、发明专利、软件系统	北京大学	王厚峰、李素建、万小军、孙斌、俞敬松、王雷	2014.1-2018.12

基于语言学的研究成果，从词汇、语句、篇章多层次进行自然语言的深层理解。学术创新如下：（1）提出新的主题神经模型，建模词汇关联精确表达词汇语义，词汇理解的工作被国际知名企业和高校高度评价和广泛跟随；（2）在国际学界最早提出基于依存框架的篇章结构表示，构建了第一个篇章依存精加工语料库，可以涵盖语言的非投射现象。

相关工作被美国、英国、法国、加拿大等 40 多个国际研究机构国际同行引用和跟随引用约 3000 余次。相关成果发表在领域顶级会议和期刊（包括 ACL、SIGIR、EMNLP、CIKM、AAAI、IJCAI、TKDE、TASLP 等）上发表高水平学术论文 50 多篇。科研工作被世界知名学者广泛引用，其中包括美国宾州大学教授 Mitchell Marcus(AAAI Fellow)、美国斯坦福大学教授 Chris Manning（ACM/AAAI Fellow）、美国伊利诺伊大学芝香槟分校教授 Jiawei Han（IEEE/ACM Fellow）、美国卡耐基梅隆大学教授 Jaime Carbonell(AAAI Fellow)等。和百度合作构建了最大规模的中文阅读理解语料 Dureader，已有 130 多个团队参加该语料的评测；开发了阅读理解模型，在微软阅读理解数据集 MS MARCO 和斯坦福大学阅读理解数据集 SQuAD 上取得过当时的第一名。技术成果荣获顶尖国际学术会议 ACL 2017 两项杰出论文奖（全球唯一获两个奖项的实验室）。指导过的研究生获得过 Google 奖研金，微软学者等奖励。

序号	成果名称	成果形式	第一完成单位	实验室参加人员姓名(排名)	成果产生年度
5	基于文本生成的机器写作应用	论文、发明专利、软件工具	北京大学	万小军、赵东岩、邹磊、孙薇薇、冯岩松、严睿	2014.1-2018.12

自然语言生成是自然语言处理与人工智能领域的一个重要研究方向，其目标是根据给定的信息表示（包括结构化数据、文字素材等）来产生符合人类阅读习惯的自然语言语句或篇章，自动生成流畅可读的新闻或文摘。自然语言生成技术能够广泛应用于媒体出版等行业，实现自然语言生成也是人工智能走向成熟的重要标志之一。

围绕可控自然语言生成的目标提出了一系列原创自动文摘与文本生成方法，实现长短可控、风格可控、内容可控、情感可控的文本稿件（包括新闻、摘要、综述、评论、诗歌等）的自动生成。所取得的代表性学术创新为：在长篇新闻与综述生成方面，首次提出基于直播文字的长篇报道自动生成方法，通过融合领域知识进行智能语句筛选，可实现高质量长篇新闻报道的实时生成，该成果解决了业界所面临的长文生成的难点，发表于顶级会议 ACL 2016，该成果具有极强的实用性，已经广泛适用于多款写作机器人系统。提出了混合对抗网络模型 SentiGAN，克服传统对抗网络模式崩塌的问题，实现高质量多类别情感文本的自动生成。该成果荣获人工智能领域顶级国际会议 IJCAI 2018 杰出论文奖(Distinguished Paper Award。相关成果在领域顶级与一流期刊与会议（包括 ACL、SIGIR、EMNLP、CIKM、AAAI、IJCAI、TKDE、TASLP、TOIS、CL 等）上发表高水平学术论文 50 多篇。研制了业界首个基于 Java 的文本自动摘要开源平台-PKUSUMSUM，并研制成功小明(Xiaomingbot)、小南等多款写稿机器人，应用于今日头条、南方都市报、光明网等单位，已自动撰写与发布新闻资讯 10 万多篇。其中“AI 小记者 Xiaomingbot”是国内首款既能自动写短文又能自动写长文的人工智能写稿机器人，为奥运会以及各类足球联赛、NBA 赛事等提供赛事新闻撰写服务，在今日头条平台上已自动撰写与发布体育类新闻数万篇，有效节约了成本，大大提高了写稿效率与覆盖率。上述写稿机器人受到业界广泛关注，被 Forbes、Lockerdome、Futurism、NextShark、QUARTZ、Popular Science、Fxtribune、Newsweek、Techweb、DotNews、科技日报、香港经济日报等上百家国内外媒体广泛报道，实现了人工智能技术在新闻出版行业的落地应用，在一定程度上推动了新闻出版行业的技术变革。机器写作成果荣获 2017 年度吴文俊人工智能技术发明奖。

3、承担科研任务

概述实验室评估期内承担科研任务总体情况。（600 字以内）

实验室科研人员聚焦实验室的战略定位，在以中文为核心的自然语言理解领域，积极承担国家重要科技前沿以及面向国家需求和服务于行业发展等方面的科技任务，2014-2018 年共承担科技项目 200 多项，五年到账总经费 7774 多万元。具体如下：

在面向国家需求及学科基础前沿方面，承担项目 112 项，总经费约 6321 万元 (占总经费的 81%)，包括：国家重点基础研究发展计划（973 计划）课题 1 项，国家高技术研究发展计划(863 计划)项目 1 项，课题 2 项，国家社科基金重大项目 5 项、教育部基地重大项目 1 项、科技部重点研发计划课题 1 项、任务 3 项，优秀青年基金项目 1 项、青年千人项目 1 项，以及多项国家自然科学基金面上项目。

在服务地方经济发展方面，承担项目 121 项，总经费 1453 万元（约占 19%），包括：技术转让、技术服务、科技开发、国际合作、海外合作、企事业委托等项目。

结构合理的项目经费投入，促进了实验室在研究成果与贡献、队伍建设与人才培养、以及开放交流与合作等方面的提升，取得良好的投入产出效益。

请选择主要的 25 项重点任务填写以下信息：

序号	项目/课题名称	编号	负责人	起止时间	经费(万元)	类别
1	融合三元空间的中文语言知识与世界知识获取和组织	2014CB340504	穗志方	20140101--20181201	427	973 计划-课题
2	大规模汉语语义基础资源和知识库设计构建及工具平台	2012AA011101	王厚峰	20120101--20141231	2224	863 计划-项目
3	面向基础教育的类人智能知识理解与推理关键技术	2015AA015403	赵东岩	20150101--20171231	555	863 计划-课题
4	基于海量知识的语言信息智能理解与推理关键技	2015AA015404	孙栩	20150101--20191201	71.85	863 计划-子任务

