基于知识图谱的中外自然语言处理研究的对比分析*

邱均平　方国平
(武汉大学中国科学评价研究中心　武汉 430072)

摘要: 【目的】从多角度对中外自然语言处理的发展进行对比分析。【方法】对 5 582 篇来自 CNKI、10 348 篇来自 Web of Science、5 573 篇来自与自然语言处理相关的重大国际会议文献，采用词频统计法、共现分析法相结合的方法，利用知识图谱呈现统计结果。【结果】统计结果表明，中外对自然语言处理的研究表现出极大的相似性，研究内容都集中在信息抽取、人工智能、信息检索、机器翻译、机器学习等领域。【局限】检索主题词的选取、数据清洗时的主观性给研究带来误差。【结论】对国内自然语言处理的发展提出建议。

关键词: 自然语言处理　知识图谱　信息检索　机器学习
分类号: G250

1　引　言

自然语言处理(Natural Language Processing, NLP)又称自然语言理解，发源于美国IBM沃森研究中心，是一种对自然语言信息进行处理的技术，包括自然语言理解(Natural Language Understanding, NLU) 和自然语言生成(Natural Language Generation, NLG) 两部分[1]。它是计算语言学的重要分支之一，是人工智能研究中一个十分活跃的领域[2]。成熟的自然语言处理技术可广泛用于机器自动翻译、情报检索、自动标引、自动文摘等领域。国外学者于20世纪40年代末至50年代初开始涉及该领域，国内自然语言处理研究兴起于20世纪80年代初期，虽然成绩斐然，但与国外相比仍存在差距[3]。从国内外的研究来看，自然语言处理的研究大体上经历三个时期，即60年代以关键词匹配为主流的早期、70年代以句法–语义分析为主流的中期和80年代开始走向实用化和工程化的近期[4-6]。目前，国内利用计量学的方法研究自然语言处理的相关成果并不多，其中比较显著的是中国科学技术信息研究所的祝清松[7]和中国医学科学院的李阳等[8]。他们都是采用传统文献计量的方法外加统计学的方法，研究自然语言处理年度发文量、国内主要研究机构、主要研究作者和关键词分析。但他们的研究都局限于国内自然语言处理技术的研究，并未涉足于国际；呈现分析结果的时候也是采用传统的表格形式，并且都是关键词词频作为衡量指标的主要依据。最重要的是，考虑到NLP研究领域的特殊性，即前沿和代表性的工作多在主流会议上发表，本文在统计文献信息的时候注意到这点，增加了不少国际上有关NLP的主流会议论文，这样在研究NLP当前的研究热点和前沿时，研究结果更具代表性和说服力。本文在前人研究的基础上，将词频统计法与共现分析法相结合，以知识图谱的形式，呈现国内外自然语言处理研究年代分布、力量分布以及主要研究内容，并作对比分析，对NLP研究领域有一定的参考意义。

2　数据来源与方法

为了全面了解自然语言处理的研究进展，本文选

收稿日期: 2014-03-17
收修改稿日期: 2014-11-03
*本文系国家社会科学基金项目“基于语义的馆藏资源深度聚合与可视化展示研究”(项目编号:11&ZD152)的研究成果之一。

取的外文文献来自于Web of Science(WoS)检索平台，检索年限为1990年至2013年，检索式：主题= “Natural Language Processing” or “NLP”；中文数据库为中国知网(CNKI)的中国学术文献网络出版总库，检索式为：主题词=“自然语言处理”，检索文献的年限为1990年至2013年。经过人工清洗，剔除明显与研究主题无关的记录，得到CNKI 5 582篇，WoS 10 348篇(其中包含968篇中国作者和9 380篇外国作者)，为了使研究更具科学性和代表性，本文特意增加5 573篇关于NLP的4个国际主流会议(ACL、COLING、IJCNLP和EMNLP)中被录用的论文，其中中国作者495篇，外国作者5 078篇，把国内外文献分别汇总，得到中国作者文献数7 045篇，外国作者文献数14 458篇，并作统计分析。

本文研究方法包括统计分析法、对比分析法、共现分析法。研究中使用的工具有统计软件Excel、可视化软件UCINET及其自带插件NetDraw。

3 研究年代分布

根据检索结果，笔者分别统计国内外作者每年的发文量，根据这些数据绘制折线图，如图 1 所示：

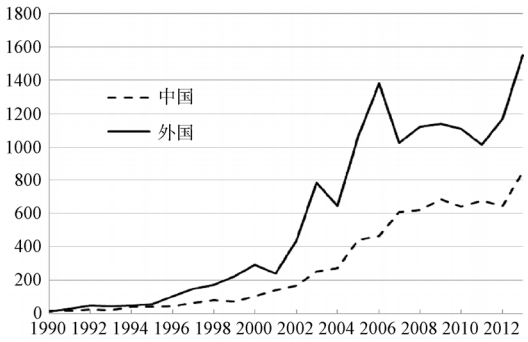

图 1 国内外自然语言处理研究文献时间分布

从图 1 中可以看出，国内对于自然语言处理技术的研究，表现为先上升再平稳再上升的整体趋势；国外则表现为先上升再下降再平稳最后再上升的趋势。整体趋势很相似，因为它们都分别经历了缓慢增长阶段、急剧增长阶段和成熟阶段，只是国外个别年份发文量有所下降。通过分析图谱，可以进一步得出以下两个结论：

(1) 国内自然语言处理的缓慢发展期长于国外(国内到 1999 年，国外到 1995 年)，并且每年文献增长量(即曲线斜率，含负增长量)小于国外。由于自然语言处理既属于人文科学研究的范畴又属于自然科学研究的内容，它涉及到语言学和计算机科学两门学科的知识，所以，可以从两个方面分析国内自然语言处理技术比国外发展缓慢的原因：

①从语言学上分析，自然语言处理技术最早来源于对西文的处理，处理西文还算比较成熟，然而面对一词多义、同音异义、笔画复杂的中文，显得有些棘手。因此，在对汉字进行预处理(包括汉字编码、汉语分词等)时，花费了很长的时间，到真正开始进行自然语言处理的研究，已经明显晚于国外[9-13]；

②从计算机科学角度分析，自然语言处理技术的发展与计算机的发展息息相关，国内计算机无论是从访问速度和存储容量都落后于国外，这对诸如语音识别、语音处理等对访问速度和存储容量要求较高的研究是不利的。

