自然语言处理研究报告

清华-中国工程院知识智能联合研究中心

2018年7月

目录

第1章 自然语言处理概念篇
第2章 自言语言处理技术篇
第3章 自然语言处理人才篇
第4章 自然语言处理应用篇
第5章 自然语言处理趋势篇

前 言

自然语言处理（NLP）是人工智能的一个重要应用领域，也是新一代计算机必须研究的课题。它的主要目的是克服人机对话中的各种限制，使用户能用自己的语言与计算机对话。

本研究报告对自然语言进行了简单梳理：

- 首先对自然语言处理进行**定义**，接着对自然语言的**发展历程**进行了梳理，对我国自然语言处理**现状**进行了简单介绍，对自然语言处理**业界**情况进行介绍。
- 其次对自然语言处理研究中的重要**技术**进行介绍。
- 然后利用AMiner大数据对自然语言处理领域**专家**进行深入挖掘，对国内外自然语言处理**知名实验室**及其主要负责人进行介绍。
- 自然语言处理在现实生活中应用广泛，目前的应用集中在**语言学、数据处理、认知科学以及语言工程**等领域，在介绍相关应用的基础上，对机器翻译未来的**发展趋势**做出了相应的预测。

目录

第1章 自然语言处理概念篇

第2章 自言语言处理技术篇

第3章 自然语言处理人才篇

第4章 自然语言处理应用篇

第5章 自然语言处理趋势篇

本章目录

第1节 自然语言处理概念

第2节 自然语言处理发展历程

第3节 我国自然语言处理现状

第4节 自然语言处理业界发展

什么是NLP？

自然语言处理

- 用计算机对自然语言的形、音、义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工。

两个流程

- 自然语言理解&自然语言生成

自然语言理解

- 计算机能够以自然语言文本来表达给定的意图

自然语言生成

- 计算机能够理解自然语言文本的意义

自然语言处理表现形式

- 机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等。

NLP判别标准

在人工智能领域或者是语音信息处理领域中，学者们普遍认为采用**图灵试验**可以判断计算机是否理解了某种自然语言，具体的判别标准有以下几条：

- 问答：机器人能正确回答输入文本中的有关问题；
- 文摘生成：机器有能力生成输入文本的摘要；
- 释义：机器能用不同的词语和句型来复述其输入的文本；
- 翻译：机器具有把一种语言翻译成另一种语言的能力。

NLP发展历程

20世纪50年代-70年代	• 图灵测试的提出——自然语言处理思想的开端 • 基于规则的方发——理性主义思潮
20世纪70年代以后	• 语料库不断丰富 • 基于统计的方法——IBM华生实验室起了推动作用 • 理性主义思潮向经验主义思潮过渡
2008年至今	• 深度学习与自然语言处理相结合

我国NLP发展现状

20世纪90年代以后，中国NLP研究高速发展，呈现出**商品化、创新化**的特征。

研究内容： 基础性研究（消除歧义、语法形式化等）

应用性研究（信息检索、文本分类、机器翻译等）

语音和文本是两类研究的重点

智能检索类研究近年逐渐升温

研究周期： 技术开发周期较短（1-3年）

语言资源库搭建较为困难（10年左右）

国家扶持力度大： 国家自然科学基金、社会科学基金、863项目、973项目等

NLP业界发展

自然语言处理

微软亚洲研究院

- 语音翻译：2017年全面采用神经网络机器翻译
- 机器翻译：将知识图谱纳入神经网络机器翻译规划语言理解的过程中
- 人机对话：小冰小娜进展极大

Google

- 机器翻译：2017年宣布实现完全基于attention的transformer网络架构
- 知识图谱：自动挖掘新知识的准确程度、文本中命名实体的识别等技术处于领先地位
- 语音识别：2012年将神经网络应用于这一领域

Facebook

- 机器翻译：2017年使用全新的卷积神经网络进行翻译，以9倍于以往循环神经网络的速度实现了当时最高的准确率
- 文本处理：基于2016年发布的FastText，开发了有效的方法和轻量级工具
- 语音识别：2018年初开发了wav2letter，这是一个简单高效的端到端自动语音识别（ASR）系统

百度

- 机器翻译：发布了世界上首个线上神经网络翻译系统，并获得2015年度国家科技进步奖

阿里巴巴

- 电商平台中构建知识图谱实现智能导购
- 全网用户兴趣挖掘
- 客服场景中打造机器人客服

腾讯

- 机器翻译：2017年翻译君上线“同声传译”新功能
- 基于文智API可以实现搜索、推荐、舆情、挖掘等功能
- AI Lab研究领域包括计算机视觉、语音识别、自然语言处理、机器学习等

京东

- 京东AI开放平台：由模型定制化平台和在线服务模块构成，在线服务模块包括计算机视觉、语音交互、自然语言处理和机器学习等
- 合作机构：南京大学、斯坦福大学等院校

科大讯飞

- 2017年，晓译翻译机1.0plus将神经网络翻译系统由在线系统转化为离线系统
- 2015年在由美国国家标准技术研究院组织的机器翻译大赛中取得全球第一的成绩

目录

第1章 自然语言处理概念篇

第2章 自言语言处理技术篇

第3章 自然语言处理人才篇

第4章 自然语言处理应用篇

第5章 自然语言处理趋势篇

本章目录

第1节 自然语言处理基础技

第2节 自然语言处理应用技术

NLP技术分类

基础技术

- 词法与句法分析
- 语义分析
- 语篇分析
- 知识图谱
- 语言认知模型
- 语言知识表示与深度学习

应用技术

- 机器翻译
- 信息检索
- 情感分析
- 自动问答
- 自动文摘
- 信息抽取
- 信息推荐与过滤
- 文本分类与聚类
- 文字识别

NLP基础技术

词法分析

- 主要任务：词性标注和词义标注
- 词性标注方法：基于规则和基于统计

句法分析

- 主要任务：判断句子的句法结构和成分，明确各成分的相互关系
  分类：完全句法分析、浅层句法分析
- 策略：“先句法后语义”、“句法语义一体话”（占主流）

语义分析

- 根据句子的句法结构和句子中每个实词的词义推导出来能够反映这个句子意义的某种形式化表示

语用分析

- 人对语言的具体运用，是对自然语言的深层理解。

篇章分析

- 对段落和整篇文章进行理解和分析

NLP应用技术（一）

机器翻译

概念

- 通过特定的计算机程序将一种书写形式或声音形式的自然语言，翻译成另一种书写形式或声音形式的自然语言

方法

- 基于理性的研究方法—基于规则的方法
- 基于经验的研究方法—基于统计的方法
- 与深度学习相结合

分类

- 语音翻译—亚马逊的Alexa、苹果的Siri、微软的Cortana等、语音同传技术的应用
- 图像翻译—谷歌等公司拥有能够让用户搜索或者自动整理没有识别标签的照片的技术
- 医疗创业公司利用计算机阅览X光照片、MRI和CT照片
- 对机器人、无人机以及无人驾驶汽车的改进至关重要
- VR翻译等

