2018

自然语言处理研究报告

AMiner 研究报告第八期

清华-中国工程院知识智能联合实验室

2018 年 7 月

目录

图表目录

订阅公众号

摘要

自然语言处理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，包括以下内容：

自然语言处理概念。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。

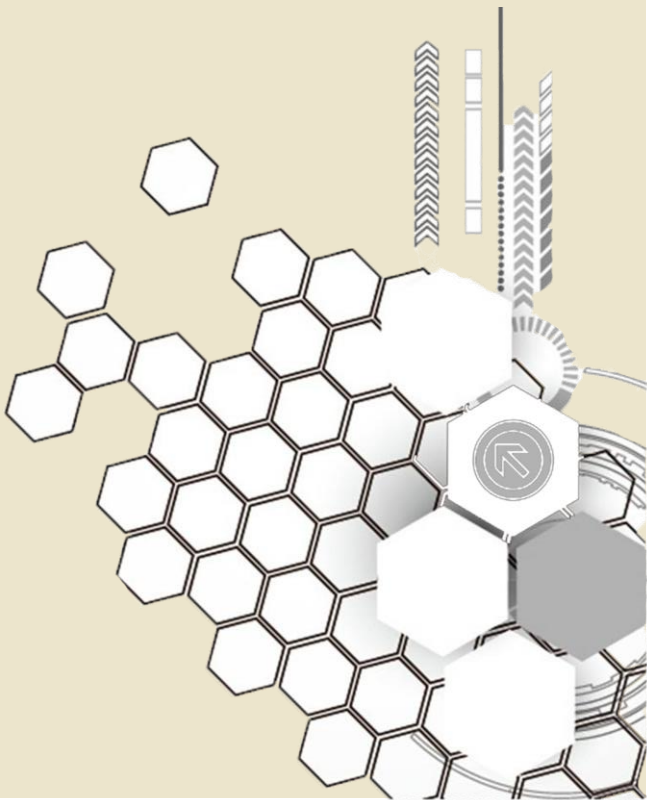
自然语言处理研究情况。依据 2016 年中文信息学会发布的中文信息处理发展报告对自然语言处理研究中的重要技术进行介绍。

自然语言处理领域专家介绍。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。

自然语言处理的应用及趋势预测。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。

1 concept

概述篇



1 概述篇

1.1 自然语言处理概念

自然语言是指汉语、英语、法语等人们日常使用的语言，是自然而然的随着人类社会发展演变而来的语言，而不是人造的语言，它是人类学习生活的重要工具。概括说来，自然语言是指人类社会约定俗成的，区别于人工语言，如程序设计的语言。在整个人类历史上以语言文字形式记载和流传的知识占到知识总量的 80%以上。就计算机应用而言，据统计，用于数学计算的仅占 10%，用于过程控制的不到 5%，其余 85%左右都是用于语言文字的信息处理。

处理包含理解、转化、生成等过程。自然语言处理，是指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。实现人机间的信息交流，是人工智能界、计算机科学和语言学界所共同关注的重要问题。自然语言处理的具体表现形式包括机器翻译、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。可以说，自然语言处理就是要计算机理解自然语言，自然语言处理机制涉及两个流程，包括自然语言理解和自然语言生成。自然语言理解是指计算机能够理解自然语言文本的意义，自然语言生成则是指能以自然语言文本来表达给定的意图。

图 1 自然语言理解层次

自然语言的理解和分析是一个层次化的过程，许多语言学家把这一过程分为五个层次，可以更好地体现语言本身的构成，五个层次分别是语音分析、词法分析、句法分析、语义分析和语用分析。

语音分析是要根据音位规则，从语音流中区分出一个个独立的音素，再根据音位形态规则找出音节及其对应的词素或词。

词法分析是找出词汇的各个词素，从中获得语言学的信息。

句法分析是对句子和短语的结构进行分析，目的是要找出词、短语等的相互关系以及各自在句中的作用。

语义分析是找出词义、结构意义及其结合意义，从而确定语言所表达的真正含义或概念。

语用分析是研究语言所存在的外界环境对语言使用者所产生的影响。

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：

第一， 问答，机器人能正确回答输入文本中的有关问题；

第二， 文摘生成，机器有能力生成输入文本的摘要；

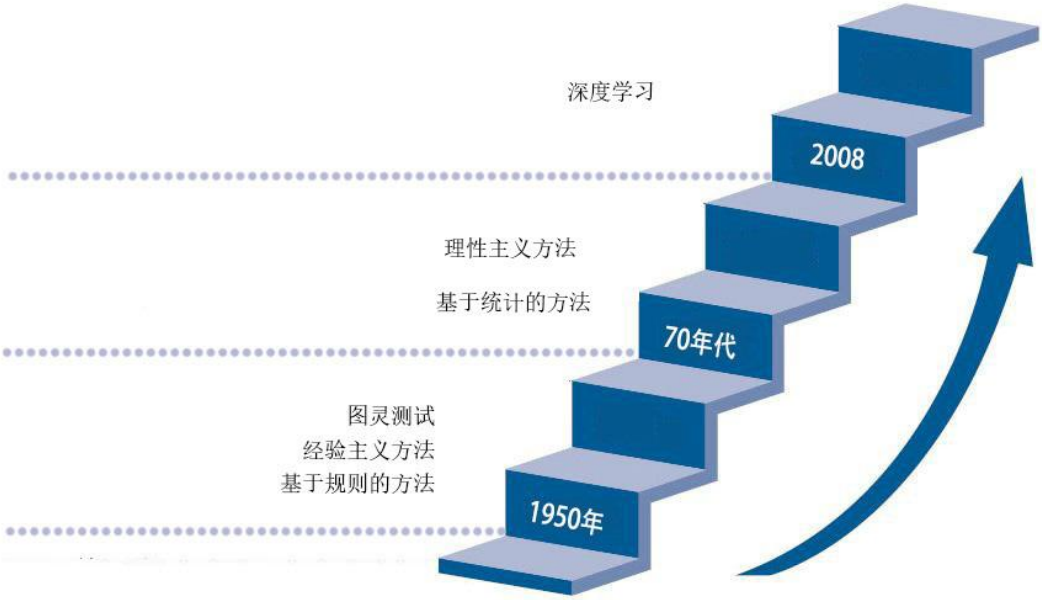
第三， 释义，机器能用不同的词语和句型来复述其输入的文本；

第四， 翻译，机器具有把一种语言翻译成另一种语言的能力。

1.2 自然语言处理发展历程

自然语言处理是包括了计算机科学、语言学心理认知学等一系列学科的一门交叉学科，

这些学科性质不同但又彼此相互交叉。因此，梳理自然语言处理的发展历程对于我们更好地了解自然语言处理这一学科有着重要的意义。



1950 年图灵提出了著名的“图灵测试”，这一般被认为是自然语言处理思想的开端，20 世纪 50 年代到 70 年代自然语言处理主要采用基于规则的方法，研究人员们认为自然语言处理的过程和人类学习认知一门语言的过程是类似的，所以大量的研究员基于这个观点来进行研究，这时的自然语言处理停留在理性主义思潮阶段，以基于规则的方法为代表。但是基于规则的方法具有不可避免的缺点，首先规则不可能覆盖所有语句，其次这种方法对开发者的要求极高，开发者不仅要精通计算机还要精通语言学，因此，这一阶段虽然解决了一些简单的问题，但是无法从根本上将自然语言理解实用化。