这两个原因的双重作用使得国内自然处理技术的发展落后于国外。

(2) 武汉大学邱均平[14]的《信息计量学》一书中在分析文献信息逻辑增长规律优点时提到，文献经过急剧增长的阶段后会趋向平稳，短暂平稳后又开始进入增长阶段。一方面，这意味着某一知识领域在取得重大进展后已进入相对成熟的阶段；另一方面也可能意味着该知识领域正面临新的突破，将产生出更新的分支领域，内容更新的文献又将进入一个新的急剧增长时期，此后又会进入一个稳定时期。无论从国内还是国外数据来分析，自然语言处理技术的发展符合学科领域的发展规律，自然语言处理技术的发展也从未停止，一直呈螺旋式的发展趋势，虽然发展中难免会碰见一些问题(障碍)，但是一旦得到解决和突破，又会出现新的理论研究和创新。

4 自然语言处理的研究力量分布

4.1 国家/地区力量分布

一个学科的发展往往具有一定的地域性，同一个国家或地区的学者交流频繁，信息流动大，能进一步带动整个国家或地区某学科的发展。某国家或地区某学科领域的作者数越多，发文量越大，则该国家或地区在该领域的研究实力越强。因此，可以统计每个国家的总发文量来了解该国家在该领域内的科研实力，统计结果如表 1 所示。

从数据统计来看，在统计的 106 个发文量大于 1 的国家中，发文量大于 2 000 的只有美国，而且发文量远高于其他国家，高达 2 857 篇，这说明美国在自然语

表 1　自然语言处理研究的国家/地区力量分布

国家/地区	数量	国家/地区	数量
USA	2 857	AUSTRALIA	216
CHINA	1 463	NETHERLANDS	204
GERMANY	666	BRAZIL	185
ENGLAND	631	SOUTH KOREA	163
FRANCE	456	SCOTLAND	155
JAPAN	519	POLAND	148
SPAIN	518	SWITZERLAND	123
CANADA	406	PORTUGAL	111
ITALY	360	GREECE	108
INDIA	243	TURKEY	97

言处理技术研究领域居于世界核心地位；发文量大于 1 000 篇且小于 2 000 篇的只有中国，为 1 463 篇，也远超其他国家(美国除外)，其中 968 篇来自于 WoS 检索，495 篇来自国际主流会议(上述 4 大会议收录的论文集)，这说明中国在自然语言处理技术领域除了加强自身的研究外，还加强了与国际的交流与合作。此外，在国际主流会议中录用的文章数量也呈现增长的趋势，这说明中国的很多研究成果得到了世界的认可。从表 1 可以看出德国、英国、法国、日本、西班牙、加拿大等国家的研究也较为活跃，发文总量均大于 400 篇。从国家研究力量分布可以发现，目前自然语言处理的研究主要集中在发达国家，而发展中国家研究相对比较匮乏(中国除外)。发达国家强大的计算硬件设备、雄厚的科研经费以及语言种类(尤其是英语)等条件为展开自然语言处理的研究提供很大的便利，在这种环境下，产出的科研成果就可能多于发展中国家。

4.2　研究机构分布

(1) 国内自然语言处理的研究机构分布

从文献计量学的角度来看，一个科研机构在某领域内的文献产出量的多少，在某种程度上代表该科研机构在该领域内科研实力的强弱。因此，只需要统计各科研机构的总发文量，就可以大致了解科研单位在该领域内的整体科研实力。

统计 7 045 篇(含 CNKI 的 5 582 篇，WoS 的 968 篇，国际主流会议的 495 篇)中国作者所属机构，只取发文量排名前 20 的机构(若发文量相同，则排名不分先后)作为研究对象，结果如表 2 所示。

表 2　国内自然语言处理的研究机构分布

国内机构	发文量	国内机构	发文量
哈尔滨工业大学	406	浙江大学	97
中国科学院	358	武汉大学	89
北京大学	229	新疆大学	85
北京邮电大学	214	重庆大学	82
大连理工大学	175	华中师范大学	82
清华大学	172	东北大学	80
上海交通大学	170	北京理工大学	73
苏州大学	168	厦门大学	68
山西大学	127	南京大学	62
复旦大学	121	吉林大学	62

从表 2 看出，发文量排名前 10 的研究机构中，哈尔滨工业大学和中国科学院发文量都高于 300 篇，发文量分别为 406 篇和 358 篇，由此说明，哈尔滨工业大学和中国科学院是国内自然语言处理研究的两大核心之地，在自然语言处理研究领域非常活跃，不仅如此，哈尔滨工业大学还拥有多个与 NLP 相关的实验室，如智能技术与自然语言处理研究室，它是国内最早研究自然语言处理的科研团体之一，自 20 世纪 80 年代初期以来，先后开展了俄汉机器翻译、固定段落问答、自动文摘、文本纠错、汉字智能输入、语音识别与合成、语料库多级加工、语言模型、信息检索、问答系统等有关自然语言处理的多项研究，此外还有信息检索实验室、语言语音教育部–微软重点实验室、机器智能与翻译研究室等相关实验室，强大的科研队伍和丰厚的师资力量使得哈尔滨工业大学在该领域内成果非常突出；中国科学院计算技术研究所也有专门的自然语言处理研究组，它是国际上机器翻译领域较为知名的研究团队之一，此外，中国科学院自动化研究所、声学研究所对自然语言处理的研究也较为活跃。从科研机构分布还可以看出，国内自然语言处理技术的研究以理工科类重点院校见长，其中前 10 名中有 6 所高校以理工科闻名，依次是哈尔滨工业大学、北京邮电大学、大连理工大学、清华大学、上海交通大学、复旦大学，分析其原因，还是归结于自然语言处理的研究主要是建立在计算机科学的基础上，而上面几所理工科重点院校也正是国内计算机专业较强的几所院校。

为了了解具体的科研单位，笔者进一步统计了所有的院、系、实验室、研究所，发现有 2 344 个科研单

位曾在自然语言处理领域内发过文章，排名前 20 的科研单位及其发文量如表 3 所示:

表 3　国内自然语言处理的具体科研单位分布

机构	发文量	机构	发文量
北京大学计算语言学研究所	162	苏州大学计算机科学与技术学院	45
哈尔滨工业大学计算机科学与技术学院	157	北京科技大学信息工程学院	45
上海交通大学计算机科学与工程系	82	上海交通大学计算机系	42
清华大学计算机科学与技术系	72	南京大学计算机软件新技术国家重点实验室	38
中国科学院计算技术研究所	66	四川大学计算机学院	35
复旦大学计算机科学与工程系	55	山西大学计算机与信息技术学院	35
中国科学院声学研究所	50	南京师范大学文学院	33
新疆大学信息科学与工程学院	48	东北大学自然语言处理实验室	33
中国科学院研究生院	47	中国科学技术信息研究所	33
中国科学院自动化研究所模式识别国家重点实验室	47	山西大学计算机科学系	32

其中北京大学计算语言学研究所和哈尔滨工业大学计算机科学与技术学院的发文量远高于其他科研机构。此外，还发现自然语言处理的研究力量主要分布在 3 个学科，分别是计算机科学、信息科学、语言学。前 20 名中包含 11 个计算机科学与技术、3 个信息科学、2 个语言学科研单位。即自然语言处理作为一种计算机技术，主要在计算机科学领域内研究的比较多；其次，自然语言作为一种信息载体，在信息科学(信息传播)也颇有研究；最后，自然语言本质还是属于一种对语言的研究，因此在语言学学科内也不乏研究。

(2) 国外自然语言处理的研究机构分布

统计 WoS 中检索到的 9 380 篇和 5 078 篇国际主流会议国外作者所属机构及其发文量，结果如表 4 所示。

表 4 仅显示前 20 名排名，其中发文量大于 200 的仅有 1 所，是美国加利福尼亚大学，发文量高达 222 篇，远高于第二名卡内基·梅隆大学(139 篇)。在前 20 名排名中，美国占据 12 所(含 11 所高校和 IBM 公司)，英国占据 5 所，日本、西班牙、瑞士各占 1 所，可见，在国外，自然语言研究的核心机构主要集中在美国。从机构类型来看，国外对自然语言研究除了在高校和科研单位外，还有一些企业研究也比较活跃，比如 IBM 公司，发文 82 篇，排名第 8；从科研单位层次来看，大部分是世界著名大学及知名企业，强大的科研队伍极大地推动了国外自然语言处理技术的发展；从科研机构所属国家来看，排名前 20 科研机构全部来自于发达国家，这与上文中研究力量分布研究结果相吻合。

表 4　国外自然语言处理的研究机构分布

机构	发文量	机构	发文量
UNIVERSITY OF CALIFORNIA	222	UNIVERSITY SYSTEM OF MARYLAND	69
CARNEGIE MELLON UNIVERSITY	139	UNIVERSITY OF PENNSYLVANIA	66
COLUMBIA UNIVERSITY	109	UNIVERSITAT D ALACANT	66
HARVARD UNIVERSITY	92	UNIVERSITY OF MANCHESTER	65
UNIVERSITY OF LONDON	87	UNIVERSITY OF PITTSBURGH	63
UNIVERSITY OF EDINBURGH	87	STANFORD UNIVERSITY	61
PENNSYLVANIA COMMONWEALTH SYSTEM OF HIGHER EDUCATION PCSHE	86	UNIVERSITY OF SHEFFIELD	60
INTERNATIONAL BUSINESS MACHINES IBM	82	UNIVERSITY OF ILLINOIS SYSTEM	59
UNIVERSITY OF TOKYO	81	UNIVERSITY OF GENEVA	58
MASSACHUSETTS INSTITUTE OF TECHNOLOGY MIT	78	UNIVERSITY OF CAMBRIDGE	57

4.3　作者分布

一个学科领域的发展和每个科研工作者的努力是分不开的，一个作者发文量的多少可以看出这位作者在该领域的研究地位。为此，本文分别统计了中国作者和国外作者的发文量，结果显示，国内发文量大于 5 篇的作者数有 176 人，国外有 528 人，发文量排名前 20 的作者分别见表 5 和表 6 所示。

表 5　国内自然语言处理的高产作者分布

作者	发文量	所属机构	作者	发文量	所属机构
刘挺	84	哈尔滨工业大学	孙茂松	30	清华大学
李生	70	哈尔滨工业大学	张全	29	中国科学院
刘群	63	中国科学院	王厚峰	26	北京大学
周国栋	49	苏州大学	朱靖波	25	东北大学
赵铁军	48	哈尔滨工业大学	宗成庆	25	中国科学院
俞士汶	44	北京大学	秦兵	23	哈尔滨工业大学
赵军	40	中国科学院	冯志伟	22	教育部语言文字应用研究所
陆汝占	31	上海交通大学	吴立德	22	复旦大学
王晓龙	31	哈尔滨工业大学	朱巧明	20	苏州大学
黄萱菁	30	复旦大学	何婷婷	19	华中师范大学

表 6　国外自然语言处理的高产作者分布

作者	发文量	所属机构
Friedman, Carol	57	Columbia University
Biegler, Lorenz T	54	Carnegie Mellon University
George Hripcsak	27	Columbia University
Wendy Chapman	22	University of California, San Diego
Grossmann IE	22	Carnegie Mellon University
Denny JC	25	Vanderbilt University
Fuji Ren	23	University of Tokushima
Mirella Lapata	20	University of Edinburgh
James R. Curran	20	University of Sydney
Hahn U	20	Univ Freiburg
Chew Lim Tan	19	National University of Singapore
Ren F	19	Univ Tokushima
Dan Klein	19	University of California at Berkeley
Trausan-Matu, S	18	Univ Politehn Bucuresti
LIU HF	18	Mayo Clin
Rosso P	17	Columbia Univ
Lovis C	17	Univ Hosp Geneva
Eduard Hovy	17	Carnegie Mellon University
Giorgio Satta	16	University of Padova
Taira RK	16	Univ Calif Los Angeles