	术*					
5	基于中国语言及方言的语言接触类型和演化建模研究	14ZBD102	陈保亚	20140101--20181231	80	国家社科基金重大项目
6	中国有声语言及口传文化保护与传承的数字化方法及其基础理论研究	10&ZD125	孔江平	20110101--20151231	160	国家社科基金重大项目
7	面向网络文本的多视角语义分析方法、语言知识库及平台建设研究	12&ZD227	王厚峰	20130101--20171231	80	国家社科基金重大项目
8	汉语国际教育背景下的汉语意合特征研究与大型知识库和语料库建设	12&ZD175	袁毓林	20121012--20171012	222	国家社科基金重大项目
9	基于多学科视域的认知研究	12&ZD119	周北海	20121001--20191231	72	国家社科基金重大项目
10	基于语音多模态的语言本体研究	17JJD740001	孔江平	20170101--20201201	80	教育部基地重大项目
11	图数据管理关键技术及系统	2016YFB1000603	邹磊	20160701--20190601	640	科技部重点研发计划-课题
12	融合大数据与人类常识的开放域多语言知识图谱构建*	2017YFB1002101	穗志方	20171001--20210930	312.6	科技部重点研发计划-子任务
13	基于异构图计算机的数据管理与分析系统*	2018YFB1003504	邹磊	20180501--20210430	230	科技部重点研发计划-子任务
14	基于大数据的类人智能关键技术与系统（二期）*	2018YFB1005100	冯岩松	20181201--20211130	198	科技部重点研发计划-子任务
15	数据库理论与系	61622201	邹磊	20170101-	130	优秀青年科学

	统			-20191201		基金
16	文本语言表达到概念关系的映射方法研究与资源建设	61305089	穗志方	20140101-20171201	79	国家自然基金面上项目
17	基于深层学习的汉语句法语义分析研究	61273318	常宝宝	20130101--20161201	80	国家自然基金面上项目
18	大规模汉语历时语料库建设及词汇语义变迁研究	61472017	胡俊峰	20150101--20181201	80	国家自然基金面上项目
19	基于网络异构文本数据融合的热点话题发现及其内容摘要研究	61273278	李素建	20130101--20161201	80	国家自然基金面上项目
20	面向科技文献的引用摘要生成关键技术研究	61572049	李素建	20160101--20191201	75	国家自然基金面上项目
21	基于隐含知识挖掘与时间敏感的知识图谱补全关键技术研究	61772040	穗志方	20180101--20211231	60	国家自然基金面上项目
22	命名实体消歧与多源实体知识获取方法研究	61370117	王厚峰	20140101--20171201	81	国家自然基金面上项目
23	基于汉语话题的句际关系自动分析研究	61371129	吴云芳	20140101--20171201	80	国家自然基金面上项目
24	基于文档的智能问答的关键技术研究与资源建设	61773026	吴云芳	20180101--20211231	72	国家自然基金面上项目
25	面向信息处理的汉语语素体系构建及应用研究	16BYY137	刘扬	20160701--20201201	19	国家社科基金一般项目

注：请依次以国家重大科技专项、“973”计划（973）、“863”计划（863）、国家自然科学基金（面上、重点和重大、创新研究群体计划、杰出青年基金、重大科研计划）、国家科技

（攻关）、国防重大、国际合作、省部重大科技计划、重大横向合作等为序填写，并在类别栏中注明。只统计项目/课题负责人是实验室人员的任务信息。只填写所牵头负责的项目或课题。**若该项目或课题为某项目的子课题或子任务，请在名称后加*号标注。**佐证材料放入附件二。

4、发展思路与潜力

简要介绍实验室的优势与存在的不足、今后五年的建设目标、发展思路和保障举措等。（800 字以内）

1) 主要优势：

科研工作密切结合国家需求：自然语言理解是人工智能领域的重要研究课题，围绕这一问题，实验室从理论探索、资源构建、模型设计、技术研发、应用落地等方面展开全方位研究。与习近平总书记在中央政治局会议上的讲话精神和国务院《新一代人工智能发展规划》制订的自然语言处理技术重大需求结合紧密。

研究方向符合国际学科前沿：自然语言理解可使计算机实现从感知智能到认知智能的飞跃，是人工智能发展的高级阶段，是融合计算机科学、语言学、认知心理等多学科的新型交叉性学科，新的学术问题和挑战不断涌现，具有很好的发展前景。

学科交叉特色鲜明：实验室研究人员学科背景包括计算机科学与技术、中文、外语、逻辑学等领域，是典型的交叉学科。实验室通过学术交流、联合申请项目、联合发表论文促进成员之间的合作与学科之间的融合，在语言学、计算机科学、认知科学等多学科的传统积累基础上，充分利用学科交叉与融合的优势，探索语言深度理解的内涵，构建语言理解的多元认知理论基础。实验室的建设也促进了各自研究方向的发展。

科研队伍精干：人数不多但队伍组织结构清晰，年龄、学科分布合理，保证了多学科融合的顺利实施与推进。科研队伍包括长江学者 1 名、万人计划领军人才 1 名、青年长江学者 2 名、青年千人 1 名、万人计划青年拔尖人才 2 名、优青 1 名、求是杰出青年学者 1 名，以及北京大学人文杰出学者和王选青年学者多名。实验室为青年人才的发展提供了良好的发展条件，青年人才获得多项人才奖励和科研奖励。

2) 主要不足：

实验室发展迅速导致办公和科研空间紧张，实验室由北京大学校内多家单位组成，物理空间比较分散。北京大学已经成立了人工智能研究院，实验室的研究方向作为人工智能的一个重要分支，也将全面纳入北京大学人工智能研究院的发展规划之中，空间和人员都将在此统一框架之下得以集中解决。

3) 五年目标、发展思路和保障举措：

发展目标：自然语言理解是人工智能发展的终极目标之一，围绕这一目标，争取在计算语言学理论基础、语言资源与语言知识工程、语言复杂系统处理模型、语言深度理解关键技术、语言智能信息处理应用等方面取得重大原创性成果，推动计算语言学及相关学科发展；为实现国家人工智能战略贡献力量。争取再引入 1 名千人、2-3 名青千；培养 2-3 名长江教授、2-3 名杰青、4-5 名优青。