NLP应用技术（二）

信息检索

概念

- 从相关文档集合中查找用户所需信息的过程

原理

- “存”：对信息进行收集、标引、描述、组织，进行有序的存放
- “取”：按照某种查询机制从有序存放的信息集合（数据库）中找出用户所需信息或获取其线索
- 检索成功：将用户输入的检索关键词与数据库中的标引词进行对比，二者匹配成功时检索成功
- 检索结果按照与提问词的关联度输出，供用户选择，用户采用“关键词查询+选择性浏览”的交互方式获取信息。

NLP应用技术（三）

情感分析

概念

- 通过计算技术对文本的主客观性、观点、情绪、极性的挖掘和分析，对文本的情感倾向做出分类判断

应用

- 评论机制的App中应用较为广泛
- 互联网舆情分析中情感分析起着举足轻重的作用
- 选举预测、股票预测等领域

NLP应用技术（四）

自动问答

概念

- 利用计算机自动回答用户所提出的问题以满足用户知识需求的任务。

分类

- 检索式问答：通过检索和匹配回答问题，推理能力较弱
- 知识库问答：web2.0的产物，用户生成内容是其基础，Yahoo！Answer、百度知道等是典型代表
- 社区问答：正在逐步实现知识的深层逻辑推理

工作流程

- 首先要正确理解用户所提出的问题，
- 抽取其中关键的信息，在已有的语料库或者知识库中进行检索、匹配，
- 将获取的答案反馈给用户

NLP应用技术（五）

自动文摘

概念

- 运用计算机技术，依据用户需求从源文本中提取最重要的信息内容，进行精简、提炼和总结，最后生成一个精简版本

特点

- 压缩性
- 内容完整性
- 可读性

分类

- 基于统计的机械式文摘：简单容易实现，是目前主要被采用的方法，但是结果不尽如人意
- 基于意义的理解式文摘：建立在对自然语言的理解的基础之上的，接近于人提取摘要的方法，难度较大

NLP应用技术（六）

社会计算

概念

- 在互联网的环境下，以现代信息技术为手段，以社会科学理论为指导，帮助人们分析社会关系，挖掘社会知识，协助社会沟通，研究社会规律，破解社会难题

社会媒体

- 文本属性：草根性，字数少、噪声大、书写随意、实时性强
- 社会属性：社交性，在线、交互
- 检索成功：将用户输入的检索关键词与数据库中的标引词进行对比，二者匹配成功时检索成功
- 典型社会媒体：Twitter、Facebook、微信、微博

应用

- 金融市场采用社会计算方法探索金融风险和危机的动态规律
- 社会安全：把握舆情、引导舆论
- 军事方面：许多国家加大投入力度扶持军事信息化的发展

NLP应用技术（七）

信息抽取

概念

- 从文本中抽取出特定的事实信息。这些被抽取出来的信息通常以结构化的形式直接存入数据库，可以供用户查询及进一步分析使用，为之后构建知识库、智能问答等提供数据支撑

原理

- 利用自然语言处理的技术，包括命名实体识别、句法分析、篇章分析与推理以及知识库等，对文本进行深入理解和分析完成信息抽取工作

应用

- 信息抽取技术对于构建大规模的知识库有着重要的意义，但是目前由于自然语言本身的复杂性、歧义性等特征，而且信息抽取目标知识规模巨大、复杂多样等问题，使得信息抽取技术还不是很完善

目录

第1章 自然语言处理概念篇

第2章 自言语言处理技术篇

第3章 自然语言处理人才篇

第4章 自然语言处理应用篇

第5章 自然语言处理趋势篇

本章目录

第1节 全球学者情况概览

第2节 华人库学者情况概览

本章节所用概念说明

- **全球学者概况所用数据**是由AMiner基于发表于国际期刊会议的学术论文，对自然语言处理领域全球h-index排序top1000的学者进行计算分析所得。
- **华人库学者概况所用数据**是由AMiner基于论文数据整理了自然语言处理华人专家库，其中包括了来自NUS、HKUS、THU、PKU、FDU等知名高校以及百度、科大讯飞、微软等公司的367位专家学者。
- **h-index：**国际公认的能够比较准确地反映学者学术成就的指数，计算方法是该学者至多有h篇论文分别被引用了至少h次。
- **注：**在我们的AMiner完整报告中，统计了近十年在ACL、EMNLP、NAACL、COLING等4个会议在近5年累计发表10次以上论文的学者（包括刘群、刘挺、周明、黄萱菁等人），并对这些学者及其所属实验室进行介绍。完整报告请点击文末链接下载