(1) 国内自然语言处理作者分布

在国内 7 045 篇文献中，总共有 5 953 位作者，平均发文量为 1.18 篇。发文量排名前 5 的高产作者中有三名来自于哈尔滨工业大学，分别是刘挺、李生、赵铁军，这三位作者发文量都比较高，这为哈尔滨工业大学成为国内自然语言处理的研究的核心成员奠立了坚实的基础。在前 20 名中，有 4 位高产作者来自于中国科学院，分别是中国科学院计算技术研究所自然语言处理研究组的刘群、中国科学院自动化研究所的赵军、中国科学院声学研究所语言语音及交互信息技术部的张全、中国科学院自动化研究所的宗成庆，此外，高产作者刘群不仅在国内发文量较高，而且在自然语言处理领域内国际顶级会议中发文非常活跃，在所有国内作者中排名第一，高达 43 篇。由此可见，中国科学院无论是从“硬件”(诸如下属研究机构数量、研究环境)还是从“软件”(诸如师资力量、队伍专业化、研究成果)来说，都是国家自然语言处理技术研究的核心团队。虽然表 5 中显示大部分作者都来自于高校，但是，一些互联网企业研究也颇为活跃，排名前 30 的高产作者有三名来自企业，分别是微软亚洲研究院自然语言处理研究组负责人周明、百度基础技术领域首席科学家王海峰、百度自然语言处理技术负责人吴华，这些作者在国际主流会议中发文量比较多，许多研究成果都很前沿。

为了进一步了解国内作者的合作情况，选取排名前 50 的作者，建立共现矩阵，利用 UCINET 绘制作者合著网络知识图谱，取 K 值为 10，表示两个作者合作发文量大于 10 才在图中显示，弱合作不会在图谱中显示，图谱如图 2 所示:

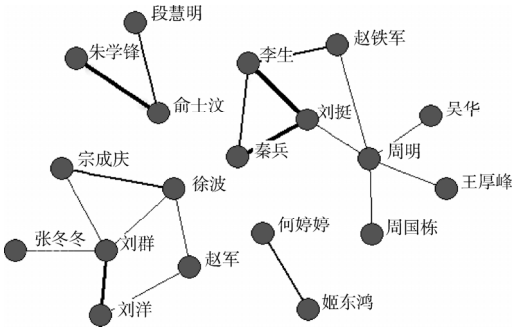

图 2　国内自然语言处理高产作者合著网络图谱

图 2 中线条的粗细代表合著者一起发文量的多少，线条越粗表示合作越紧密。可以看出，目前国内主要有三大合著比较紧密的学术团体：第一为中国科学院

的学术团体，以刘群为核心，包括宗成庆、徐波、赵军、张冬冬、刘洋，其中刘群和刘洋合作最为紧密，共同发文 36 篇，该学术团体发文多在信息抽取、问答系统、机器翻译、文本自动分类等领域；第二为哈尔滨工业大学为主的学术团体，其中还包括百度公司、微软亚洲研究院、苏州大学、北京大学的高产作者成员，其中刘挺、李生、秦兵三位作者的合作最为紧密，共同发文 25 篇。从这个学术团体的良好合作可以看出：自然语言处理技术的研究既要加强内部信息共享(这样有利于加强信息整合、节约研究时间和成本，从而提高整个所属机构的研究水平)，又要加强外部信息交流(吸收其他机构的经验，借鉴他人的新理论和技术，拓宽视野，促进整个学科的发展)，既要注重理论研究，又要兼顾实践创新，如加强高校与企业的合作；第三为北京大学学术团体，包括俞士汶、朱学锋、段慧明，三位作者共同发文 12 篇，他们的研究主要是从语言学角度入手构建语料库，对现代汉语语料库的系统构建做出了很大的贡献。还有武汉大学姬东鸿和华中师范大学的何婷婷合作也较为紧密。

(2) 国外自然语言处理作者分布

在国外 14 458 篇论文中，总共有 8 903 位作者，平均发文量为 1.62 篇，比国内稍高。其中发文量最多的作者是来自于美国哥伦比亚大学 Friedman, Carol 教授，她主要致力于自然语言处理、人工智能、信息检索、文本挖掘、信息抽取等方面的研究，在国外自然语言处理领域内具有一定的影响力。通过查找前 20 名作者信息，发现国外作者研究力量分布较均匀，核心作者并没有集中在某几所高校，大部分都来自不同的科研单位，而中国高产作者比较集中。此外，作者所属学科来源也比较广泛，除了传统的计算机科学，还有生物信息科学、化学工程等学科领域。为了了解作者合著情况，与国内相似，利用 UCINET 绘制出作者合著网络，K 值为 10，如图 3 所示。

从图 3 中可以看出国外合著网络中高频合作相对比较密集，说明国外更注重科研之间的合作。从作者研究领域来看，既有同一个研究领域，比如 Biegler, LT 和 Kumar, A 都是研究化学信息学的作者；又有来自于不同领域的作者合作，比如作者 Haug, PJ 主要从事化学信息学研究，而 Chapman, WW 主要从事生物信息学；此外还有不同国家之间的作者合作，如

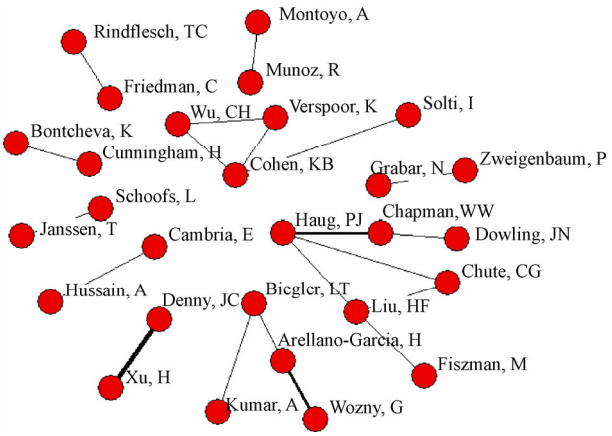

图 3 国外自然语言处理高产作者合著网络图谱

Cambria, E 来自于新加坡，Hussain, A 来自于英国。自然语言处理是一门交叉学科，涉及到信息科学、计算机科学、语言学等众多学科，既要加强学科内部信息共享，又要增进学科之间合作；既要兼顾国内合作，又要着眼国外交流。如果只注重某个领域内部(比如计算机学科)的合作，对整个自然语言处理领域的发展是不利的，这应该成为国内外共同关注的问题。