发展思路：针对人工智能、语言智能和自然语言理解等前沿科学问题和国家重大目标，围绕语言、认知与计算的交叉与融合开展创新性的研究，加强实验室内部及与国内外其他机构的密切合作；充分发挥实验室开放基金和青年基金的作用；充分利用教学与科研相互融合的优势，争取承担更多国家重要科技任务。

保障措施：1) 进一步提升承担国家重大任务的能力，围绕总体定位，开展战略性、系统性和前瞻性研究；2) 通过千人计划、长江计划和北大百人计划等人才计划进一步加强队伍建设，适当增加体量，五年后固定研究人员达到 50-55 名和高级技术人员 3-5 名；3)争取更多的研究办公空间，实验室面积达到 4000 平米以上，改善实验室条件。

三、研究队伍建设

1、队伍建设总体情况

简述实验室队伍的总体情况，包括总人数，队伍结构，40 岁以下研究骨干比例及作用。简要介绍评估期内队伍建设、人才引进情况，以及吸引、培养优秀中青年人才的措施及取得的成绩。（800 字以内）

实验室共有全职人员 45 名，其中教学科研人员 44 人，教学辅助人员 1 人，行政管理 0 人。正高级职称 20 人，约占总人数 44%；副高级职称 18 人，约占总人数 40%；中级职称 6 人，约占总人数 13%。队伍组织结构清晰，年龄、学科分布合理。

40 岁及以下的全职人员 11 人，约占总人数的 25%，青年成员在实验室的教学科研工作中发挥了重要的作用，例如在评估期内，万小军教授的研究论文分别获得国际计算语言学顶级会议 ACL Outstanding Paper Award 和 IJCAI Distinguished Paper Award；孙栩研究员在面向大规模复杂结构的语言理解模型方面的研究成果获得了 Google、Microsoft、IBM 等国际知名公司企业的广泛关注。孙栩研究员于 2014 年入选中组部“青年千人计划”；2015 年获得香港求是基金会“求是杰出青年学者奖”，是该年度计算机领域唯一获奖学者。实验室多位青年学者获得青年长江、国家“万人计划”青年拔尖人才、国家自然科学基金委优秀青年基金、北京大学人文杰出青年学者奖、北京大学王选青年学者奖等多项青

年人才奖励。

在人才引进方面，实验室成果显著。评估期内（2014-2018），共引进 10 人，平均年龄 39 岁；其中，万人计划青年拔尖人才 2 名（周韧、王彦晶）。这些引进人才极大地充实了现有研究队伍，增强了实验室的创新活力。

在人才培养方面，实验室始终以队伍建设为工作核心，高度重视对年轻教师的培养，给予多方面的支持，包括积极从学校争取实验室空间，为新引进人员提供实验和办公条件、保证硕士和博士生的招收名额。此外，实验室建立了激励措施，奖励成员在论文发表、获奖等方面的成果，给予不同额度的科研经费支持等等。

在学校、学院和实验室的大力扶持下，实验室人才辈出。2014—2018 年间，袁毓林教授获得长江学者称号和万人计划领军人才称号，詹卫东教授和董秀芳教授获得青年长江学者称号，孙栩研究员获得青年千人称号，王彦晶副教授和周韧副教授获得万人计划青年拔尖人才称号，邹磊教授获得国家优秀青年基金。此外，孙栩研究员获得求是杰出青年学者称号，苏琪副教授、朱彦副教授、邹磊教授、詹卫东教授、董秀芳教授等多人获得北京大学人文杰出学者和王选青年学者奖。实验室为青年人才的发展提供了良好的发展条件，青年人才获得多项人才奖励和科研奖励。

2、实验室主任和学术带头人

简要列举实验室主任及学术带头人学术简历。（学术带头人为各研究方向带头人，每个学术简历不超过 200 字）

穗志方：实验室主任，北京大学信息科学与技术学院 教授、博士生导师。在中文语言资源及领域知识图谱构建方面，研究适合汉语的可计算的语言知识库理论体系、高效的知识库建设方法以及多层次语言知识的挖掘方法。作为课题负责人，承担国家 973 课题、863 项目、国家自然科学基金、国家社会科学基金、中国出版集团项目等多项科研项目。在计算语言学顶级国际会议 ACL、COLING、EMNLP 上发表学术论文 100 余篇，作为主要成员，制订国家标准 2 项。研究成果“综合型语言知识库”获 2011 年度国家科技进步二等奖和 2010 年度中国电子学会电子科学技术奖一等奖。

袁毓林：实验室副主任、北京大学中文系教授、博士生导师，教育部长江学者特聘教授，第三批国家“万人计划”哲学社会科学领军人才。主要研究理论语言学和汉语语言学，特别是从认知心理学和计算机处理自然语言的角度研究汉语句法和语义问题。正在进行面向内容计算和信息检索的语义知识资源的研究和建设工作。先后主持和承担多个科研项目，包括国家社科基金重大项目、教育部人

文社科重大项目等。多次获教育部人文社科优秀成果二等奖。担任过十个国内外学术期刊的编委。在《中国社会科学》、《中国语文》、《当代语言学》和《中文信息学报》等刊物发表论文一百余篇，出版中文著作 10 部，英语著作 2 部。

詹卫东：实验室副主任、北京大学中文系教授、博士生导师，入选 2017 年度“青年长江学者奖励计划”。主要研究现代汉语形式语法、中文信息处理、语言知识工程、语言文字应用规范。正在进行汉语构式知识库与语料库建设工作。先后主持和承担多个科研项目，包括教育部人文社科重大项目、国家社科基金项目、国家 973 项目、863 项目等。在《中国语文》《语言科学》《中文信息学报》等刊物发表论文多篇，出版专著一部。主持制定国家语言文字标准并主持编写标准解读本，参与编写《现代汉语》《计算语言学》《汉语参考语法》等教材多部。

常宝宝：实验室副主任、博士，副教授。主要研究领域为计算语言学、句法语义分析、自然语言生成等。作为负责人先后主持多个科研项目，包括多个国家自然科学基金及国家社科基金等项目。作为主要完成人，先后获得国家科技进步二等奖 1 项(排名第 3)以及包括教育部科技进步一等奖在内的省部级科研奖励 3 项。在国内外学术会议和期刊上发表学术论文近百篇，其中在 ACL、EMNLP、COLING、IJCAI、AAAI 等计算语言学或人工智能国际顶级会议上发表论文 40 余篇。多次担任 ACL、EMNLP、COLING、AAAI、IJCAI 等顶级国际会议程序委员会委员。担任 2017 年度全国计算语言学学术会议程序委员会副主席，2018 年国际词汇语义学会议程序委员会主席。还担任了《中文信息学报》编委，中国中文信息学会计算语言学专业委员会委员、中国人工智能学会自然语言理解专业委员会委员等社会工作。