NLP全球学者情况概览

NLP学者全球分布图

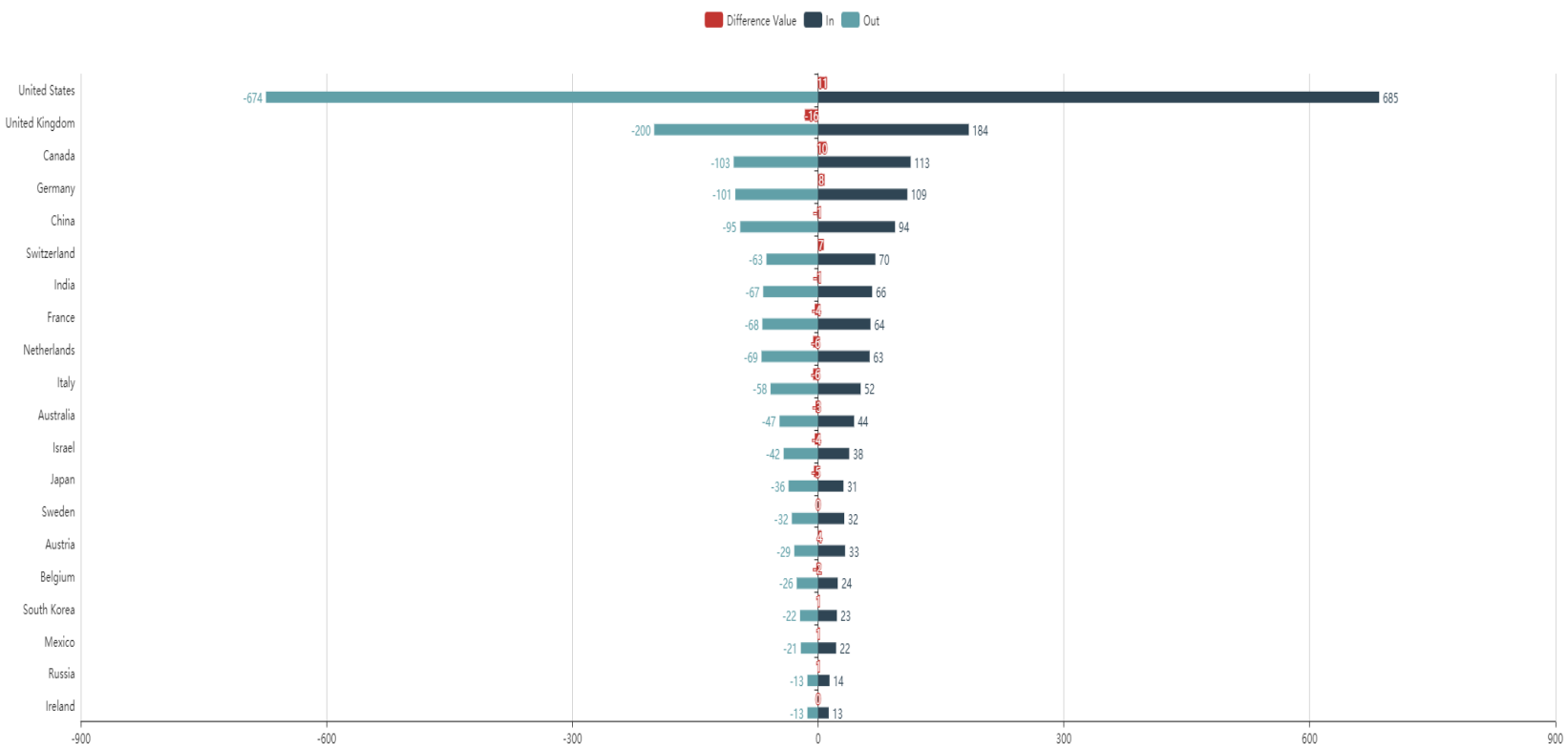
从国家来看：

- 美国自然语言处理研究学者聚集最多
- 英国、德国、加拿大和意大利紧随其后

从地区来看：

- 美国东部是自然语言处理人才的集中地
- 西欧、美国西部等其他先进地区也吸引了大量研究者

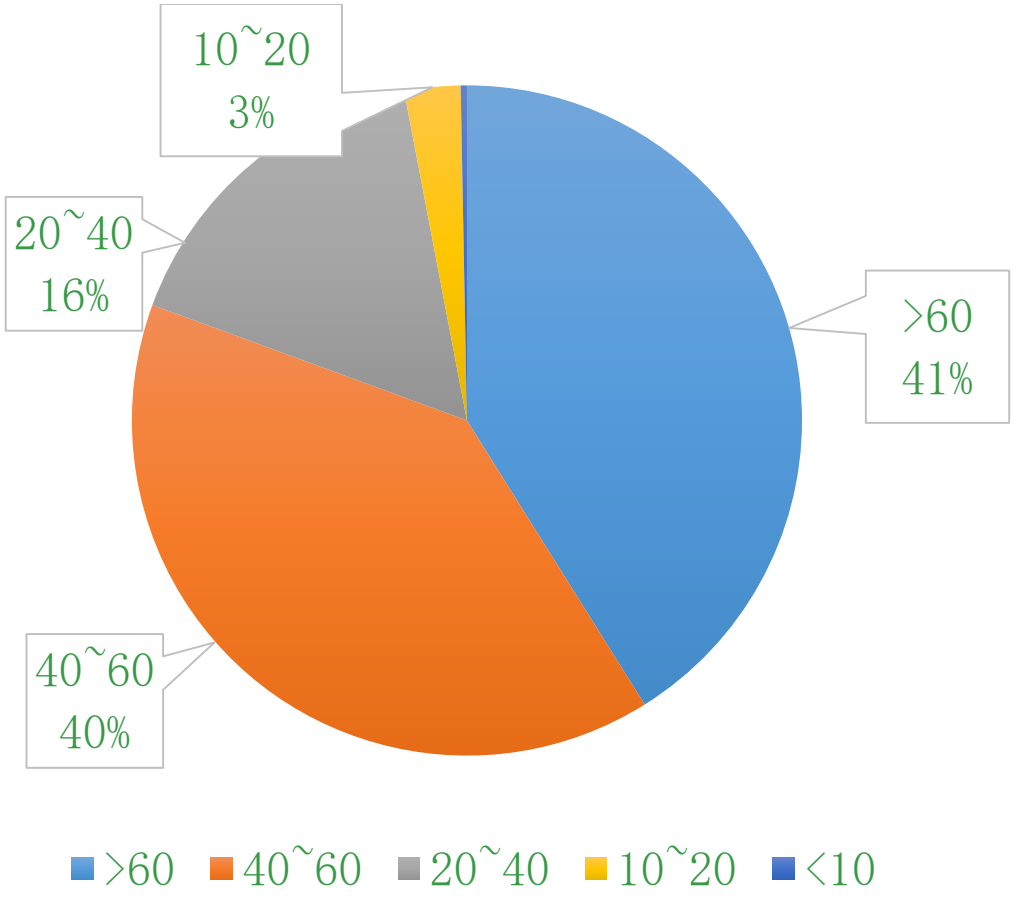
NLP全球学者情况概览



NLP学者顺逆差图

- 各国自然语言处理顶尖人才的流失和引进相对比较均衡
- 美国是自然语言处理领域人才流动大国，人才输入和输出幅度都大幅度领先，人才流入略大于流出。
- 英国、德国、加拿大和中国等国落后于美国，其中英国和加拿大有轻微的顶尖人才流失现象。