70 年代以后随着互联网的高速发展，丰富的语料库成为现实以及硬件不断更新完善，自然语言处理思潮由理性主义向经验主义过渡，基于统计的方法逐渐代替了基于规则的方法。贾里尼克和他领导的 IBM 华生实验室是推动这一转变的关键，他们采用基于统计的方法，将当时的语音识别率从 70%提升到 90%。在这一阶段，自然语言处理基于数学模型和统计的方法取得了实质性的突破，从实验室走向实际应用。

从 2008 年到现在，在图像识别和语音识别领域的成果激励下，人们也逐渐开始引入深度学习来做自然语言处理研究，由最初的词向量到 2013 年 word2vec，将深度学习与自然语言处理的结合推向了高潮，并在机器翻译、问答系统、阅读理解等领域取得了一定成功。深度学习是一个多层的神经网络，从输入层开始经过逐层非线性的变化得到输出。从输入到输出做端到端的训练。把输入到输出对的数据准备好，设计并训练一个神经网络，即可执行预想的任务。RNN 已经是自然语言护理最常用的方法之一，GRU、LSTM 等模型相继引发了一轮又一轮的热潮。

1.3 我国自然语言处理现状

20 世纪 90 年代以来，中国自然语言处理研究进入了高速发展期，一系列系统开始了大规模的商品化进程，自然语言处理在研究内容和应用领域上不断创新。

目前自然语言处理的研究可以分为基础性研究和应用性研究两部分，语音和文本是两类研究的重点。基础性研究主要涉及语言学、数学、计算机学科等领域，相对应的技术有消除歧义、语法形式化等。应用性研究则主要集中在一些应用自然语言处理的领域，例如信息检索、文本分类、机器翻译等。由于我国基础理论即机器翻译的研究起步较早，且基础理论研究是任何应用的理论基础，所以语法、句法、语义分析等基础性研究历来是研究的重点，而且随着互联网网络技术的发展，智能检索类研究近年来也逐渐升温。

从研究周期来看，除语言资源库建设以外，自然语言处理技术的开发周期普遍较短，基本为 1-3 年，由于涉及到自然语言文本的采集、存储、检索、统计等，语言资源库的建设较为困难，搭建周期较长，一般在 10 年左右，例如北京大学计算语言所完成的《现代汉语语法信息词典》以及《人民日报》的标注语料库，都经历了 10 年左右的时间才研制成功。

自然语言处理的快速发展离不开国家的支持，这些支持包括各种扶持政策和资金资助。国家的资金资助包括国家自然科学基金、社会科学基金、863 项目、973 项目等，其中国家自然科学基金是国家投入资金最多、资助项目最多的一项。国家自然科学基金在基础理论研究方面的投入较大，对中文的词汇、巨大、篇章分析方面的研究都给予了资助，同时在技术方面也给予了大力支持，例如机器翻译、信息检索、自动文摘等。除了国家的资金资助外，一些企业也进行了资助，但是企业资助项目一般集中在应用领域，针对性强，往往这些项目开发周期较短，更容易推向市场，实现由理论成果向产品的转化。

1.4 自然语言处理业界发展

- **微软亚洲研究院**

微软亚洲研究院 1998 年成立自然语言计算组，研究内容包括多国语言文本分析、机器翻译、跨语言信息检索和自动问答系统等。这些研究项目研发了一系列实用成果，如 IME、对联游戏、Bing 词典、Bing 翻译器、语音翻译、搜索引擎等，为微软产品做出了重大的贡献，并且在自然语言处理顶级会议，例如 ACL（Association for Computational Linguistics）、COLING（International Conference on Computational Linguistics）等会议上发表了许多论文。

2017 年微软在语音翻译上全面采用了神经网络机器翻译，并新扩展了 Microsoft Translator Live Feature，可以在演讲和开会时，实时同步在手机端和桌面端，同时把讲话者的话翻译成多种语言。其中最重要的技术是对于源语言的编码以及引进的语言知识，微软将句法知识引入到神经网络的编码、解码中，得到了更好的翻译。同时，微软还表示，将来要将知识图谱纳入神经网络机器翻译中规划语言理解的过程中。

在人机对话方面微软也取得了极大的进展，如小娜现在已经拥有超过 1.4 亿用户，在数以十亿计的设备上与人们进行交流，并且覆盖了十几种语言。还有聊天机器人小冰，正在试图把各国语言的知识融合在一起，实现一个开放语言自由聊天的过程，目前小冰实现了中文、日文和英文的覆盖，有上亿用户。

- **Google**

Google 是最早开始研究自然语言处理技术的团队之一，作为一个以搜索为核心的公司，Google 对自然语言处理更为重视。Google 拥有着海量数据，可以搭建丰富庞大的数据库，可以为其研究提供强大的数据支撑。Google 对自然语言处理的研究侧重于应用规模、跨语

言和跨领域的算法，其成果在 Google 的许多方面都被使用，提升了用户在搜索、移动、应用、广告、翻译等方面的体验。

机器翻译方面，2016 年 Google 发布 GNMT 使用最先进的训练技术，能够实现机器翻译质量的最大提升，2017 年宣布其机器翻译实现了完全基于 attention 的 transformer 机器翻译网络架构，实现了新的最佳水平。

Google 的知识图谱更是遥遥领先，例如自动挖掘新知识的准确程度、文本中命名实体的识别、纯文本搜索词条到在知识图谱上的结构化搜索词条的转换等，效果都领先于其他公司，而且很多技术都实现了产品化。

语音识别方面，Google 一直致力于投资语音搜索技术和苹果公司的 siri 竞争，2011 年收购语言信息平台 SayNow，把语音通信、点对点对话、以及群组通话和社交应用融合在一起，2014 年收购了 SR Tech Group 的多项语音识别相关专利，自 2012 年以来将神经网络应用于这一领域，使语音识别错误率极大降低。

- **Facebook**

Facebook 涉猎自然语言处理较晚，Facebook 在 2013 年收购了语音对语音翻译（speech-to-speech translation）研发公司 Mobile Technologies，开始组建语言技术组。该团队很快就投入到其第一个项目——翻译工具——的研发，到 2015 年 12 月，Facebook 用的翻译工具已经完全转变为自主开发。Facebook 语言技术小组不断改进自然语言处理技术以改善用户体验，致力于机器翻译、语音识别和会话理解。2016 年，Facebook 首次将 29 层深度卷积神经网络用于自然语言处理，2017 年，Facebook 团队使用全新的卷积神经网络进行翻译，以往循环神经网络 9 倍的速度实现了当时最高的准确率。