陆俭明：北京大学中文系教授、博士生导师。现任国家语委咨询委员会委员、北京语言大学对外汉语研究中心学术委员会主任，以及南京大学、武汉大学、北京师范大学、北京语言大学等 17 所高等院校兼职教授。曾任世界汉语教学学会会长、国际中国语言学学会会长、中国语言学会副会长、北京大学汉语语言学研究中心主任、北京大学计算语言学研究所副所长、北京大学文科学术委员会委员、新加坡教育部课程发展署华文顾问等职。独立完成、出版的著作和教材共 10 部，主编或与他人合作编写论文集和教材 12 部；发表学术论文、译文、序文等 350 余篇，内容涉及现代汉语的本体研究和应用研究。他曾主持和参与多个省部级以上的科研项目。他在学界被誉为 20 世纪中国现代汉语语法研究八大家之一。先后获得中国第一届高等学校教学名师奖、北京大学国华杰出学者奖、香港理工大学 2000 年度大陆杰出学人奖等多个奖项。

俞士汶：北京大学计算语言学研究所教授。研究领域为计算语言学和自然语言处理。作为第一完成人的研究成果有以《现代汉语语法信息词典》为基础的“综合型语言知识库”，于 2011 年获中国国家科学技术进步奖二等奖。同年获中国中文信息学会的终身成就奖。著作 8 本（其中 2003 年出版的《计算语言学概论》

2016 年获评北京大学优秀教材），第一作者论文 120 余篇。为计算语言学学科领域培养了一大批高端人才，于 2012 年获北京大学访问学者优秀导师荣誉。“综合型语言知识库”于 2013 年再次获北京大学首届产学研结合特别贡献奖。

陈保亚：北京大学博雅特聘教授，主要研究语言学以及语言与人类复杂系统，汉藏语区域茶马古道的发现与命名人之一，提出语言接触的无界有阶模型、语言习得与认知的单位与规则还原程序。主持教育部基地十三五规划项目，并多次主持国家重大课题、重点课题、教育部重大课题、国家自然科学基金和国际项目，在 SSCI 等国际索引期刊发表论文十余篇。以第一人身份获王力语言学奖一等奖两次，国家教学成果奖一等奖一次、教育部人文社科优秀成果奖二等奖一次，北京市哲学社会科学优秀成果奖一等奖一次。担任语言与人类复杂系统国际会议主席一次，担任中国西南地区国际汉藏语研讨会主席两次。

孔江平：北京大学特聘人文教授、语言学实验室主任、教育部重点文科基地“中国语言学研究中心”研究员及管委会成员和北京大学、香港中文大学、台湾大学联合系统“语言与人类复杂系统联合研究中心”常务副主任。美国嗓音学会会员，中国声学学会高级会员。中国中文信息学会理事，中国艺术医学协会理事，中国民族语言学会理事。中国语言学会语音学分会副主任，中国中文信息学会语音信息专业委员会副主任，中国民族语言文字信息技术国家民委-教育部重点实验室学术委员会委员。JCL（Journal of Chinese Linguistics)特约编辑，《中国语音学报》创刊人之一，副主编，创刊号执行编辑；《语言学论丛》编委；《民族语文》编委；《语言与翻译》编委；《听力学及言语疾病杂志》编委；《中国听力语言康复科学杂志》编委。主持中国哲学社会科学重大投标项目、教育部人文社科基地重大项目等多个项目，获得全国高等教育 2009 年人文社会科学科研成果奖著作类二等奖、2009 年北京大学王力语言学二等奖。

王厚峰：北京大学信息科学技术学院教授，北京大学计算语言所所长。主要研究方向为语篇分析、自动问答、人机对话、观点挖掘和知识资源建设。曾担任国家 863 项目首席专家和国家社科基金重大项目首席专家，并主持国家自然科学基金面上项目、国家社科基金面上项目以及企业横向合作项目 10 余项，研究成果多次有偿转让。发表学术论文 80 余篇，包括自然语言处理、知识挖掘和人工智能的一流学术会议，如，ACL、KDD、AAAI、IJCAI 等。

孙栩：北京大学信息学院研究员、博士生导师。2010 年于日本东京大学获得计算机博士学位。先后在日本东京大学、微软公司美国雷蒙德研究院、美国康奈尔大学、香港理工大学担任研究职位。研究方向为自然语言处理和机器学习，特别是自然语言生成、面向语言的深度学习。在 ACL、ICML、NIPS、ICLR、EMNLP、COLING 等发表多篇论文。2014 年入选中组部“青年千人计划”。获得香港求是科技基金会“求是杰出青年学者奖”、中国电子学会科学技术奖一等奖、COLING 2018 最佳论文奖。

万小军：北京大学计算机科学技术研究所研究员，博士生导师，语言计算与互联网挖掘研究室负责人，在北京大学获得学士、硕士与博士学位。研究方向为自然语言处理与文本挖掘，主要研究内容包括自动文摘与文本生成、情感分析与语义计算等。曾担任计算语言学顶级国际期刊 Computational Linguistics 编委，目前担任 TACL 执行编辑、Natural Language Engineering 编委，担任自然语言处理领域顶级国际会议 EMNLP-IJCNLP 2019 程序委员会主席，10 多次担任相关领域重要国际会议领域主席(Area Chair)或高级程序委员(SPC)，包括 ACL、NAACL、EMNLP、IJCAI、AAAI 等。荣获 ACL2017 Outstanding Paper Award、IJCAI 2018 Distinguished Paper Award、2017 年吴文俊人工智能技术发明奖、CCF NLPCC 青年新锐奖等荣誉或奖励。与字节跳动、南都、三菱综研、科学网等单位合作推出小明、小南、小柯等多款 AI 写作机器人，受到国内外一百多家媒体的广泛报道。

赵东岩：北京大学计算机科学技术研究所研究员，博士生导师。主要研究方向为自然语言处理、语义数据管理、智能服务技术。近年来承担国家自然科学基金、863/重点研发计划等国家级项目 15 项、主持 7 项；发表学术论文 100 余篇（包括 ACL、AAAI、IJCAI、SIGKDD、SIGIR、SIGMOD、VLDB，AI Journal、TKDE、VLDB Journal 等顶级会议和期刊 50 余篇）；授权发明专利 20 项；先后七次获得国家和省部级奖励，包括 2006 年度国家科技进步二等奖（排名第一）；个人获第十届中国青年科技奖（2007 年）和北京市第七届“科技之光”技术创新特别奖等荣誉。兼任中国计算机学会（CCF）杰出会员，CCF 中文信息技术专委会秘书长。