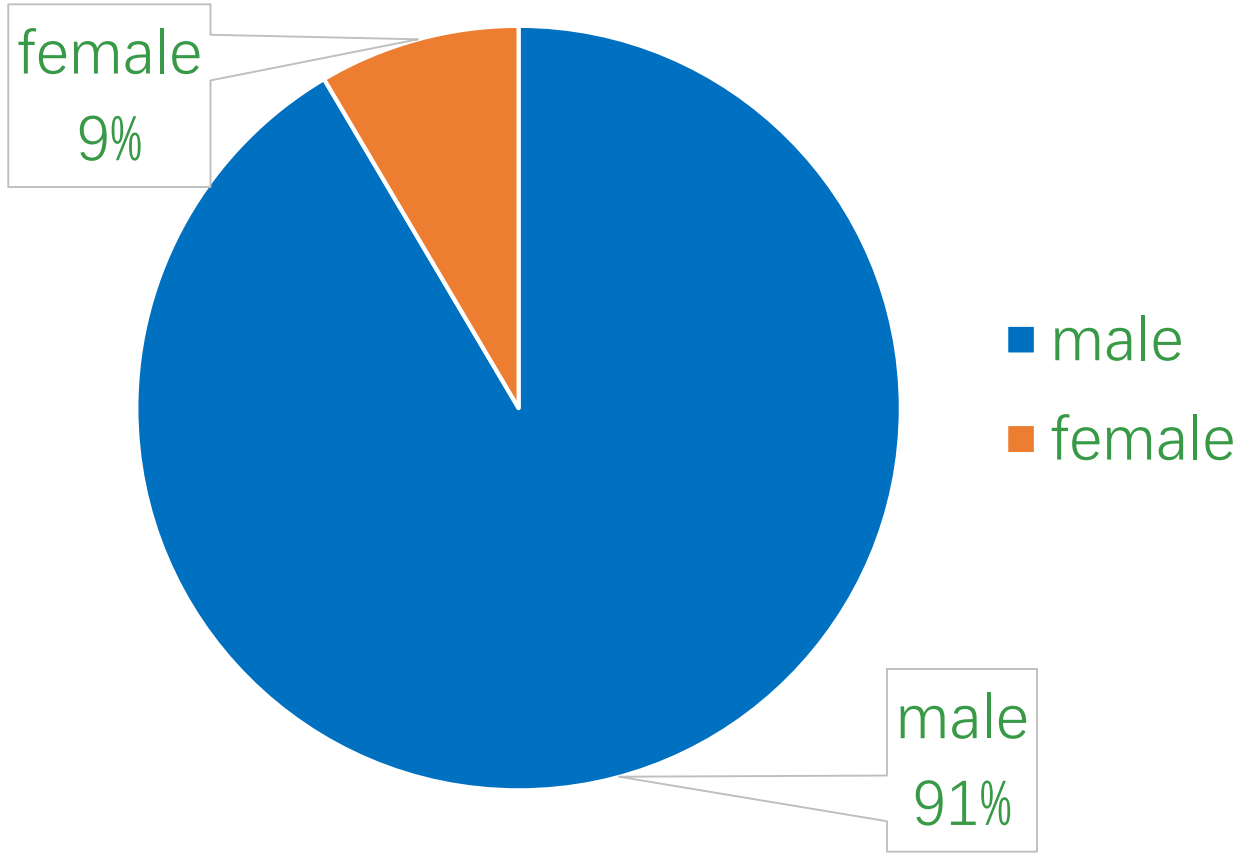
NLP全球学者情况概览



自然语言处理顶尖学者h-index分布

- 全球自然语言处理顶尖学者的h-index平均数为**59**
- h-index指数大于60的学者最多占比**41%**
- h-index指数在40到60之间的学者次之，占比**40%**

NLP全球学者情况概览



自然语言处理顶尖学者性别分布

- 全球自然语言处理顶尖学者男性占比91%，女性占9%

NLP华人库学者情况概览

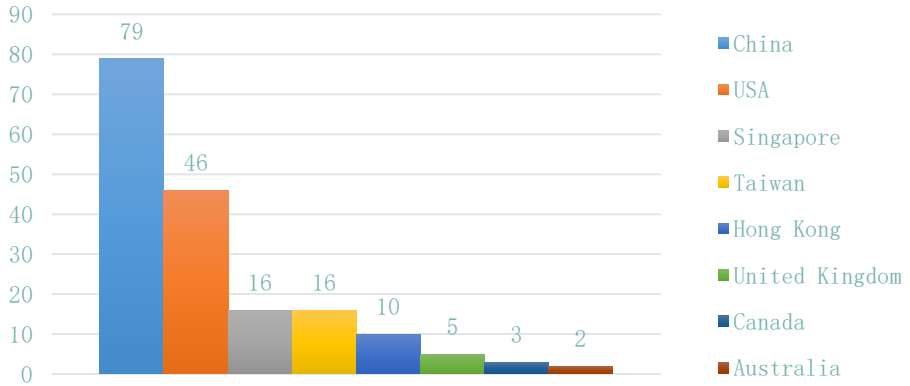
AMiner自然语言处理华人库专家全球分布

AMiner自然语言处理华人库专家国内分布

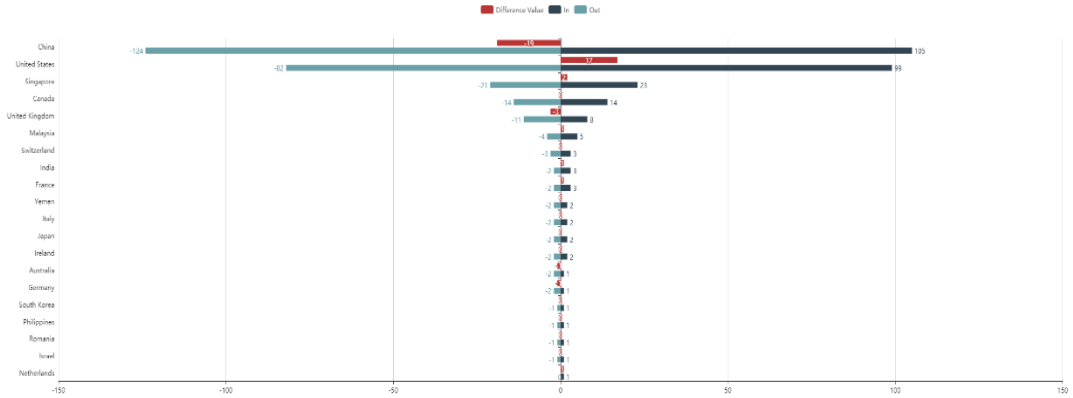
- 自然语言处理领域中华人专家在中国最多，美国次之。
- 从地区来看，中国大陆是自然语言处理华人人才的最主要聚集地，尤其是北京、哈尔滨及东南沿海地区等具有自然语言处理学术基础的地区。
- 美国东部和西部等其他地区排在其后。