2015 年，Facebook 相继建立语音识别和对话理解工具，开始了语音识别的研发之路。2016 年 Facebook 开发了一个响应“Hey Oculus”的语音识别系统，并且在 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。Facebook 针对文本处理还开发了有效的方法和轻量级工具，这些都基于 2016 年发布的 FastText 即预训练单词向量模型。

- **百度**

百度自然语言处理部是百度最早成立的部门之一，研究涉及深度问答、阅读理解、智能写作、对话系统、机器翻译、语义计算、语言分析、知识挖掘、个性化、反馈学习等。其中，百度自然语言处理在深度问答方向经过多年打磨，积累了问句理解、答案抽取、观点分析与聚合等方面的一整套技术方案，目前已经在搜索、度秘等多个产品中实现应用。篇章理解通

过篇章结构分析、主体分析、内容标签、情感分析等关键技术实现对文本内容的理解，目前，篇章理解的关键技术已经在搜索、资讯流、糯米等产品中实现应用。百度翻译目前支持全球28 种语言，覆盖 756 个翻译方向，支持文本、语音、图像等翻译功能，并提供精准人工翻译服务，满足不同场景下的翻译需求，在多项翻译技术取得重大突破，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。

对百度自然语言处理部做出重要贡献的人物不可不提王海峰、吴华等人。王海峰是百度现任副总裁，负责百度搜索引擎、手机百度、百度信息流、百度新闻、百度手机浏览器、百度翻译、自然语言处理、语音搜索、图像搜索、互联网数据挖掘、知识图谱、小度机器人等业务。是 ACL 50 多年历史上唯一出任过主席（President）的华人，也是迄今为止最年轻的ACL Fellow。同时，王海峰博士还在多个国际学术组织、国际会议、国际期刊兼任各类职务。吴华是百度自然语言处理部技术负责人，她所领导的团队在自然语言处理和机器翻译方面取得重大突破，同时她主持研发的多项 NLP 核心技术应用于搜索、Feed、Duer OS 等百度产品。吴华署名的专利达 40 余件、重要学术论文 50 余篇，在 IJCAI、ACL 等国际会议上多次发声。

- **阿里巴巴**

阿里自然语言处理为其产品服务，在电商平台中构建知识图谱实现智能导购，同时进行全网用户兴趣挖掘，在客服场景中也运用自然语言处理技术打造机器人客服，例如蚂蚁金融智能小宝、淘宝卖家的辅助工具千牛插件等，同时进行语音识别以及后续分析。阿里的机器翻译主要与其国家化电商的规划相联系，可以进行商品信息翻译、广告关键词翻译、买家采购需求以及即时通信翻译等，语种覆盖中文、荷兰语、希伯来语等语种，2017 年初阿里正式上线了自主开发的神经网络翻译系统，进一步提升了其翻译质量。

- **腾讯**

AI Lab 是腾讯的人工智能实验室，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。其研发的腾讯文智自然语言处理基于并行计算、分布式爬虫系统，结合独特的语义分析技术，可满足自然语言处理、转码、抽取、数据抓取等需求，同时，基于文智API 还可以实现搜索、推荐、舆情、挖掘等功能。在机器翻译方面，2017 年腾讯宣布翻译君上线“同声传译”新功能，用户边说边翻的需求得到满足，语音识别+NMT 等技术的应用保证了边说边翻的速度与精准性。

- **京东**

京东在人工智能的浪潮中也不甘落后。京东 AI 开放平台基本上由模型定制化平台和在线服务模块构成，其中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。京东 AI 开放平台计划通过建立算法技术、应用场景、数据链间的连接，构建京东 AI 发展全价值链，实现 AI 能力平台化。

按照京东的规划，NeuHub 平台将作为普惠性开放平台，不同角色均可找到适合自己的场景，例如用简单代码即可实现对图像质量的分析评估。从业务上说，平台可以支撑科研人员、算法工程师不断设计新的 AI 能力以满足用户需求，并深耕电商、供应链、物流、金融、广告等多个领域应用，探索试验医疗、扶贫、政务、养老、教育、文化、体育等多领域应用，

聚焦于新技术和行业趋势研究，孵化行业最新落地项目。同时，京东人工智能研究院与南京大学、斯坦福大学等院校均有合作。

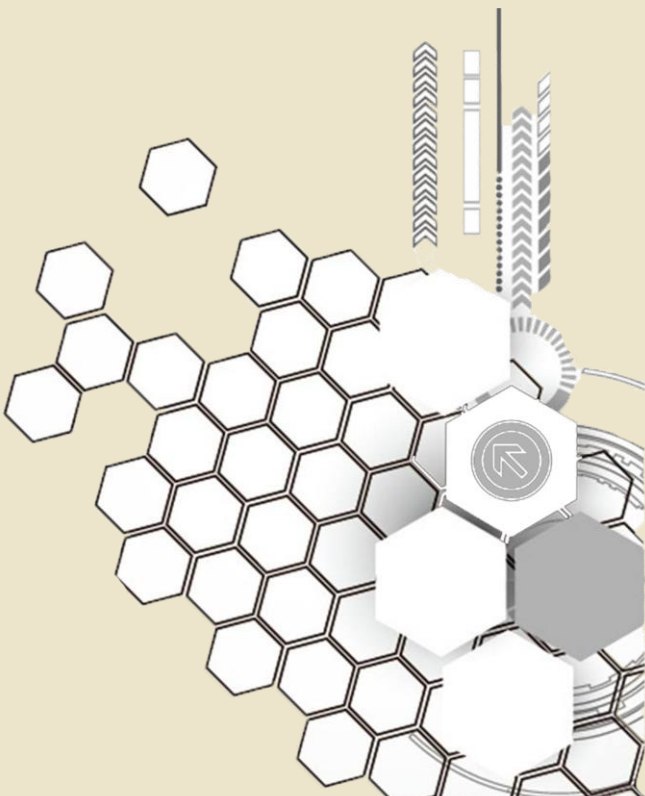
- **科大讯飞**

科大讯飞股份有限公司成立于 1999 年，是一家专业从事智能语音及语言技术、人工智能技术研究、软件及芯片产品开发、语音信息服务及电子政务系统集成的国家级骨干软件企业。科大讯飞作为中国智能语音与人工智能产业领导者，在语音合成、语音识别、口语评测、自然语言处理等多项技术上拥有国际领先的成果。是我国以语音技术为产业化方向的“国家 863 计划成果产业化基地”、“国家规划布局内重点软件企业”、“国家高技术产业化示范工程”，并被原信息产业部确定为中文语音交互技术标准工作组组长单位，牵头制定中文语音技术标准。