3、流动人员情况

简要列举评估期内实验室流动人员概况，包括人数、引进流动人员的政策、流动人员对实验室做出的代表性贡献（限五个以内典型案例）等。（600 字以内）

实验室流动人员主要包括博士后和访问学者，以博士后为主。评估期内，出站博士后 6 人，在站博士后 8 人。

实验室给予博士后多方面的支持，如提供实验平台、办公条件以及提供实验室开放基金等。博士后积极参与合作导师的科研活动，在参与重大项目研究等方面做出了较大贡献，涌现出一批优秀博士后。若干优秀博士后人员的主要业绩简介如下：

博士后卢达威（合作导师袁毓林）参与了国家社科基金重大项目和国家高科技 973 项目的研究工作，主持了国家社科基金项目，在计算语言学研究方面取得了许多创新性成果，发表文章 10 余篇，现为中国人民大学助教授。

博士后王恩旭（合作导师袁毓林）参与了国家社科基金重大项目和国家高科

技 973 项目的研究工作，主持了国家社科基金项目，在汉语语法学和词汇学研究方面取得了许多创新性成果，发表文章 10 余篇，现为济南大学副教授。

博士后刘钰（合作导师邹磊）为北京大学博雅博士后，作为课题骨干参与了国家自然科学基金重点项目、国家重点研发计划等多个课题的研究工作，主持了一项北京大学医学交叉种子基金项目，在图近似算法和真实图模型等研究领域取得了一定创新性成果，已发表数据库领域国际顶级会议 VLDB 长文两篇。

四、学科发展与人才培养

1、学科发展

简述实验室所依托学科的发展情况，从科学研究和人才培养两个方面分别介绍对学校学科建设发挥的支撑作用，以及推动学科交叉与新兴学科建设的情况。（800 字以内）

北京大学的计算语言学是北大计算机学科和语言学科深度融合、凸显文理交叉特色的研究方向。五年来，实验室的发展也带动了北京大学相关学科的提升。在评估期内，北京大学的计算机学科和语言学科稳步发展，持续保持在全国的领先地位，在科研成果与水平、学科声誉和人才培养质量等方面在全国名列前茅。在 2017 年教育部组织的全国高校学科评估中，北京大学的计算机科学与技术、中国语言文学、外国语言文学均被评为 A+学科。2014-2018 年期间，在全球院校计算机科学领域实力排名 CS ranking 中，北京大学的自然语言处理名列全球第 1。在 2018 QS 世界大学学科排名中，北京大学的语言学排名全球第 10，全国第 1。

在科学研究方面，实验室聚焦自然语言理解这一核心主题，重点研究语言、认知与计算在语言理解中的相互关系以及以中文为核心的自然语言处理理论与方法体系。具体来说，在语言学理论方向研究人类语言的认知机制与语义表征；在语言资源与语言知识工程方向，研发支撑中文自然语言理解的知识资源基础设施；在语言复杂系统处理模型方向，研究自然语言理解中的大规模复杂结构学习问题；在语言深度理解关键技术方向，研究自然语言的分析与生成技术；在语言智能信息处理应用系统方向，研究基于文本理解与生成的机器写作应用。

这些研究不仅大大推动了北京大学计算语言学学科的发展，也促进了主要以理论研究成果为追求目标的基础学科如语言学、逻辑学、认知心理学等更加积极地与计算语言学进行跨学科碰撞，产生新的有价值的研究课题，比如将汉语构词法的理论研究应用于中文大规模概念词典中语义构词模式的标注，并在标注基础上生成人造语料用于词义的向量空间表示，将语言学知识融入神经网络深度学习模型，提升计算机在词语相似度计算方面的表现，为语言学研究与机器学习模型

研究之间的交叉结合研究模式，开辟了值得探索的方向。

在人才培养方面，实验室根据学科发展需要，不断完善人才培养体系和机制，积极培养和引进急需人才。实验室通过在北京大学设立中文系应用语言学专业、软微学院语言工程系、跨学科联合培养博士等方式，建立了本科—硕士—博士—访问学者这样一套完整的计算语言学人才培养体系。5 年共毕业博士生 38 名，硕士生 110 名；多人获得本领域国内外顶级学术会议优秀论文奖。过去 5 年，实验室培养了一批优秀人才，包括：教育部长江学者、青年长江学者、国家“万人计划”领军人才、青年拔尖人才、国家自然科学基金委优秀青年基金、求是杰出青年学者、CCF 自然语言处理与中文计算青年新锐等。这些优秀人才对推动计算语言学的学科发展发挥了重要作用。

2、科教融合推动教学发展

简要介绍实验室人员承担依托单位教学任务情况，主要包括开设主讲课程、编写教材、教改项目、教学成果等，以及将本领域前沿研究情况、实验室科研成果转化为教学资源的情况。（500 字以内）

重点实验室的固定人员均为学院一线在职教师，承担本科和研究生的教学任务。其中，本科生课程 45 门，总学时达到 10294 学时；承担研究生课程 70 余门，总学时达到 12212 学时；有效地将实验室的前沿科研成果与教学相结合，在传授基础知识的同时培养学生的学术创新能力。此外，实验室成员还积极进行教学改革工作，陈保亚、汪峰、董秀芳等教学成果《教学、实践、科研相结合的语言学培养模式》，同时获得国家级教学成果一等奖和省部级教学成果一等奖，北京大学优秀教材 2 项，编写教材 2 部。

实验室成员注重探索科教融合的新形式。结合北京大学基础学科拔尖学生培养计划和本科生科研基金项目，利用实验室基础研究方向众多、交叉领域新课题选择面宽等科研优势，实验室成员把科研与教学活动有机结合起来，在高年级本科生和研究生常规的课堂教学活动之外，注意激发学生的研究兴趣，引导和培养学生的科研创新精神。在科研选题、文献资料的收集和分析、研究方案的设计和制定、实验室分析和数据整理、资料的综合分析和论文写作的各个环节，实现教师与学生的互动。学生科研项目完成后，通过答辩可以获得 2-4 个学分。