NLP华人库学者情况概览

AMiner自然语言处理华人库专家地区统计



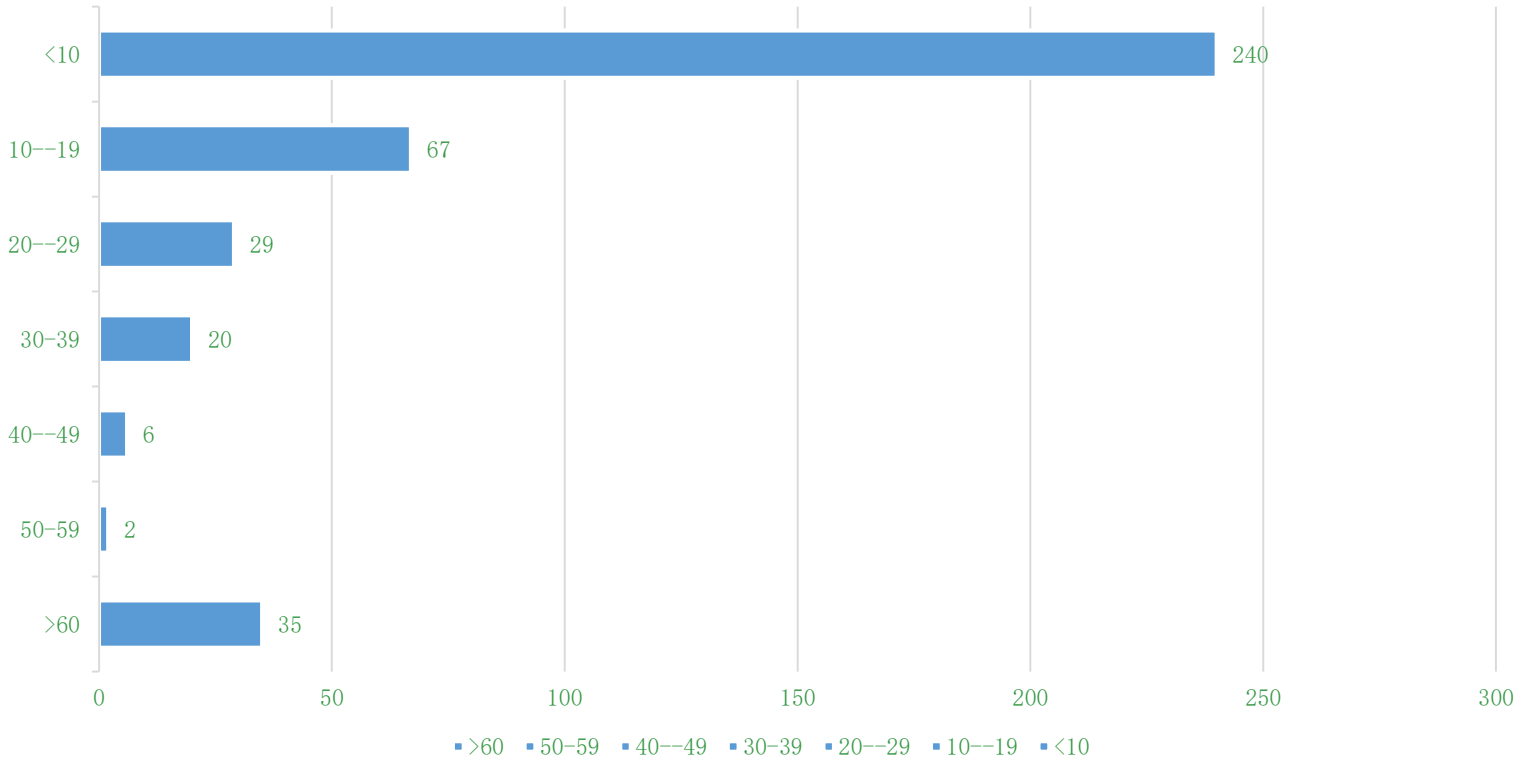
AMiner自然语言处理华人库专家迁徙图



华人专家在中国流出量大于流入量，美国则正好相反。

这也说明就自然语言处理领域而言，中国对人才的吸引力要小于美国。

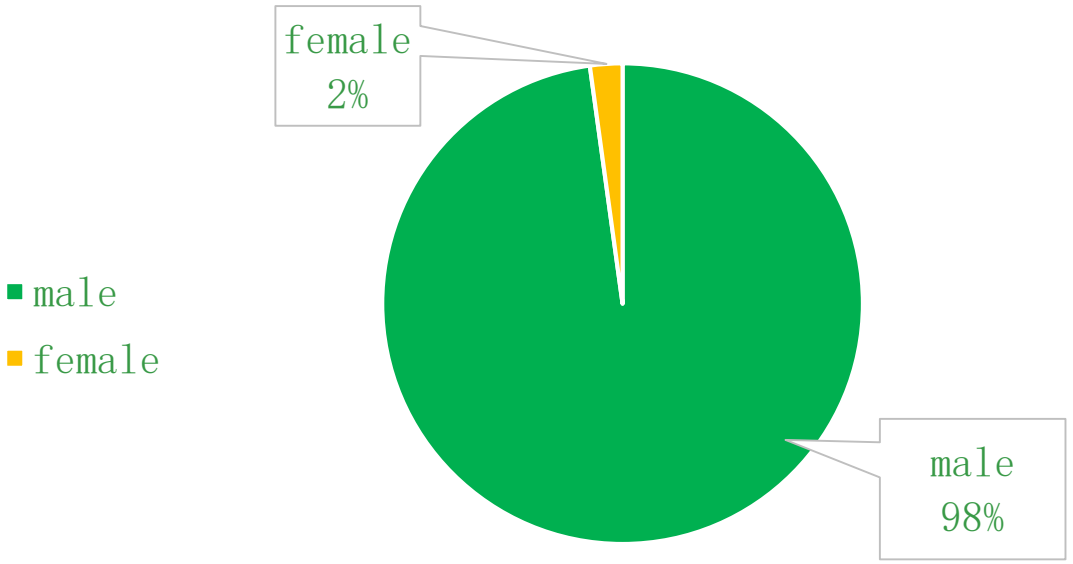
NLP华人库学者情况概览



自然语言处理华人库专家h-index分布

- h-index指数的平均数为14，这一数值远远低于自然语言处理全球top1000学者h-index指数平均数。
- h-index指数<10的专家人数最多，占比60%
- 10-19次之，占比17%
- >60的专家占比仅占9%
- 自然语言处理华人专家整体水平低于自然语言处理领域全球top1000的学者，尤其是在h-index指数>60的学者方面有所欠缺。