科大讯飞成立之时就开始在语言和翻译领域布局项目。基于深度神经网络算法上的创新和突破，科大讯飞在 2014 年国际口语翻译大赛 IWSLT 上获得中英和英中两个翻译方向的全球第一名；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。2017 年科大讯飞还推出了多款硬件翻译产品，其中晓译翻译机 1.0plus 将神经网络翻译系统由在线系统转化为离线系统，实现在没有网络的情况下提供基本的翻译服务。

2 talent

技术篇



2 技术篇

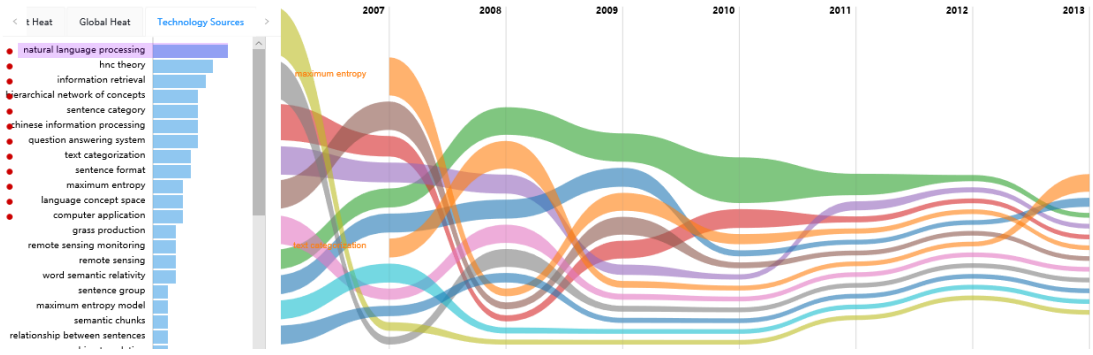


图 2 自然语言处理技术起源

自然语言处理的研究领域极为广泛，各种分类方式层出不穷，各有其合理性，我们按照中国中文信息学会 2016 年发布的《中文信息处理发展报告》，将自然语言处理的研究领域和技术进行以下分类，并选取其中部分进行介绍。

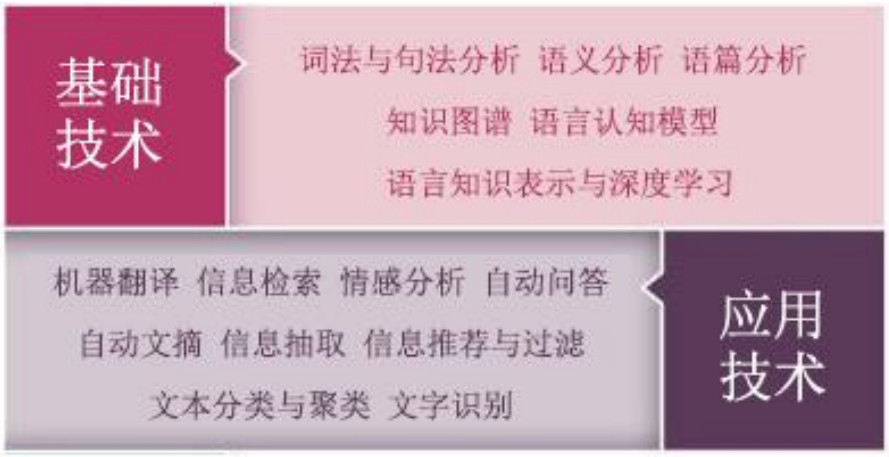


图 3 自然语言处理技术分类

2.1 自然语言处理基础技术

这一小节重点介绍自然语言处理的基础研究方面，自然语言的基础技术包括词汇、短语、句子和篇章级别的表示，以及分词、句法分析和语义分析以及语言认知模型和知识图谱等。

2.1.1 词法、句法及语义分析

词法分析的主要任务是词性标注和词义标注。词性是词汇的基本属性，词性标注就是在给定句子中判断每个词的语法范畴，确定其词性并进行标注。解决兼类词和确定未登录词的词性问题是标注的重点。进行词性标注通常有基于规则和基于统计的两种方法。一个多义词往往可以表达多个意义，但其意义在具体的语境中又是确定的，词义标注的重点就是解决如何确定多义词在具体语境中的义项问题。标注过程中，通常是先确定语境，再明确词义，方法和词性标注类似，有基于规则和基于统计的做法。

判断句子的句法结构和组成句子的各成分，明确它们之间的相互关系是句法分析的主要任务。句法分析通常有完全句法分析和浅层句法分析两种，完全句法分析是通过一系列的句法分析过程最终得到一个句子的完整的句法树。句法分析方法也分为基于规则和基于统计的

方法，基于统计的方法是目前的主流方法，概率上下文无关文法用的较多。完全句法分析存在两个难点，一是词性歧义；二是搜索空间太大，通常是句子中词的个数 n 的指数级。浅层句法分析又叫部分句法分析或语块分析，它只要求识别出句子中某些结构相对简单的成分如动词短语、非递归的名词短语等，这些结构被称为语块。一般来说，浅层语法分析会完成语块的识别和分析、语块之间依存关系的分析两个任务，其中语块的识别和分析是浅层语法分析的主要任务。

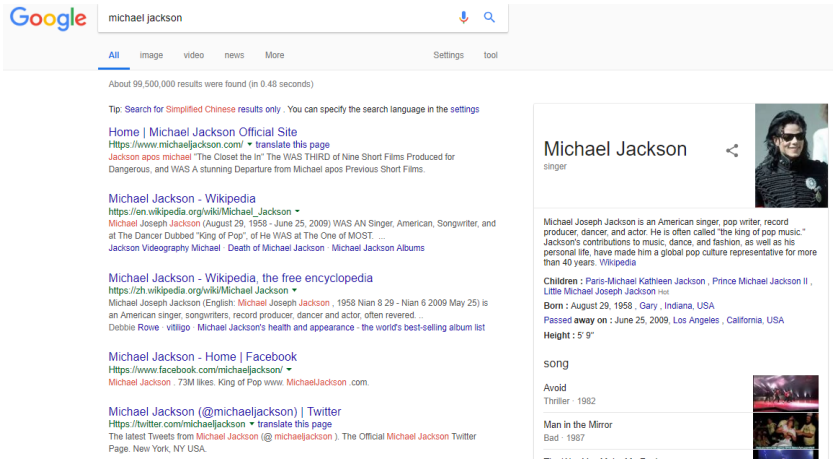
语义分析是指根据句子的句法结构和句子中每个实词的词义推导出来能够反映这个句子意义的某种形式化表示，将人类能够理解的自然语言转化为计算机能够理解的形式语言。句子的分析与处理过程，有的采用“先句法后语义”的方法，但“句法语义一体化”的策略还是占据主流位置。语义分析技术目前还不是十分成熟，运用统计方法获取语义信息的的研究颇受关注，常见的有词义消歧和浅层语义分析。