5 自然语言处理的研究内容分布

5.1 基于词频统计的研究内容分布

关键词是一篇文献主要内容的结晶，如果在若干篇文献中含有某个关键词的频次越多，说明这个关键词是这些文献的研究热点。为此，利用 Excel 自编 VBA 小程序[15]分别统计了国内 7 045 篇(外文文献对关键词进行翻译后统一中文合并)和国外 14 458 篇文献的所有关键词，分别得 9 472 个和 14 988 个原始关键词，对这些关键词进行人工核查，对同义词、中英文表示等进行去重叠加；删除本身研究对象，如自然语言处理；删除诸如企业、国际会议等明显无关词语，英文文献做词汇词根处理，最后分别得到 9 418 和 14 924 个有效关键词，这些关键词中低频关键词占据很大的数量，比如国外 14 924 个关键词中，有 11 297 个关键词词频为 1。基于此，可以得出检索得到的文献内容耦合性大，各篇文献含有多个相同的关键词，研究内容相似度较大，这对本文的研究是有利的，因为本文需要的数据就是那些少部分的高频关键词，这些高频词汇很大程度上代表着文章的主要研究内容。把有效关键词按词频逆序排列，国内排名前 10 的词频分别为信息检

索(227)、语料库(196)、本体(184)、机器翻译(181)、人工智能(177)、信息抽取(170)、中文信息处理(148)、机器学习(138)、命名实体识别(133)、问答系统(129)；国外排名前 10 的词频分别为 Information Retrieval (317)、Machine Learning(308)、Information Extraction (247)、Ontology(242)、Text Mining(216)、Machine Translation (202)、Semantic Web(178)、Artificial Intelligence(165)、Question Answering(163)、Corpus (128)。

通过对比国内外关键词统计，可以发现国内外在自然语言处理领域内研究的内容极具相似性，都分别集中在信息检索、机器翻译、机器学习、人工智能、本体、信息抽取、问答系统、语料库等几个领域，可以说国内在研究内容上与国外是接轨的。与之不同的是，国内词频较高的还有“中文信息处理”，这说明，自然语言处理作为一门技术，解决的是语言问题，除了研究与国外相关研究热点外，还要顾及自身需要解决的问题，目前国内在这方面做得很好，比如关键词中的“维吾尔语”、“藏语”都充分说明，国内在研究自然语言处理的同时善于利用技术来解决自身的语言问题。中国科学院自动化研究所宗成庆指出，我国拥有 56 个民族，几十种语言共存，藏、蒙、维、哈、朝等这些少数民族语言更是在被广泛使用，如何实现多语言信息的自由交流是一项重要而艰巨的任务[16]。这就需要一种技术(自然语言处理技术)把这些以各种语言文字表达的文本内容准确、高效地破解，挖掘和抽取其中的有用信息。利用技术解决国家实际问题，也是科研工作者们追求的共同目标。

5.2　基于关键词共现的研究热点分布

利用词频统计法统计关键词只能比较粗略地看出某个领域内的主要研究内容，如果需更进一步了解主要研究热点(前沿)，需要用关键词共现法[17-18]。关键词共现源于文献同被引，如果某个指定关键词与其他任意关键词在若干篇文献中出现的频率越大，说明这个指定的关键词必然是这些相似文献共同研究的热点问题，因此可以通过分别建立国内外关键词的共现矩阵，来了解国内外在自然语言处理领域内的研究热点。取词频排名前 100 的关键词，建立关键词网络共现图谱，如图 4 和图 5 所示。

(1) 国内自然语言处理的研究热点

从图 4 可以清晰地看出目前国内自然语言的研究热点主要集中在以下 5 点:

①中文信息处理。图 4 中与“中文信息处理”共现频次高的关键词有中文信息、汉语自动分词、自动分词、中文分词、分词、词义消歧等，可见，汉语分词技术是中文信息处理的研究热点。香港理工大学学者在 COLING 会议中发表的一篇题为《Book Review: Introduction to Chinese Natural Language Processing》[19]的文章中指出，汉语与其他语言的差别在于词法、句法和语义，其中最主要的不同在于词法。因此如何进行汉语分词，一直是自然语言处理技术在中文信息处理中的重要研究热点同时也是难点之一。