NLP华人库学者情况概览



自然语言处理华人库专家性别统计

- AMiner自然语言处理华人库367位专家中:
- 男性专家占98%,
- 女性专家占2%
- 二者比例约为49:1

目录

第1章 自然语言处理概念篇

第2章 自言语言处理技术篇

第3章 自然语言处理人才篇

第4章 自然语言处理应用篇

第5章 自然语言处理趋势篇

本章目录

第1节 知识图谱

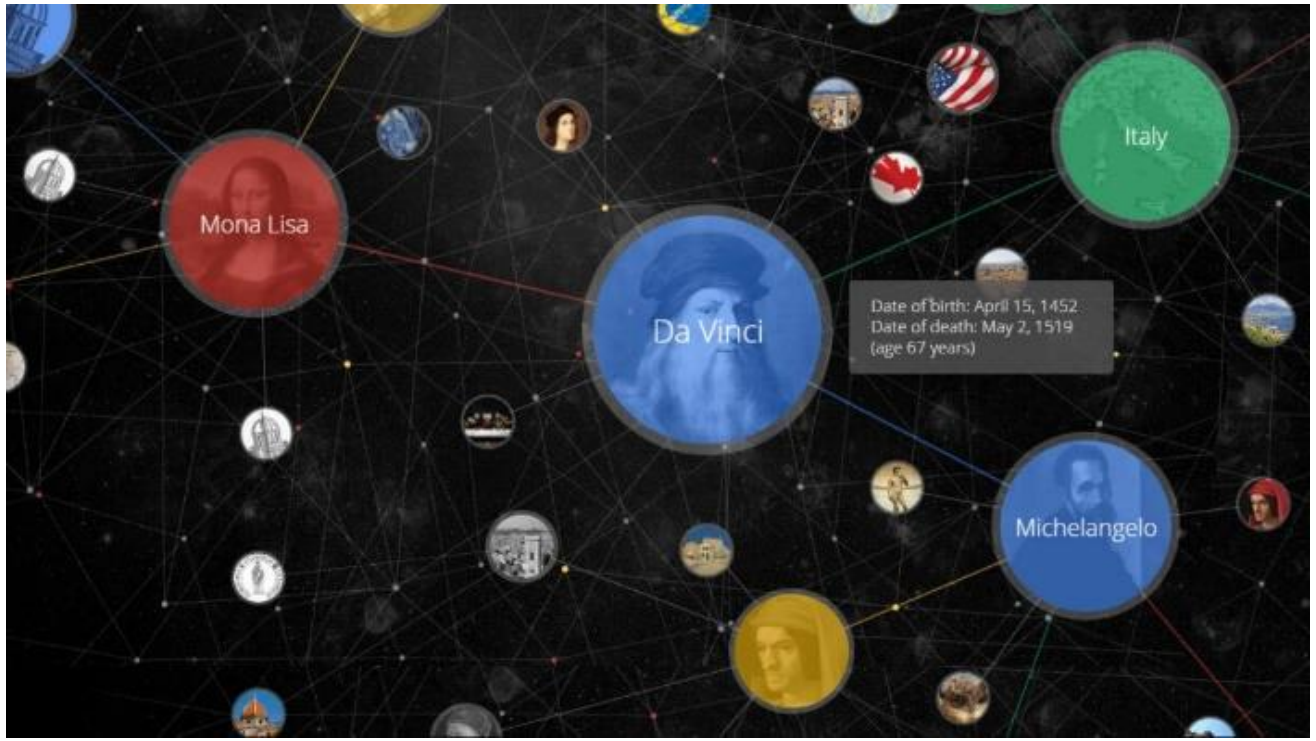
第2节 机器翻译

第3节 聊天机器人

第4节 搜索引擎

第5节 推荐系统

知识图谱



- **语义搜索：**利用建立大队莫知识库对搜索关键词和文档内容进行语义标注，改善搜索结果，如谷歌、百度等在搜索结果中嵌入知识图谱
- **知识问答：**基于知识库的问答，通过对提问句子的语义分析，在将其解析为结构化的询问，在已有的知识库中获取答案
- **基于知识的大数据分析决策：**一般起到辅助决策作用。Netflix公司利用其订阅用户的注册信息以及观看行为构建的知识图谱来决定《纸牌屋》拍摄

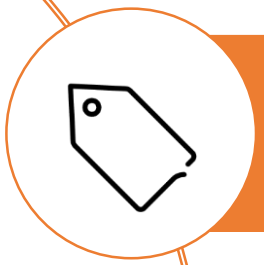
机器翻译

- **科大讯飞**：晓译翻译机1.0plus将世界上最先进的神经网络翻译系统优化为离线系统
- **阿里巴巴**：2017年初正式上线自主开发的神经网络翻译系统
- **腾讯**：2017年翻译君上线同声传译新功能
- **搜狗**：2017年乌镇世界互联网大会上展示机器同传技术；2018年上线翻译宝，在硬件领域开始探索

聊天机器人

- **概念：**能通过聊天app、聊天窗口或语音唤醒app进行交流的计算机程序，是被用来解决客户问题的智能数字化助手
- **特点：**成本低、高效且持续工作
- **对话机器人：**Siri、小娜等
- **智能问答系统：**电商网站的应用如京东客服jimi等

文本分类

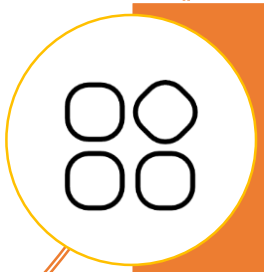


概念

- 根据文档的内容或者属性，将大量的文档归到一个或多个类别的过程

关键问题

- 如何构建一个分类函数或分类模型，并利用这一分类模型将未知文档映射到给定的类别空间



应用

- 垃圾电子邮件检测
- 门户网站每天产生的信息分繁杂多，文本分类技术尤为重要

搜索引擎

- **涉及技术：**词义消歧、句法分析、指代消解等
- **功能：**不单单是帮助用户找到答案，还能帮助用户找到所求，连接人与实体世界的服务
- **基本模式：**自动化地聚合足够多的内容，对之进行解析、处理和组织，响应用户的搜索请求找到对应结果返回