自然语言处理的基础研究还包括语用语境和篇章分析。语用是指人对语言的具体运用，研究和分析语言使用者的真正用意，它与语境、语言使用者的知识涵养、言语行为、想法和意图是分不开的，是对自然语言的深层理解。情景语境和文化语境是语境分析主要涉及的方面，篇章分析则是将研究扩展到句子的界限之外，对段落和整篇文章进行理解和分析。

除此之外，自然语言的基础研究还涉及词义消歧、指代消解、命名实体识别等方面的研究。

2.1.2 知识图谱

2012 年 5 月，Google 推出 Google 知识图谱，并将其应用在搜索引擎中增强搜索能力，改善用户搜索质量和搜索体验，这是“知识图谱”名称的由来，也标志着大规模知识图谱在互联网语义搜索中的成功应用。搜索关键词，google 会在右侧给出与关键词相关的搜索结果。



知识图谱，是为了表示知识，描述客观世界的概念、实体、事件等之间关系的一种表示形式。这一概念的起源可以追溯至语义网络——提出于 20 世纪五六十年代的一种知识表示形式。语义网络由许多个“节点”和“边”组成，这些“节点”和“边”相互连接，“节点”表示的是概念或对象，“边”表示各个节点之间的关系，如下图。

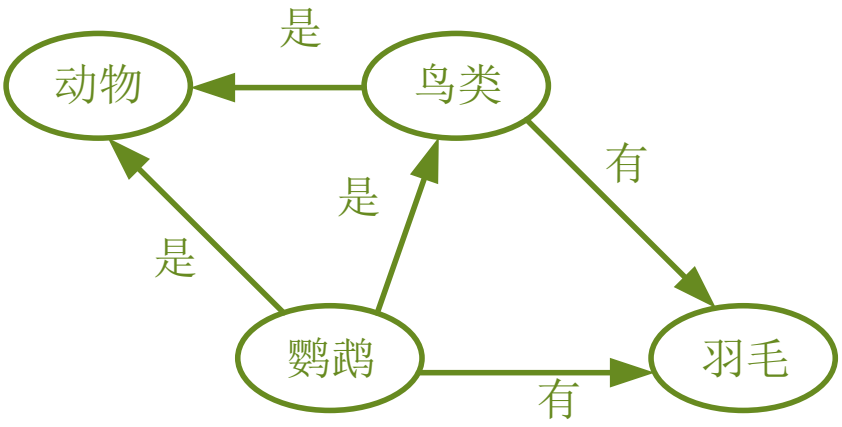


图 4 语义网络示意图

知识图谱在表现形式上与语义网络比较类似，不同的是，语义网络侧重于表示概念与概念之间的关系，而知识图谱更侧重于表述实体之间的关系。现在的知识网络被用来泛指大规模的知识库，知识图谱中包含的节点有以下几种：

实体：指独立存在且具有某种区别性的事物。如一个人、一种动物、一个国家、一种植物等。具体的事物就是实体所代表的内容，实体是知识图谱中的最基本元素，不同的实体间有不同的关系。

语义类：具有同种特性的实体构成的集合，如人类、动物、国家、植物等。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。

内容：通常是实体和语义类的名字、描述、解释等，变现形式一般有文本、图像、音视频等。

属性（值）：主要指对象指定属性的值，不同的属性类型对应于不同类型属性的边。

关系：在知识图谱上，表现形式是一个将节点（实体、语义类、属性值）映射到布尔值的函数。

除语义网络之外，70 年代的专家系统以及 Tim Berners Lee 提出的语义网和关联数据都可以说是知识图谱的前身。

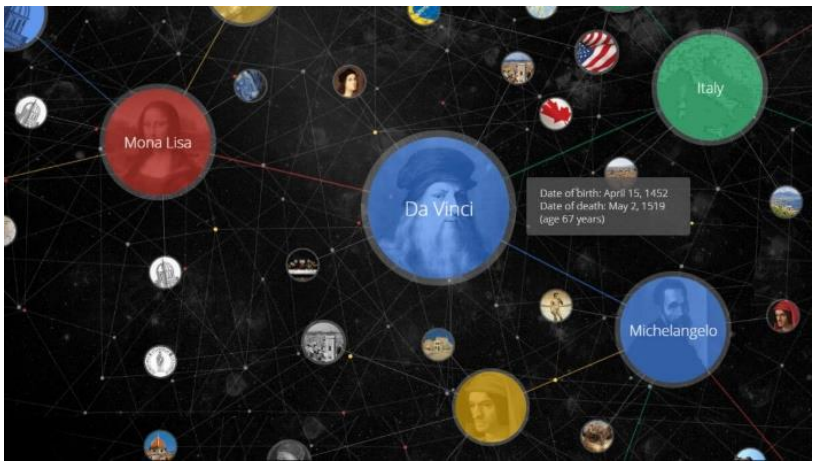


图 5 知识图谱示意图

知识图谱表示、构建和应用涉及很多学科，是一项综合的复杂技术。知识图谱技术既涉及自然语言处理中的各项技术，从浅层的文本向量表示、到句法和语义结构表示被适用于资

源内容的表示中，分词和词性标注、命名实体识别、句法语义结构分析、指代分析等技术被应用于自然语言处理中。同时，知识图谱的研究也促进了自然语言处理技术的研究，基于知识图谱的词义排岐和语义依存关系分析等知识驱动的自然语言处理技术得以建立。

2.2 自然语言处理应用技术

自然语言处理应用技术包括机器翻译、信息检索、情感分析、社会媒体处理等。

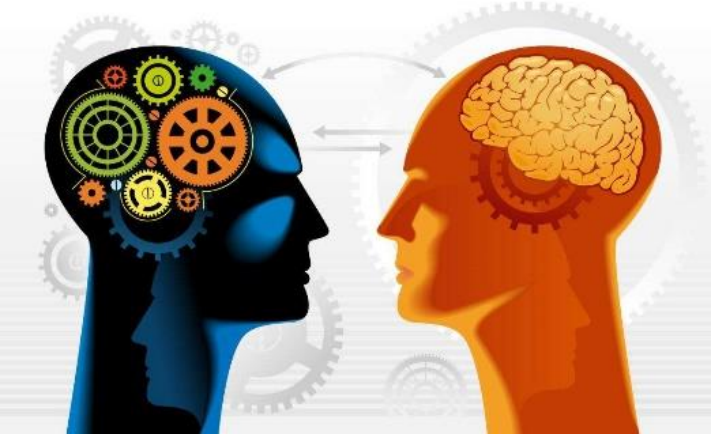
2.2.1 机器翻译

机器翻译（Machine Translation）是指运用机器，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。机器翻译是一门交叉学科（边缘学科），组成它的三门子学科分别是计算机语言学、人工智能和数理逻辑，各自建立在语言学、计算机科学和数学的基础之上。