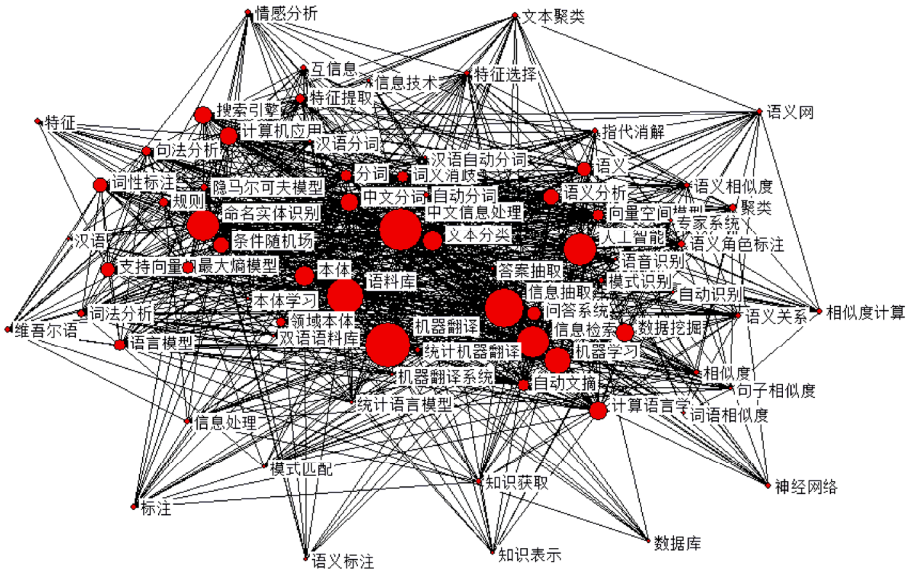

图 4　国内高频关键词共现网络图谱

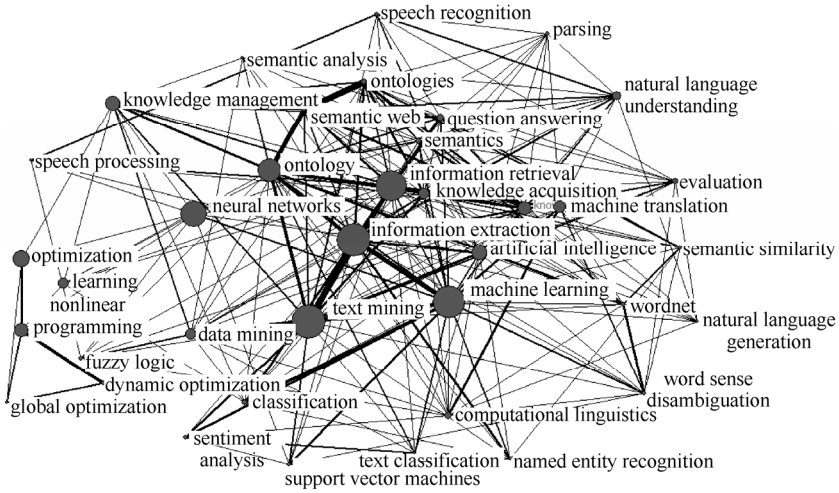

图 5　国外高频关键词共现网络图谱

②语料库。自然语言处理的资源建设主要是语料库与知识库的建设。语料库是存放语言材料的仓库，自然语言处理领域的语料库则是按照一定原则组织在一起的大规模真实自然语言数据的集合。语料库主要用于研究自然语言规律，特别是统计语言学模型的训练及相关系统的评价与评测[20]。

③信息抽取。从图 4 可以看出，与信息抽取聚类的关键词有答案抽取、问答系统、知识获取、知识库、专家系统。因此目前信息抽取的研究热点集中在如何从已有的知识库(专家库)中自动抽取答案，形成智能的问答系统，这种研究也是为了突破不断兴起的社会化问答平台人工作答的瓶颈[21-22]。仅从国际主流会议中国内作者发文量来统计，目前有 12 篇论文是关于“Question Answering”，其中有 5 篇文献是关于自动问答系统的研究。

④命名实体识别。它是指识别出文本中特定的实体，是信息抽取、机器翻译、自动问答等多种自然语言处理技术的基础。但是由于受中文自身特点的限制，中文命名实体识别一直相当困难。为了促进其他技术和应用的发展，研究中文命名实体的识别技术很有意义，也非常重要。与“命名实体识别”聚类的关键词有规则、条件随机场、隐马尔可夫模型、最大熵模型。孙镇和王惠临在《命名实体识别研究进展综述》[23]一文中提到，当前中文命名实体识别的方法主要有规则、条件随机场、隐马尔可夫模型、最大熵模型。可见，目前关于中文命名实体识别的研究主要集中在方法领域内，仅从国际 NLP 主流会中国内作者发表的论文来统计，有关主题为“Named Entity Recognition”的发文高达 21 篇，位居主流会议论文统计榜首。

⑤机器翻译。机器翻译一直是自然语言处理技术研究的热点之一，长久不衰。从关键词聚类可以看出，目前机器翻译研究热点有两个，分别是机器翻译系统构建(包括系统评价)和统计机器翻译[21]。机器翻译系统构建已经产品化，市面上有很多免费的产品，比如 Google、百度、有道等，不仅如此，由中国科学院团队研发的维汉、藏汉、蒙汉等少数民族语言翻译和英汉翻译系统取得了很好的测评效果[16]。由此可见国内在机器翻译系统构建领域内的研究还是很活跃；统计机器翻译也是目前国内自然语言处理领域的研究热点之一，仅从国际 NLP 主流会中国内作者发表的论文来统计，有关主题为“Statistical Machine Translation”的发文高达 29 篇。统计机器翻译(Statistical Machine Translation)是机器翻译的一种，也是目前非限定领域机器翻译中性能较佳的一种方法。统计机器翻译的基本思想是通过对大量的平行语料进行统计分析，构建统计翻译模型，进而使用此模型进行翻译[21]。

(2) 国外自然语言处理的研究热点

从共现网络可以看出，国外对自然语言处理技术的研究主要集中在信息抽取、机器学习、文本挖掘、本体、信息检索、知识表示、语义网、人工智能、知识获取等方面。有些研究热点词与国内相同，关键词“信息检索”是国内外在自然语言处理研究领域内共同的主要研究内容，这说明自然语言处理作为一种语言处理技术，在信息检索中具有重要的应用[24-31]。除此之外，图 5 中还有人工智能、神经网络、语义分析、语义相似性等关键词也较为突出，这些关键词都说明，国外在自然语言处理领域内的研究已经由原来的语形研究上升到语义或者语用层次的研究[32-33]。文本挖掘、数据挖掘、信息抽取都是数据获取的重要方法，这些方法对语料库、知识库的建立都起着重要作用[34-36]。

对比图 4 和图 5 可以看出，国内的图谱要比国外的密集一些，分析其可能的原因有以下两点：第一，国外的低频词比国内多。在国外统计的 14 924 个关键词中，光词频为 1 的关键词就有 11 297 个，此外还有词频为 2，词频为 3 等，到达指定阈值的关键词数量

就偏少，图谱就显得略稀疏；第二，国内外期刊文献在选取关键词的作者主观依据、关键词数量也可能有偏差。

6　结　语

本文利用词频统计法以及可视化的方法，对比分析了国内外自然语言处理领域内的年代分布、研究力量分布、研究内容分布。从整体上来看，国内与国外在自然语言处理研究领域具有很大的相似性，包括年代分布、研究力量分布、研究内容分布，同时也看出了国内在自然领域内的卓越成就，尤其是在国际顶级会议中的发文量(包括录用的长短论文)逐年增加，这说明，国内的研究成果得到了世界的赞同与认可。但是本文在研究过程中也存在以下不足:

(1) 在获取数据来源时，使用“自然语言处理”作为主题词搜索，实际上 NLP 是个很大的范围，用主题词“自然语言处理”很难涵盖“机器翻译”、“信息检索”等相关领域，如需做深入研究，还需要区分和界定这些关键词之间的联系，比如信息检索哪些方面用到了自然语言处理技术，它们之间交叉部分主要在哪里，这些问题对笔者以后深入研究提供了一些思维脉络;

(2) 选择的期刊数据库来源也有限，一些自然语言处理刊文量较大的国内期刊，比如《情报学报》，并没有收录到 CNKI; 国外还有 EI、ISTP 等数据库也收录了部分的文献，因此数据来源难免有偏颇之处。

基于上文统计分析，笔者对国内自然语言处理技术的发展提出如下建议:

(1) 从年代分布来看，国内应该继续加强与国际合作交流，比如积极参加 NLP 相关的国际学术会议。虽然中国在自然语言处理技术方面取得不少成果，但跟国际相比还存在一定差距，从年代分析来看，中国要想在该领域内取得更进一步的发展就必须善于学习和借鉴国际一些先进技术或者成熟的理论，以免走重复路、走弯路，可以把更多的时间用到解决中国的实际问题中，比如热点关键词中的“中文信息处理”、“中文分词”、“维吾尔语”等充分体现中国需要运用自然语言处理技术解决众多自身的语言问题。既要积极地走出去，又要合理地引进来。

(2) 从研究力量分布来看，国内应加强与自然语言处理相关的不同学科之间的作者合作，不能仅仅停留在某个学科内，而应该进行较多的跨学科合作。自然语言处理和计算机科学联系最为紧密，并且和信息科学、语言学的研究密不可分，如果各学科各成一家，不能及时了解相互的研究进展，那么对整个自然语言处理的发展是很不利的。

(3) 抓住机遇，与时俱进。自然语言处理技术是随着时代的发展不断发展的，随着云计算和大数据时代的到来，自然语言处理技术将会面对新的挑战，但挑战与机遇共存，科研者们应该善于运用工具，提出新的对策及应变方法。抓住机遇，乘势而上，自然语言处理技术一定会不断有新的突破，也一定会为社会、经济发展乃至整个民族振兴做出卓越的贡献[19]。

(致谢: 感谢匿名外审专家们对本文的前期多次修改提出的宝贵意见，在接受他们意见的同时，对他们的严谨治学的科研态度也表示非常钦佩！此外，也感谢编辑部的多次修改意见！)

参考文献：

[1] Allen J. 自然语言理解[M]. 第二版. 刘群，张华平，骆卫华，等译. 北京: 电子工业出版社，2005. (Allen J. Natural Language Understanding [M]. The 2nd Edition. Translated by Liu Qun, Zhang Huaping, Luo Weihua, et al. Beijing: Publishing House of Electronics Industry, 2005.)

[2] 杨国文. 自然语言理解[J]. 外语教学与研究，1987(3) : 28-31, 81. (Yang Guowen. On Understanding Natural Language [J]. Foreign Language Teaching and Research, 1987(3): 28-31, 81.)

[3] 冯志伟. 自然语言处理的学科定位[J]. 解放军外国语学院学报，2005, 28(3): 1-8. (Feng Zhiwei. Academic Position of Natural Language Processing [J]. Journal of PLA University of Foreign Languages, 2005, 28(3): 1-8.)

[4] 冯志伟. 自然语言处理的历史与现状[J]. 中国外语，2008, 5(1): 14-22. (Feng Zhiwei. The Past and Present of Natural Language Processing [J]. Foreign Languages in China, 2008, 5(1): 14-22.)

[5] 曹佩. 论自然语言处理[J]. 信息与电脑，2010(5): 187. (Cao Pei. On the Natural Language Processing [J]. China Computer and Communication, 2010(5): 187.)

[6] 殷杰，董佳蓉. 论自然语言处理的发展趋势[J]. 自然辩证法研究，2008, 24(3): 31-37. (Yin Jie, Dong Jiarong. The Development Trend of the Natural Language Processing [J]. Studies in Dialectics of Nature, 2008, 24(3): 31-37.)

[7] 祝清松. 我国自然语言处理研究的文献计量分析[J]. 情报

杂志，2009，28(S2)：32-34. (Zhu Qingsong. Bibliometric Analysis of Natural Language Processing in China [J]. Journal of Information, 2009, 28(S2): 32-34.)

[8] 李阳，许培扬．我国自然语言处理研究文献计量分析[J]. 中华医学图书情报杂志，2012，21(2)：65-70. (Li Yang, Xu Peiyang. Research on Natural Language Processing in China: A Bibliometric Analysis [J]. Chinese Journal of Medical Library and Information Science, 2012, 21(2): 65-70.)

[9] 田瑛．运用语义分析解决自然语言处理中的英语歧义问题[J]．语文学刊(外语教育与教学)，2009(5)：14-15. (Tian Ying. The Use of Natural Language Processing Semantic Analysis to Resolve Ambiguity in English [J]. Journal of Language and Literature Studies, 2009(5):14-15.)

[10] 吴巧玲．中文分词算法在自然语言处理技术中的研究及应用[J]．信息与电脑，2011(12)：39-40. (Wu Qiaoling. Chinese Word Segmentation Algorithm and Its Application in Natural Language Processing Techniques [J]. China Computer & Communication, 2011(12): 39-40.)

[11] 许坤，冯岩松，赵东岩，等.面向知识库的中文自然语言问句的语义理解[J]．北京大学学报：自然科学版，2014，50(1)：85-92. (Xu Kun, Feng Yansong, Zhao Dongyan, et al. Automatic Understanding of Natural Language Questions for Querying Chinese Knowledge Bases [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 85-92.)

[12] 才让加．面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究[J]．中文信息学报，2011，25(6)：157-161. (Tse Ring'rgyal. Research on Large-scale Sino-Tibetan Bilingual Corpus Construction for Natural Language Processing [J]. Journal of Chinese Information Processing, 2011, 25(6): 157-161.)

[13] 孟维娟．自然语言处理中的歧义[J]．上海电机学院学报，2006，9(S1)：16-19. (Meng Weijuan. Simple Analysis of Ambiguity in Natural Language Processing [J]. Journal of Shanghai Dianji University, 2006, 9(S1): 16-19.)

[14] 邱均平．信息计量学[M]．武汉：武汉大学出版社，2007. (Qiu Junping. Informetrics [M]. Wuhan: Wuda Publishing House, 2007.)

[15] 化柏林．用 VBA 实现文献计量分析研究中的数据预处理技术[J]．现代图书情报技术，2007(3)：69-72. (Hua Bolin. Implementation of Preprocess Technology in Bibliometric and Analytic Research via VBA [J]. New Technology of Library and Information Service, 2007(3): 69-72.)

[16] 中国科学报：《让机器会"说"多种语言》[EB/OL]. [2014-01-03]. http://www.ia.cas.cn/xwzx/mtsm/201401/t20140103_4009934.html. (Chinese Science News: Let the Machine "Say" in Many Languages [EB/OL]. [2014-01-03]. http://www.ia.cas.cn/xwzx/mtsm/201401/t20140103_4009934.html.)