推荐系统

起源

- 1992年Goldberg提出的Tapestry，这是一个个性化邮件推荐系统，第一次提出了协同过滤的思想

技术

- 数据、算法、人机交互、数据挖掘技术、信息检索技术以及计算统计学等

应用

- 音乐电影的推荐、电子商务产品推荐、个性化阅读、社交网络好友推荐等场景

目录

第1章 自然语言处理概念篇

第2章 自言语言处理技术篇

第3章 自然语言处理人才篇

第4章 自然语言处理应用篇

第5章 自然语言处理趋势篇

本章目录

第1节 文本理解与推理：浅层分析向深度理解迈进

第2节 对话机器人：实用化、场景化

第3节 NLP+行业：与专业领域深度结合

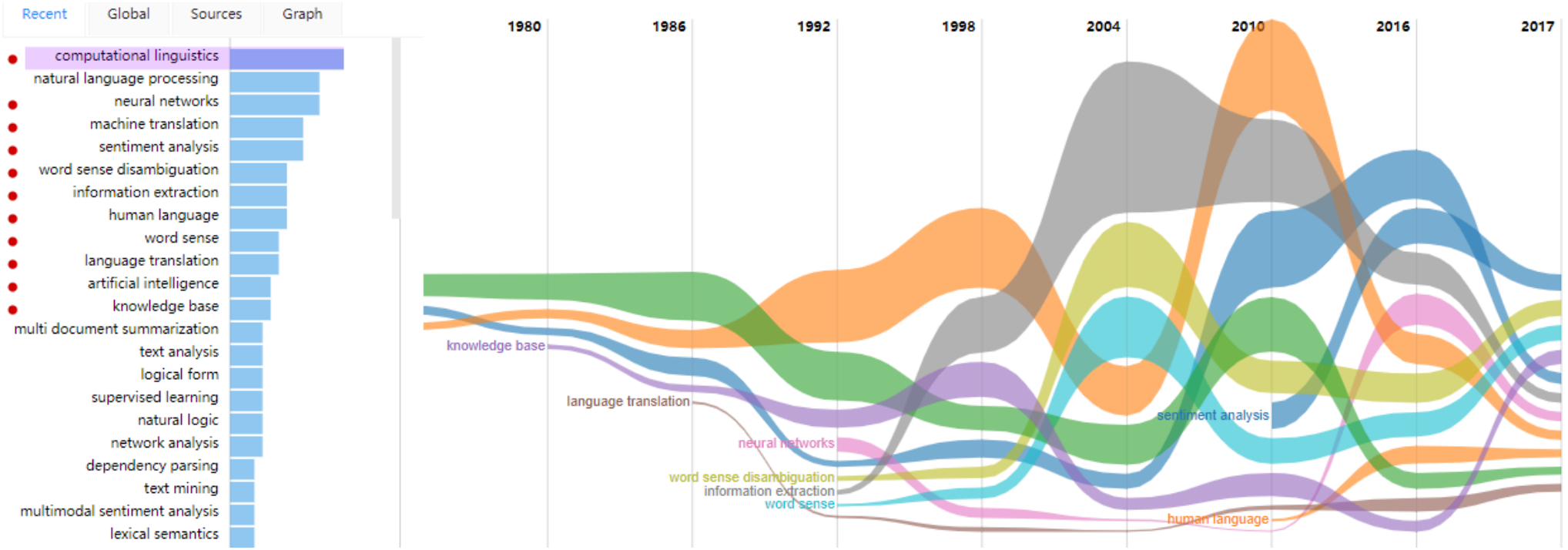
第4节 学习模式：先验语言模式与深度学习结合

第5节 文本情感分析：事实性文本到情感性文本

NLP热点图说明

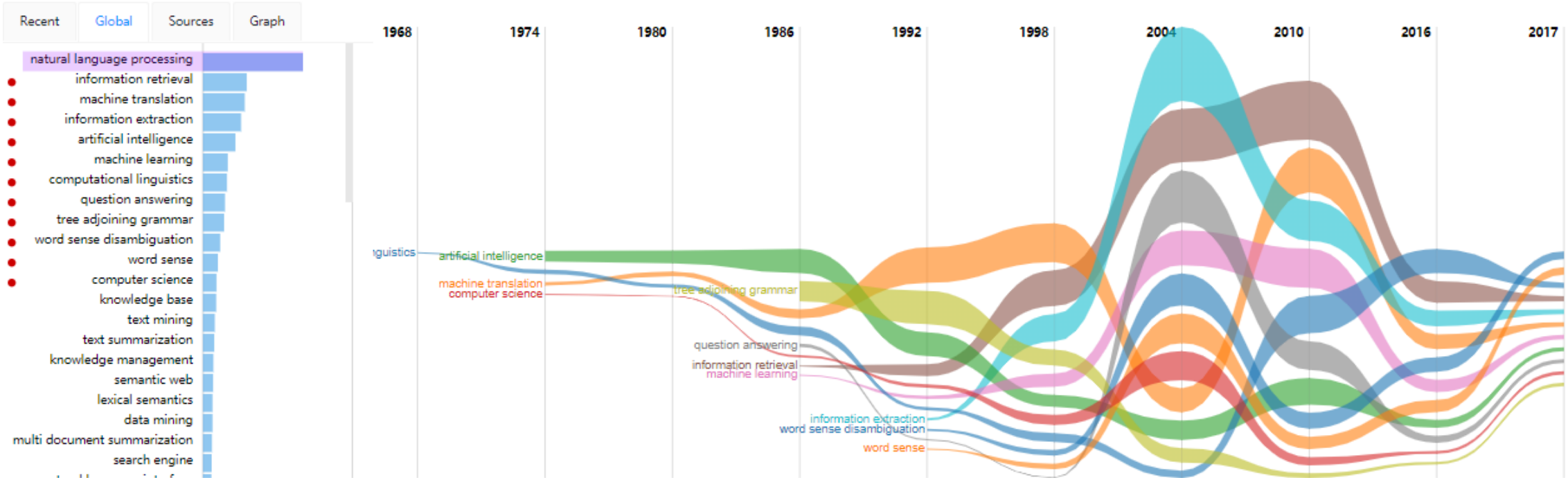
- **绘制方法：**通过对1994-2017年间自然语言处理领域论文的挖掘，总结出二十多年来，自然语言处理的领域关键词主要集中在计算机语言、神经网络、情感分析、机器翻译、词义消歧、信息提取、知识库和文本分析等领域。
- **目的**：旨在基于历史的科研成果数据的基础上，对自然语言处理热度甚至发展趋势进行研究。
- **含义：**图中，每个彩色分支表示一个关键词领域，其宽度表示该关键词的研究热度，各关键词在每一年份（纵轴）的位置是按照这一时间点上所有关键词的热度高低进行排序。