机器翻译的方法总体上可以分为基于理性的研究方法和基于经验的研究方法两种。

所谓“理性主义”的翻译方法，是指由人类专家通过编撰规则的方式，将不同自然语言之间的转换规律生成算法，计算机通过这种规则进行翻译。这种方法理论上能够把握语言间深层次的转换规律，然而理性主义方法对专家的要求极高，不仅要求其了解源语言和目标语言，还要具备一定的语言学知识和翻译知识，更要熟练掌握计算机的相关操作技能。这些因素都使得研制系统的成本高、周期长，面向小语种的翻译更是人才匮乏非常困难。因此，翻译知识和语言学知识的获取成为基于理性的机器翻译方法所面临的主要问题。

所谓“经验主义”的翻译方法，指的是以数据驱动为基础，主张计算机自动从大规模数据中学习自然语言之间的转换规律。由于互联网文本数据不断增长，计算机运算能力也不断加强，以数据驱动为基础的统计翻译方法逐渐成为机器翻译的主流技术。但是同时统计机器翻译也面临诸如数据稀疏、难以设计特征等问题，而深度学习能够较好的缓解统计机器翻译所面临的挑战，基于深度学习的机器翻译现在正获得迅速发展，成为当前机器翻译领域的热点。



机器翻译技术较早的被广泛应用在计算机辅助翻译软件上，更好地辅助专业翻译人员提升翻译效率，近几年机器翻译研究发展更为迅速，尤其是随着大数据和云计算技术的快速发展，机器翻译已经走进人们的日常生活，在很多特定领域为满足各种社会需求发挥了重要作用。按照媒介可以将机器翻译分为文本翻译、语音翻译、图像翻译以及视频和 VR 翻译等。

目前，文本翻译最为主流的工作方式依然是以传统的统计机器翻译和神经网络翻译为主。Google、Microsoft 与国内的百度、有道等公司都为用户提供了免费的在线多语言翻译系统。将源语言文字输入其软件中，便可迅速翻译出目标语言文字。Google 主要关注以英语为中心的多语言翻译，百度则关注以英语和汉语为中心的多语言翻译。另外，即时通讯工具如 Googletalk、Facebook 等也都提供了即时翻译服务。速度快、成本低是文本翻译的主要特点，而且应用广泛，不同行业都可以采用相应的专业翻译。但是，这一翻译过程是机械的和僵硬的，在翻译过程中会出现很多语义语境上的问题，仍然需要人工翻译来进行补充。

语音翻译可能是目前机器翻译中比较富有创新意思的领域，吸引了众多资金和公众的注意力。亚马逊的 Alexa、苹果的 Siri、微软的 Cortana 等，我们越来越多的通过语音与计算机进行交互。应用比较好的如语音同传技术。同声传译广泛应用于国际会议等多语言交流的场景，但是人工同传受限于记忆、听说速度、费用偏高等因素门槛较高，搜狗推出的机器同传技术主要在会议场景出现，演讲者的语音实时转换成文本，并且进行同步翻译，低延迟显示翻译结果，希望能够取代人工同传，实现不同语言人们低成本的有效交流。科大讯飞、百度等公司在语音翻译方面也有很多探索。如科大讯飞推出的“讯飞语音翻译”系列产品，以及与新疆大学联合研发的世界上首款维汉机器翻译软件，可以准确识别维吾尔语和汉语，实现双语即时互译等功能。

图像翻译也有不小的进展。谷歌、微软、Facebook 和百度均拥有能够让用户搜索或者自动整理没有识别标签照片的技术。图像翻译技术的进步远不局限于社交类应用。医疗创业公司可以利用计算机阅览 X 光照片、MRI（核磁共振成像）和 CT（电脑断层扫描）照片，阅览的速度和准确度都将超过放射科医师。而且更图像翻译技术对于机器人、无人机以及无人驾驶汽车的改进至关重要，福特、特斯拉、Uber、百度和谷歌均已在上路测试无人驾驶汽车的原型。

除此之外还有视频翻译和 VR 翻译也在逐渐应用中，但是目前的应用还不太成熟。

机器翻译这一话题 AMnier 研究报告系列第五期《人工智能之机器翻译研究报告》中有详细阐述，具体内容可查看：https://static.aminer.cn/misc/article/translation.pdf。

2.2.2 信息检索

信息检索是从相关文档集合中查找用户所需信息的过程。先将信息按一定的方式组织和存储起来，然后根据用户的需求从已经存储的文档集合当中找出相关的信息，这是广义的信息检索。信息检索最早提出于 20 世纪 50 年代，90 年代互联网出现以后，其导航工具——搜索引擎可以看成是一种特殊的信息检索系统，二者的区别主要在于语料库集合和用户群体的不同，搜索引擎面临的语料库是规模浩大、内容繁杂、动态变化的互联网，用户群体不再是具有一定知识水平的科技工作者，而是兴趣爱好、知识背景、年龄结构差异很大的网民群体。

信息检索包括“存”与“取”两个方面，对信息进行收集、标引、描述、组织，进行有序的存放是“存”。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的过程是“取”。信息检索的基本原理是将用户输入的检索关键词与数据库中的标引词进行对比，当二者匹配成功时，检索成功。检索标识是为沟通文献标引和检索关键词而编制的人工语言，通过检索标识可以实现“存”“取”的联系一致。检索结果按照与

提问词的关联度输出，供用户选择，用户则采用“关键词查询+选择性浏览”的交互方式获取信息。

以谷歌为代表的“关键词查询+选择性浏览”交互方式，用户用简单的关键词作为查询提交给搜索引擎，搜索引擎并非直接把检索目标页面反馈给用户，而是提供给用户一个可能的检索目标页面列表，用户浏览该列表并从中选择出能够满足其信息需求的页面加以浏览。这种交互方式对于用户来说查询输入是简单的事，但机器却难以通过简单的关键词准确的理解用户的真正查询意图，因此只能将有可能满足用户需求的结果集合以列表的形式提供给用户。

目前互联网是人们获取信息的主要来源，网络上存放着取之不尽、用之不竭的信息，网络信息有着海量、分布、无序、动态、多样、异构、冗余、质杂、需求各异等特点。人们不再满足于当前的搜索引擎带来的查询结果，下一代搜索引擎的发展方向是个性化（精确化）、智能化、商务化、移动化、社区化、垂直化、多媒体化、实时化等。

2.2.3 情感分析

情感分析又称意见挖掘，是指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。情感分析是自然语言理解领域的重要分支，涉及统计学、语言学、心理学、人工智能等领域的理论与方法。情感分析在一些评论机制的 App 中应用较为广泛，比如某酒店网站，会有居住过的客人的评价，通过情感分析可以分析用户评论是积极还是消极的，根据一定的排序规则和显示比例，在评论区显示。这个场景同时也适用于亚马逊、阿里巴巴等电商网站的商品评价。

除此之外，在互联网舆情分析中情感分析起着举足轻重的作用，话语权的下降和网民的大量涌入，使得互联网的声音纷繁复杂，利用情感分析技术获取民众对于某一事件的观点和意见，准确把握舆论发展趋势，并加以合理引导显得极为重要。