[17] 寇继虹，楼雯．基于知识图谱的E-learning研究的可视化分析[J]．电化教育研究，2010(9)：20-25. (Kou Jihong, Lou Wen. Visual Analysis of E-learning Based on Knowledge Map [J]. E-education Research, 2010(9): 20-25.)

[18] 杨皓东，江凌，李国俊．国内自然语言处理研究热点分析——基于共词分析[J].图书情报工作，2011，55(10)：112-117. (Yang Haodong, Jiang Ling, Li Guojun. The Hotspot of Natural Language Processing in China: Based on Co-word Analysis [J]. Library and Information Service, 2011, 55(10): 112-117.)

[19] Wong K, Li W, Xu R, et al. Book Review: Introduction to Chinese Natural Language Processing [J]. Computational Linguistics, 2010, 36(4): 777-780.

[20] 李生．自然语言处理的研究与发展[J]．燕山大学学报，2013，37(5)：377-384. (Li Sheng. Research and Development of Natural Language Processing [J]. Journal of Yanshan University, 2013, 37(5): 377-384.)

[21] 王献昌，史晓东，陈火旺．机器翻译与自然语言处理的现状与趋势[J]．计算机科学，1992，19(3)：1-3. (Wang Xianchang, Shi Xiaodong, Chen Huowang. The Current Situation and Trend of Machine Learning and Natural Language Processing [J]. Computer Science, 1992, 19(3): 1-3.)

[22] Liu J, Wang Q, Lin C, et al. Question Difficulty Estimation in Community Question Answering Services [C]. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA. Association for Computational Linguistics, 2013: 85-90.

[23] 孙镇，王惠临．命名实体识别研究进展综述[J]．现代图书情报技术，2010(6)：42-47. (Sun Zhen, Wang Huilin. Overview on the Advance of the Research on Named Entity Recognition [J]. New Technology of Library and Information Service, 2010(6): 42-47.

[24] 崔新华．自然语言处理在信息检索中的应用研究[J]．贵阳学院学报：自然科学版，2012，7(3)：37-40. (Cui Xinhua. Natural Language Processing Applications in Formation Retrieval Research [J]. Journal of Guiyang College: Natural Science, 2012, 7(3): 37-40.)

[25] 左远清，周洞汝，王波．自然语言处理在搜索引擎信息检索中的应用[J].现代计算机，2002(7)：28-29，44. (Zuo Yuanqing, Zhou Dongru, Wang Bo. Application of Natural Language Processing in Information Retrieve by Search Engineer [J]. Modem Computer, 2002(7): 28-29, 44.)

[26] 于志敏，张文德．基于自然语言处理的信息检索[J]．山东科技大学学报：自然科学版，2006，25(1)：122-124. (Yu

Zhimin, Zhang Wende. Information Retrieval Based on Natural Language Processing [J]. Journal of Shandong University of Science and Technology: Natural Science, 2006, 25(1): 122-124.)

[27] 蔡霞，张森. 自然语言理解在 Web 数据挖掘中的应用[J]. 计算机工程与设计, 2003, 24(11): 1-3. (Cai Xia, Zhang Sen. Practice of Web Mining Based on Nature Language Understanding [J]. Computer Engineering and Design, 2003, 24(11): 1-3.)

[28] Lewis D D, Jones K S. Natural Language Processing for Information Retrieval [J]. Communications of the ACM, 1996, 39(1): 92-101.

[29] Voorhees E M. Natural Language Processing and Information Retrieval [A].// Information Extraction [M]. Springer Berlin Heidelberg, 1999: 1-17.

[30] Doszkocs T E. Natural Language Processing in Information Retrieval [J]. Journal of the American Society for Information Science, 1986, 37(4): 191-196.

[31] 黄敏. 自然语言处理与信息检索[J]. 图书情报工作, 2001, 45(4): 41-44, 65. (Huang Min. Natural Language Processing and Information Retrieval [J]. Library and Information Service, 2001, 45(4): 41-44, 65.)

[32] 蔡艳婧，程显毅，潘燕. 面向自然语言处理的人工智能框架[J]. 微电子学与计算机, 2011, 28(10): 173-176, 180. (Cai Yanjing, Cheng Xianyi, Pan Yan. A Framework of Artificial Intelligence Oriented Natural Language Processing [J]. Microelectronics & Computer, 2011, 28(10): 173-176, 180.)

[33] Obermeier K K.Natural Language Processing Technologies in Artificial-Intelligence -The Science and Industry Perspective [M]. Ellis Horwood, 1989.

[34] Costantino M, Morgan R G, Collingham R J, et al. Natural Language Processing and Information Extraction: Qualitative Analysis of Financial News Articles [C]. In: Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering. IEEE, 1997: 116-122.

[35] Coulet A, Cohen K B, Altman R B. The State of the Art in Text Mining and Natural Language Processing for Pharmacogenomics [J]. Journal of Biomedical Informatics, 2012, 45(5): 825-826.

[36] Zhou G, Liu F, Liu Y, et al. Statistical Machine Translation Improves Question Retrieval in Community Question Answering via Matrix Factorization [C]. In: Proceedings of Annual Meeting of the Association of Computational Linguistics. 2013.

作者贡献声明：

邱均平：提出研究命题，设计研究思路及研究方法，论文起草，最终版本修订；

方国平：论文修改，文献调研，原始数据获取、清洗、分析。

(通讯作者：方国平 E-mail: 1259297235@qq.com)

The Comparative Analysis of Natural Language Processing Research at Home and Abroad Based on Knowledge Mapping

Qiu Junping　Fang Guoping
(Research Center for Chinese Science Evaluation, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper makes a comparative analysis to the development of natural language processing at home and abroad from multi-angle. [**Methods**] The literatures are from CNKI (5 582), Web of Science (10 348) and major international conferences on natural language processing (5 573). Use word frequency statistics and co-occurrence analysis as main research methods and use knowledge maps to show statistical results. [**Results**] The result shows that the study of natural language processing performance at home and abroad has a great similarity. Their research focuses on the domains of information extraction, artificial intelligence, information retrieval, machine translation, machine learning and so on. [**Limitations**] There are some limitations in this paper, such as the choice of subject term, the error resulting from the subjectivity to data cleaning. [**Conclusions**] According to the results, several recommendations are made on the development of natural language processing.

Keywords: Natural language processing　Knowledge mapping　Information retrieval　Machine learning