NLP近期热点图



情绪分析、词义消歧、知识库和计算机语言学是近期研究热点

NLP全局热点图



词义消歧、词义理解、计算机语言学、信息检索和信息提取是自然语言处理全局热点

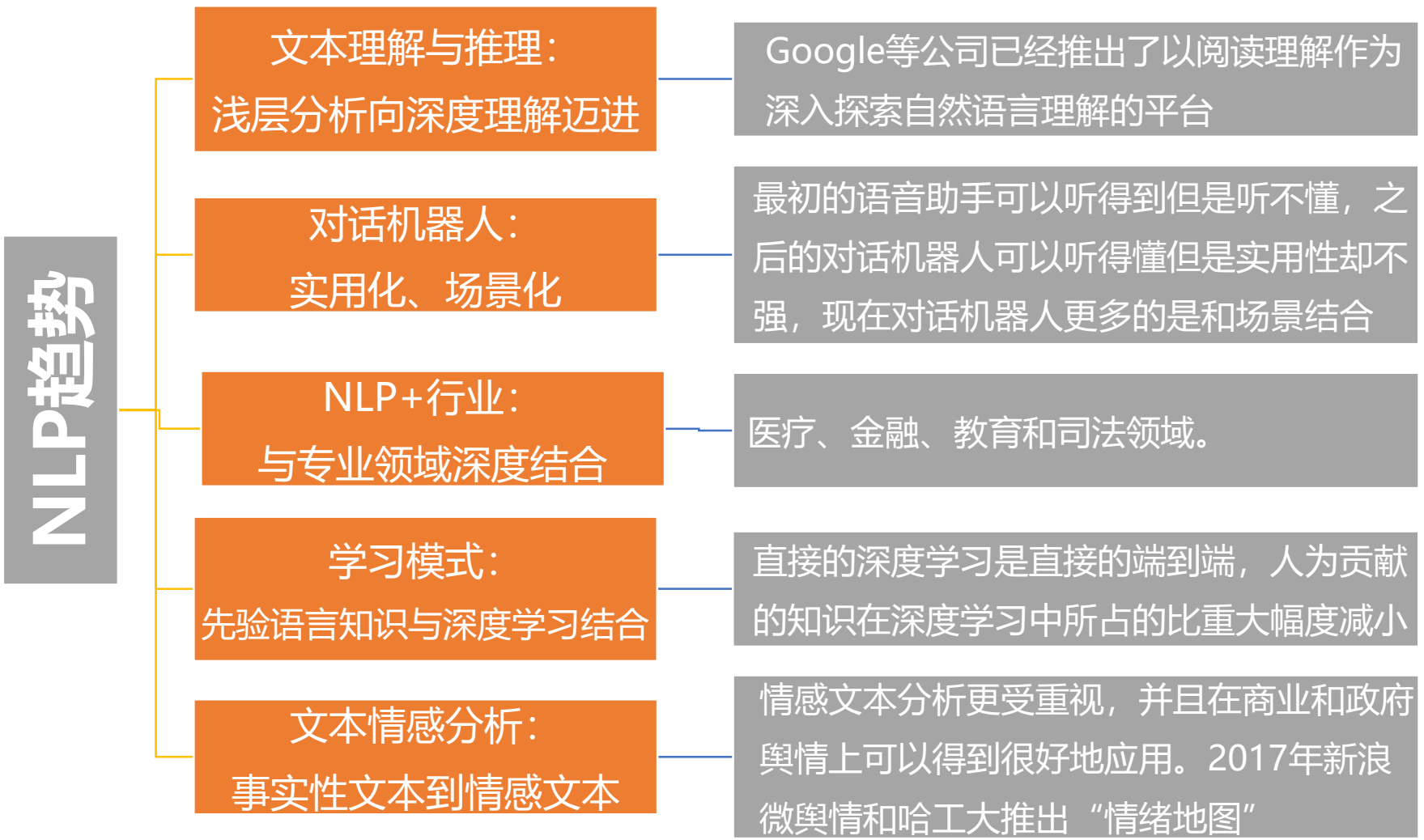
NLP趋势预测

在微博@ArnetMiner中发起了关于NLP处理未来发展趋势的投票，得到了如下结果。

文本理解与推理：浅层分析到深度理解	135（28.1%）
对话机器人：实用化场景化	83（17.3）
NLP行业：与专业领域结合	74（15.4）
学习模式：先验语言知识与深度学习结合	45（9.4%）
文本情感分析：事实性文本到情感性文本	43（9%）
语言知识：人工构建到自动构建	25(5.2%)
信息检索：跨语言、多媒体	23(4.8%
文本生成：规范文本到自由文本	15（3.1%）
NLP平台化：封闭到开放	13（2.7%）
对抗训练思想的应用	9（1.9%）

- 共有465人次参与了投票
- 文本理解与推理由浅层分析到深度理解有135人次支持，占比28.1%
- 对话机器人实用化、场景化，NLP行业与专业领域结合，学习模式由先验语言知识与深度学习结合以及文本情感分析由传统媒体到社交媒体依次排列，分别占比17.3%、15.4%、9.4%和9%。

NLP趋势预测


NLP趋势
文本理解与推理：
浅层分析向深度理解迈进
Google等公司已经推出了以阅读理解作为深入探索自然语言理解的平台
对话机器人：
实用化、场景化
最初的语音助手可以听得到但是听不懂，之后的对话机器人可以听得懂但是实用性却不强，现在对话机器人更多的是和场景结合
NLP+行业：
与专业领域深度结合
医疗、金融、教育和司法领域。
学习模式：
先验语言知识与深度学习结合
直接的深度学习是直接的端到端，人为贡献的知识在深度学习中所占的比重大幅度减小
文本情感分析：
事实性文本到情感文本
情感文本分析更受重视，并且在商业和政府舆情上可以得到很好地应用。2017年新浪微舆情和哈工大推出“情绪地图”

感谢阅读

完整报告请在“学术头条”公众号中下载