上述举措获得了良好的效果。据不完全统计，评估期内，参加本科科研的学生达到 120 多人，本科生以第一作者发表顶级会议论文 7 篇。

3、人才培养

（1）人才培养总体情况

简述实验室人才培养的代表性举措和效果，包括跨学科、跨院系的人才交流和培养，与国内、国际科研机构或企业联合培养创新人才等。（800 字以内）

计算语言学作为跨文理的大跨度交叉型学科，对于复合型人才的需求更为突出。因此实验室特别重视人才培养，在过去五年，采取多种措施，在外部环境条件允许的情况下，开拓思路，加大创新力度，探索出针对本科生、研究生在跨院系环境下的学术交流和科研合作多种培养模式。

在本科生培养方面，实验室积极参与北京大学应用语言学专业本科课程体系建设，包括开设针对语言学专业背景（包括中文系、外文系、哲学系、心理系等）本科生的《程序设计》《语言统计分析》《编译原理》《语言、逻辑与计算》等课程，以及由计算机专业和语言学专业教师合作共同建设《自然语言处理导论》本科课程。在日常课程教学的基础上，结合国家拔尖学生计划，加强本科生参与科研项目的科研实践能力。近 5 年实验室有 7 篇本科生作为第一作者撰写的学术论文发表在自然语言处理国际顶级学术会议，一些本科生毕业后即加入国际知名研究机构，如中文系 2015 级应用语言学专业本科生林子毕业后进入 Google 人工智能研究院。信息科学技术学院 2011 级本科生李嫣然毕业后进入香港理工大学攻读博士学位，于 2017 年获得 Google 女性奖学金荣誉。

在研究生培养方面，实验室成员根据研究方向和研究课题的需要，组成了多个带有文理交叉性质的研究生指导小组，不同于一般院系仅由导师个人指导学生研究的方式，改为以团队形式和研究生一起开展学术沙龙、组织前沿研究讨论班等多种教学活动形式，指导小组成员共同为不同研究方向的学生论文提供指导意见，对论文质量进行把关，鼓励学生从不同学科的视角来分析计算语言学的学术难题，按照更高的学术标准要求自己，充分了解结合学科发展趋势和本领域前沿热点，做更有创新性和领先性的研究，近 5 年实验室研究生发表论文 113 篇，核心期刊论文 40 篇，10 篇论文获奖，培养了一大批面向学科发展前沿和服务国家需求的人才。

为培养具有国际视野和实践能力的创新型人才，实验室拓展各种渠道，积极与国内外科研机构和企业合作，联合培养研究生。派多位博士研究生前往美国华盛顿大学、伦斯勒理工学院、英国爱丁堡大学、微软亚洲研究院等国际知名大学与研究机构进行合作交流。

（2）研究生代表性成果（列举不超过 5 项）

简述研究生在实验室平台的锻炼中，取得的代表性科研成果，包括高水平论文发表、国际学术会议大会发言、挑战杯获奖、国际竞赛获奖等。（每段描述 200 字以内）

实验室为研究生的培养提供了有力的设备支持和资源保障，为高水平研究生的培养奠定了坚实的基础。具体表现为：实验室的研究生论文多次在国际顶级学术会议上获得最高奖项（Distinguished Paper Award, Best Student Paper, Outstanding Paper Award），多人在本领域顶级会议及核心期刊上发表文章。此外，实验室还积极探索研究生培养的创新举措，努力为学生提供更好更多的国际交流平台，拓展学术研究的国际视野。2014-2018 年间累积派出研究生出国参加学术会议 150 余人次，其中 100 余人次在会议上作口头报告。通过这些举措，研究生培养取得了丰硕成果，涌现出一批优秀毕业生，列举如下：

葛涛：在攻读博士期间，提出并实现了一系列关键技术解决文本流知识挖掘任务中的多个共性问题。研究成果获得了 NLPCC2016 最佳学生论文奖、Scholarship for the Workshop on Corpus and Empirical Linguistics、Google scholarship for LxMLS 以及北京大学校长奖学金、北京大学优秀科研奖等多项奖励。

王科：在攻读博士期间，提出混合对抗生成模型 SentiGAN 实现多类评论文本的批量生成，能有效克服传统对抗生成网络模式崩塌的问题，该成果荣获人工智能领域顶级国际会议 IJCAI 2018 杰出论文奖(Distinguished Paper Award，top 0.2%)

杨鹏程：在攻读硕士期间，提出用于多标签分类的序列生成模型，在自然语言处理主流数据集显著提升多标签分类的准确性。该成果获自然语言处理顶级国际会议 COLING 2018 最佳论文奖 (该年度来自中国大陆唯一获奖论文)。

王义中：在攻读硕士学位期间，提出一种新的基于两阶段的篇章分析方法，在国际篇章数据集 RST-DT 上显著提升了篇章分析的性能，该成果荣获自然语言处理顶级国际会议 ACL 2017 杰出论文奖(Outstanding Paper Award，top 1.5%)。

朴敏浚：在攻读博士期间，研究基于细粒度语言学知识的“比”字句分析模型及计算应用，在核心期刊上发表了多篇高水平学术论文。博士论文获得北京大学优秀博士论文奖以及中国中文信息学会优秀博士论文提名奖。

（3）研究生参加国际会议情况（列举 10 项以内）

序号	参加会议形式	参加会议研究生	参加会议名称及会议主办方	参加会议年度	导师
1	口头报告	裴文哲	the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)	2014	常宝宝
2	口头报告	曹自强	the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)	2014	李素建
3	口头报告	葛涛	the 53nd Annual Meeting of the Association for Computational Linguistics ACL2105	2015	穗志方
4	口头报告	李立	the 2015 Conference on Empirical Methods in Natural Language Processing EMNLP 2015	2015	王厚峰
5	口头报告	姜廷松	the 2016 Conference on Empirical Methods in Natural Language Processing（EMNLP 2016	2016	穗志方
6	口头报告	李炜	the 26th International Conference on Computational Linguistics: Technical Papers（COLING 2016）	2016	吴云芳
7	口头报告	马树铭	the 55nd Annual Meeting of the Association for Computational Linguistics ACL	2017	孙栩
8	口头报告	许晶晶	the 2018 Conference on Empirical Methods in Natural Language Processing.	2018	孙栩
9	口头报告	林子	the 2018 Conference on Empirical Methods in Natural Language Processing.EMNLP 2018	2018	万小军
10	口头报告	范非凡	the 2018 Conference on Empirical Methods in Natural Language Processing.EMNLP 2018	2018	赵东岩