同时，在一些选举预测、股票预测等领域情感分析也逐渐体现着越来越重要的作用。

2.2.4 自动问答

自动问答是指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。问答系统是信息服务的一种高级形式，系统反馈给用户的不再是基于关键词匹配排序的文档列表，而是精准的自然语言答案，这和搜索引擎提供给用户模糊的反馈是不同的。在自然语言理解领域，自动问答和机器翻译、复述和文本摘要一起被认为是验证机器是否具备自然理解能力的四个任务。

自动问答系统在回答用户问题时，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。这一过程涉及了包括词法句法语义分析的基础技术，以及信息检索、知识工程、文本生成等多项技术。传统的自动问答基本集中在某些限定专业领域，但是伴随着互联网的发展和大规模知识库语料库的建立，面向开放领域和开放性类型问题的自动问答越来越受到关注。

根据目标数据源的不同，问答技术大致可以分为检索式问答、社区问答以及知识库问答三种。检索式问答和搜索引擎的发展紧密联系，通过检索和匹配回答问题，推理能力较弱。

社区问答是 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，这些社区问答数据覆盖了大量的用户知识和用户需求。检索式问答和社区问答的核心是浅层语义分析和关键词匹配，而知识库问答则正在逐步实现知识的深层逻辑推理。

纵观自动问答发展历程，基于深度学习的端到端的自动问答将是未来的重点关注，同时，多领域、多语言的自动问答，面向问答的深度推理，篇章阅读理解以及对话也会在未来得到更广阔的发展。

2.2.5 自动文摘

自动文摘是运用计算机技术，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的过程。生成的文摘具有压缩性、内容完整性和可读性。

从 1955 年 IBM 公司 Luhn 首次进行自动文摘的实验至今的几十年中，自动文摘经历了基于统计的机械式文摘和基于意义的理解式文摘两种。机械式方法简单容易实现，是目前主要被采用的方法，但是结果不尽如人意。理解式文摘是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。但是随着自然语言处理技术的发展，理解式文摘有着长远的前景，应用于自动文摘的方法也会越来越多。

自动文摘的分类方法多种多样，下表进行简单梳理：

表 1 **自动文摘分类**

分类依据	类别		
摘要功能	指示摘要	信息摘要	评价摘要
与原文档关系	抽取（extraction）	摘要（abstraction）	
对象	单文档摘要	多文档摘要	
基于用户类型	主题摘要	普通摘要	
机器学习角度	有指导的摘要	无指导的摘要	

作为解决当前信息过载的一项辅助手段，自动文摘技术的应用已经不仅仅限于自动文摘系统软件，在信息检索、信息管理等各领域都得到了广泛应用。同时随着深度学习等技术的发展，自动文摘也出现了许多新的研究和领域，例如多文本摘要、多语言摘要、多媒体摘要等。

2.2.6 社会计算

社会计算也称计算社会学，是指在互联网的环境下，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，破解社会难题的学科。社会计算是社会行为与计算系统交互融合，是计算机科学、社会科学、管理科学等多学科交叉所形成的研究领域。它用社会的方法计算社会，既是基于社会的计算，也是面向社会的计算。

社会媒体是社会计算的主要工具和手段，它是一种在线交互媒体，有着广泛的用户参与性，允许用户在线交流、协作、发布、分享、传递信息、组成虚拟的网络社区等等。近年来，

社会媒体呈现多样化的发展趋势，从早期的论坛、博客、维基到风头正劲的社交网站、微博和微信等，正在成为网络技术发展的热点和趋势。社会媒体文本属性特点是其具有草根性，字数少、噪声大、书写随意、实时性强；社会属性特点是其具有社交性，在线、交互。它赋予了每个用户创造并传播内容的能力，实施个性化发布，社会化传播，将用户群体组织成社会化网络，目前典型的社会媒体是 Twitter 和 Facebook，在我国则是微博和微信。社会媒体是一种允许用户广泛参与的新型在线媒体，通过社会媒体用户之间可以在线交流，形成虚拟的网络社区，构成了社会网络。社会网络是一种关系网络，通过个人与群体及其相互之间的关系和交互，发现它们的组织特点、行为方式等特征，进而研究人群的社会结构，以利于他们之间的进一步共享、交流与协作。

社会计算应用广泛，近年来围绕社会安全、经济、工程和军事领域得到了长足发展。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，例如美国圣塔菲研究所建立了首个人工股票市场的社会计算模型。许多发达国家都在政府资助下开展了研究项目，例如美国的 ASPEN，欧盟的 EURACE 等，并且在国家相应的经济政策制定中发挥着越来越重要的作用。通过社交媒体来把握舆情、引导舆论也是社会计算在社会安全方面发挥的一个重要作用。军事方面，许多国家更是加大投入力度扶持军事信息化的发展。

2.2.7 信息抽取

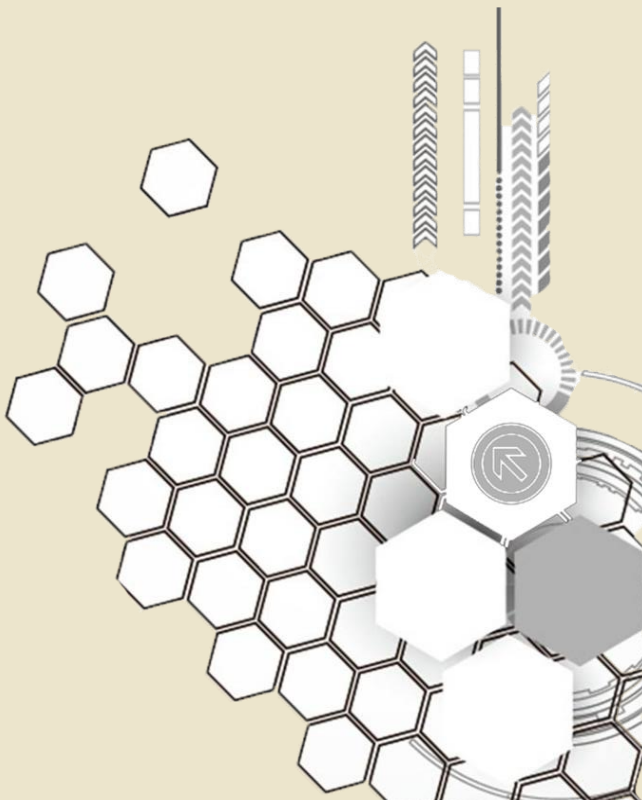
信息抽取技术可以追溯到 20 世纪 60 年代，以美国纽约大学开展的 Linguish String 项目和耶鲁大学 Roger Schank 及其同时开展的有关故事理解的研究为代表。信息抽取主要是指从文本中抽取出特定的事实信息，例如从经济新闻中抽取新发布产品情况，如公司新产品名、发布时间、发布地点、产品情况等，这些被抽取出来的信息通常以结构化的形式直接存入数据库，可以供用户查询及进一步分析使用，为之后构建知识库、智能问答等提供数据支撑。