以上均为国际自然语言处理领域顶级会议。

五、开放交流与运行管理

1、开放交流

（1）开放课题设置情况

简述实验室在评估期内设置开放课题、主任基金概况。（600 字以内）

为加强同行之间、学科之间的学术交流，促进实验室发展和高水平成果产出，扩大实验室的学术影响，实验室积极筹措经费，于 2013 年起，设立了开放基金。开放基金重点资助与本实验室研究方向密切相关、具有学术思想新颖、

处于学科发展前沿或优先发展领域的基础性、应用性研究项目。5 年期间实验室共设立 8 项开放课题,经费额度为 2.0-3.0 万元 / 项。因实验室经费有限,2016 年之后积极拓展与国内兄弟单位开放交流的更多途径，通过实验室固定成员承担的各级各类科研项目，以科研协作的形式，进一步加强与国内高校之间的合作与交流。

开放课题的实施产生了良好的效果。2014-2018 年间，这些开放课题已发表学术论文 17 篇，其中 EI 检索论文 9 篇，核心期刊论文 7 篇。通过开放课题的合作研究，一方面带动了兄弟院校的青年人才的培养工作，比如，承担实验室开放课题的郑州大学的昝红英老师、乐山师范学院的金澎老师、陈兴元老师均从副教授晋升为教授。昝红英老师成长为郑州大学计算机系的系主任，金澎老师成长为四川省学术技术带头人后备人选、互联网自然语言智能处理四川省高等学校重点实验室主任。为实验室承担国家重要科研任务储备了研究队伍和合作伙伴。已资助的开放课题共培养研究生 16 人、本科生 9 人。

（2）主办或承办大型学术会议情况

序号	会议名称	主办单位名称	会议主席	召开时间	参加人数	类别
1	语言资源构建——理论、方法与应用国际研讨会	北京大学计算语言学教育部重点实验室 北京大学中文系中国语言学研究中心 美国宾州大学语言资源联盟	穗志方	2017 年 11 月 4 日-6 日	95	全球
2	第 18 届词汇语义学国际研讨会	乐山师范学院互联网自然语言智能处理四川省高等学校重点实验室 北京大学计算语言学教育部重点实验室	俞士汶	2017 年 5 月 18 日-5 月 20 日	110	全球

3	中文信息学会暑期学校	北京大学计算语言学教育部重点实验室 中国中文信息学会	王厚峰	2015 年 7 月 24-25 日	260	全国

注：请按全球性、地区性、双边性、全国性等类别排序，并在类别栏中注明。

（3）国内外学术交流与合作情况

请列出实验室人员国内外学术交流与合作的主要活动，包括与国外研究机构共建实验室、承担重大国际合作项目或机构建设、参与国际重大科研计划、在国际重要学术会议做特邀报告的情况。请按国内合作与国际合作分类填写。（600字以内）

1) 国内合作

本实验室与国内高校、研究所以及实验室之间建立了广泛的联系，定期邀请国内知名专家赴实验室学术交流及讲座。实验室 2015 年举行的“中文信息学会暑期学校”，吸引了众多校外研究生参与。实验室设立了多项开放基金，资助校内外学者开展合作研究。实验室成员与国内同行在课题申请、共同发表学术成果等方面开展了大量实质性的合作。2014 年实验室还与江苏师范大学、清华大学、中国社会科学研究院、语言文字应用研究所合作，成立了语言能力协同创新中心。2014 年计算语言学教育部重点实验室（北京大学）联合武汉大学计算机学院、乐山师范学院，共建互联网自然语言智能处理四川省高等学校重点实验室。该实验室依托乐山师范学院，是四川省内唯一一家以自然语言处理为核心研究的学术机构。

2) 国际合作

联合举办国际会议：在自然语言信息处理旺盛应用需求的推动下，自然语言知识资源的建设历经 30 多年的高速发展，已经积累了相当丰富的数据。在当前深度学习的热潮下，如何认识语言资源数据加工的意义和价值，如何更有效地组织语言资源建设，是非常重要的议题。2017 年 11 月，实验室与美国宾州大学语言资源联盟（LDC）联合召开国际研讨会，邀请相关专业背景的专家学者，就语言资源建设的理论、方法及应用前景展开深度交流，共同推动自然语言知识资源的未来发展。来自美国宾夕法尼亚大学、香港大学、香港城市大学、香港理工大学、清华大学、上海交通大学、哈尔滨工业大学、天津大学、北京语言大学、中国科学院、中国社会科学院等国内外高校和科研机构的 80 余位专家学者围绕语言资源的构建理论、方法与应用等问题展开了深入的交流和精彩的碰撞。本次研讨会充分体现了北京大学在重视学科融合发展、国际学术交流方面的特色与优势，获得与会者高度评价，在自然语言处理领域产生了重要影响。

在国际重要学术会议做特邀报告和邀请来实验室举办讲座：实验室成员平均每年参加国际重要会议和国际学术交流活动 50 余人次。据不完全统计，自 2014 年以来有 30 余人次在国际重要学术会议做特邀报告。此外，实验室还邀请美国密西根大学端木三教授、美国伊利诺伊大学香槟分校的季姮教授、美国斯坦福大学的李纪为博士、比利时鲁汶大学的 Dirk Geeraerts 教授等大量国外知名学者来举办系列讲座和做学术报告。

（4）科学传播

简述实验室开展科学传播的举措和效果。（500 字以内）

实验室非常重视科学成果的传播，多次举办面向中学生的科普和参观活动，普及了计算语言学的学科知识，激发了同学们对计算语言学和人工智能的研究热情。

2015 年 7 月，第十届中国中文信息学会暑期学校在北京大学成功举办。语言技术暑期学校是国内语言信息处理领域最为重要的学术活动之一。本届暑期学校的成功举办，不仅让大家对自然语言处理及相关技术有了更深入的认识，而且通过交流让大家对自然语言处理技术的发展前景更加充满了信心，暑期学校获得了广大师生的普遍好评，为自然语言技术的人才培养和技术推广做出了卓越贡献，数百名学子在暑期学校中获得了来自国内外著名高校和科研机构的知名学者的当面指导，受益匪浅。

2016 年 7 月，中学生开放科学营的两批共 60 余名中学生参观了北京大学计算语言学教育部重点实验室。实验室主任穗志方教授、副主任常宝宝副教授、计算语言所所长王厚峰教授和柯永红博士后一同热情接待了同学们。利用通俗易懂的语言为同学们讲解了语言计算、自然语言处理、知识图谱、信息检索、机器翻译等概念的含义和应用场景。激发了同学们对计算语言学的学习热情。

2016 年江苏高校语言能力协同创新中心等单位举办第三届全国优秀大学生夏令营，俞士汶教授应邀做了“语言、人脑与电脑”的科普讲座。2018 年第八届汉语言文字学高级研讨班暨青年学者论坛，俞士汶教授做了“计算语言学介绍”的科普讲座。对与会的上百名青年学子进行了计算语言学的启蒙教育。

结合新课堂实践，实验室向社会开放 MOOC 课程。俞敬松副教授主持的《翻译技术原理和实践》大规模开放在线课程在 EdX, Coursera, CNMOOC, CHINESE

MOOCS 等多个广为流行的 MOOC 平台上提供，从 2013 年秋季起有大约 55,000 到 60,000 名学生注册学习，学生遍布全球。中国国内有包括北京大学、北京外国语大学、西安外国语大学在内的多所高校在其翻译和口译硕士学位项目教学使用本门公开课作为相关课程的辅助教学材料。这种模式将课后的在线学习与课上讨论结合起来，取得了很好的学习成果。

实验室还参加北京大学信息学院每年的开放日，接待优秀大学生夏令营营员，宣传和普及计算语言学基本知识。

2、运行管理

（1）实验室内部管理情况

请简要介绍实验室内部规章制度建设、网站建设、日常管理工作、自主研究选题情况、学术委员会作用，实验室科研氛围和学术风气、有无违反学术道德的事件发生。（600 字以内）

规章制度：实验室制定了详细的规章制度，涉及组织分工、日常管理、奖励办法、开放研究基金管理办法、学术研讨会、学术委员会工作条例、科研经费管理、统筹和财务报账管理制度等。实验室还建立了机房、设备、资源管理制度、消防安全管理规定。

日常管理：实验室日常管理清晰、有序。在实验室主任的主持下，常务副主任负责日常工作，其他两位副主任各司其职，实验室秘书负责实验室档案管理、网页维护和日常性行政工作。实验室经常召开工作会议，重点讨论实验室学术方向、日常运行中的对外开放、各研究方向间合作、学术交流和设备管理等重要事务。

网站与宣传：实验室设有专门的网站，为实验室成员和其他人员提供经常更新的消息和基本信息。

自主选题：实验室设立了开放课题，资助与本实验室研究方向密切相关的基础性和应用性研究项目。

学术委员会：实验室学术委员会在实验室建设中发挥了重要作用。每年召开一次学术委员会议，实验室主任汇报该年度的实验室工作进展和问题。在听取这些汇报和实验室运行情况汇报的基础上，学术委员会就实验室的研究方向、队伍建设、运行管理、学术交流等方面工作提出建设性意见和努力方向。此外，多名学术委员还在非正式会议时经常性的为实验室建设出谋划策，发挥了积极重要的作用。

学风情况：实验室具有非常浓厚的科研氛围和学术气氛，通过各种学术讨论会以及学术沙龙，各研究方向的老师们开展广泛合作，进行交叉研究。本实验室没有发生过违反学术道德的事件。

（2）主管部门和依托单位支持情况

简述主管部门和依托单位为实验室提供实验室建设和基本运行经费、相对集中的科研场所和仪器设备等条件保障的情况，在学科建设、人才引进、团队建设、研究生培养指标、自主选题研究等方面给予优先支持的情况。依托单位对实验室进行年度考核的情况。（600 字以内）

依托单位北京大学一贯高度重视实验室的发展，为实验室的发展提供了相对独立的建制，在人事和财务自主权方面能够充分考虑实验室的特点，给予全方位的支持。北京大学为实验室提供实验室建设和基本运行经费。学校科学研究部还对实验室的管理和年度考核给予指导，有专人负责重点实验室工作，参加实验室学术委员会会议，审查实验室年报，并及时提出建议和意见。依托单位还在实验室发展环境，如科研项目的立项、研究经费的争取、科研人员的用房、仪器设备的购置、人才的引进和培养、以及学术交流和国际合作、团队建设等方面，都给予大力支持。2014-2018 年间，共为实验室提供发展基金 810 万元。

北京大学信息科学技术学院与北京大学中文系为实验室提供了必要的实验室空间、仪器设备、人员和队伍建设等方面的支持。为实验室学科建设、人才引进、团队建设、研究生培养等方面的工作提供了便利的条件。

目前，本实验室主体在北京大学理科 1 号楼和人文学苑 6 号楼，后期学校将继续提供超过 600 平米的空间用于实验室教师办公室、学生用房、会议室、资料室以及服务器机房。实验室还将在未来获得更多的科研、实验和教学空间，这将为实验室的长远发展奠定坚实基础。

3、仪器设备

简述实验室大型仪器设备的使用、开放共享情况，研制新设备和升级改造旧设备等方面的情况。（500 字以内）

实验室目前没有大型仪器设备。

实验室设备包括：微机 76 台，笔记本电脑 69 台，服务器、工作站 75 台，用于配合课题研究采购的 32 导 EGI 脑电放大设备与人体生物信号采集装置、E-prime 脑电实验平台、Eyelink 1000+眼动仪及实验设计平台等设备，立足数字信号处理技术与数据挖掘技术相结合，面向脑机接口、语言认知、智慧诊疗及计算心理等多个交叉领域学科人才培养，同时也为相关课程提供公共实验平台。另外用于办公环境建设的设备，如空调、打印机、投影仪等 40 台，共约 16 万元。日常使用各种服务器（包括 IBM、DELL、至强、超微等）40 余台，除用于支撑实验室日常的工作环境外，还为完成具有较大计算量的科研实验任务提供支持。这些设备主要供实验室固定人员、访问人员、博士后以及学生使用，在较为空闲的时候也提供给北京大学信息科学技术学院其他单位的科研人员共享使用。

实验室的主要仪器设备为各种类型的计算机，普遍具有换代快、改造难的特点，我们采用多种方式延长设备的使用周期，如将老旧设备用于一些计算量小的专门任务。

实验室注重安全管理，每年对实验室管理者签订安全责任书，以减少使用中的安全隐患。

六、审核意见

实验室承诺所填内容属实，数据准确可靠。 数据审核人： 实验室主任： （单位公章） 2019 年 8 月 27 日
依托单位审核意见 依托单位负责人签字： （单位公章） 2019 年 8 月 29 日
主管部门审核意见 主管部门负责人签字: （单位公章） 年 月 日
评估机构形式审查意见 审核人: 年 月 日