信息抽取和上文提到的信息检索关系密切，但是二者之间仍存在着很大的不同。首先是二者要实现的功能不同，信息检索是要从大量的文档中找到用户所需要的文档，信息抽取则是用在文本中获取用户感兴趣或所需要的事实信息。其次是二者背后的处理技术也不同，信息检索依靠的主要是以关键字词匹配以及统计等技术，不需要对文本进行理解和分析，而信息则需要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后才能完成信息抽取工作。除了以上的不同之外，信息检索和信息抽取又可以相互补充，信息检索的结果可以作为信息抽取的范围，提高效率，信息抽取用于信息检索可以提高检索质量，更好地满足用户的需求。

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。但我们相信，在信息抽取技术经历了基于规则的方法、基于统计的方法、以及基于文本挖掘的方法等一系列技术演变之后，随着 web、知识图谱、深度学习的发展，可以为信息抽取提供海量数据源、大规模知识资源，更好地机器学习技术，信息抽取技术的问题会得到进一步解决并有长足的发展。

3 application

人才篇



3 人才篇

3.1 国外实验室及人才介绍

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，绘制了该领域顶尖学者全球分布地图。

图 6 自然语言处理人才全球分布图

根据上图，我们可以得出以下结论——从国家来看，美国是自然语言处理研究学者聚集最多的国家，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，而西欧、美国西部等其他先进地区也吸引了大量自然语言处理的研究者。

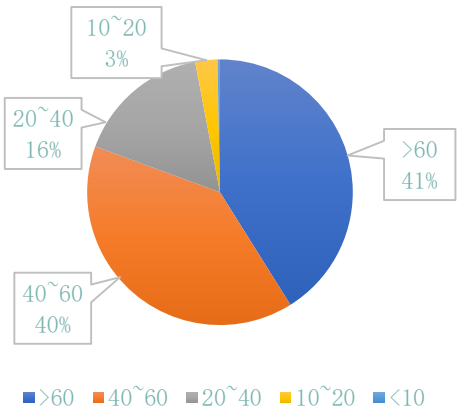


图 7 自然语言处理顶尖学者 h-index 分布

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 41%，h-index 指数在 40 到 60 之间的学者次之，占比 40%。

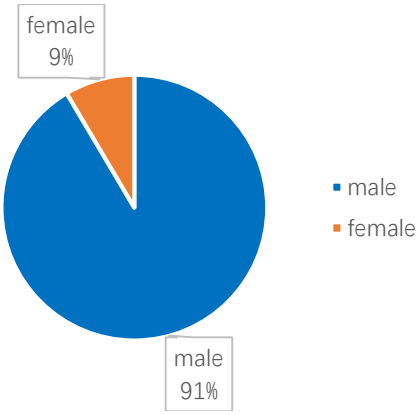


图 8 自然语言处理顶尖学者性别比

自然语言处理领域顶尖学者男性占比 91%，女性占比 9%，男女比例不均衡。

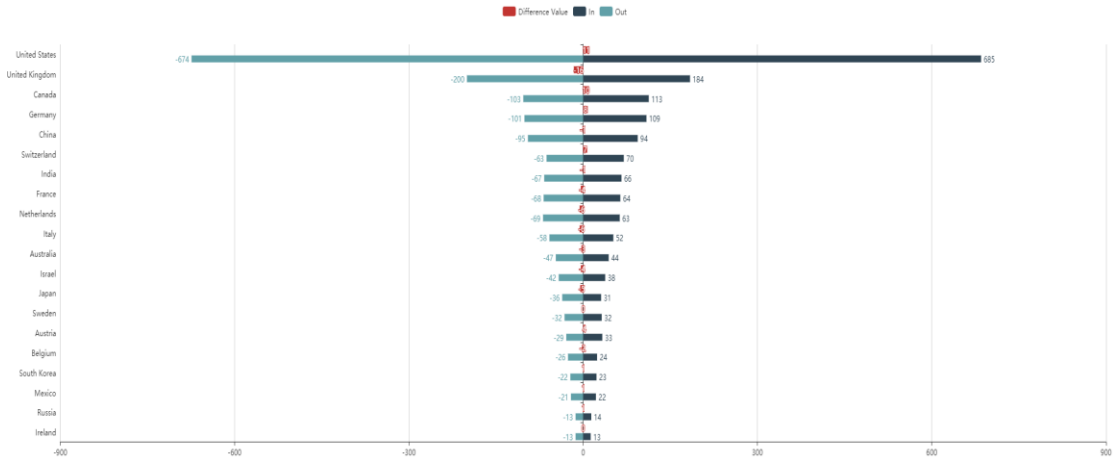


图 9 自然语言处理顶尖人才顺逆差

AMiner 对顶尖人才的迁徙路径做了分析。由上图可以看出，各国自然语言处理顶尖人才的流失和引进是相对比较均衡的，其中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，且从数据来看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。

以下选取在 ACL、EMNLP、NAACL、COLING 等 4 个会议在近 5 年累计发表 10 次以上论文的国外学者及其所在实验室做简要介绍。

- **Chris Dyer**

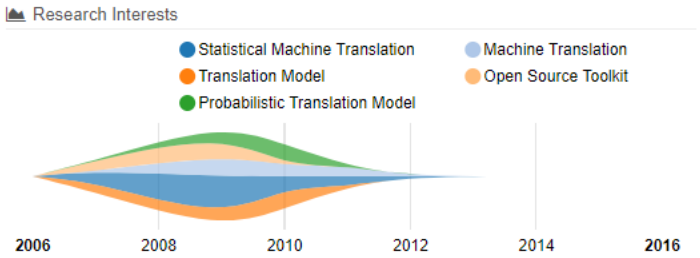
Chris Dyer

H 44 | A 78.45 | S 20.44 | c 11384 | P 77

Assistant Professor

Department of Machine Learning ,Carnegie Mellon University

Statistical Machine Translation | Machine Translation | Translation Model | Open Source Toolkit | Probabilistic Translation Model | Posterior Distribution | Minimum Error Rate Training | Parsing-based Machine Translation



Chris Dyer，卡内基梅隆大学助理教授，2010 年在马里兰大学获语言学博士学位。主要兴趣领域是机器学习、自然语言处理和语言学的交叉研究。比较感兴趣的一些课题有：机器翻译、用于语言处理的神经网络模型、语言建模、特征归纳和表示学习、大数据算法、音乐概率模型等。

2017 年 Chris Dyer 在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems.

作者：Wang Ling、Dani Yogatama、Chris DyerPhil Blunsom

收录会议：ACL

Learning to Create and Reuse Words in Open-Vocabulary Neural Language Modeling.

作者：Kazuya Kawakami、Chris Dyer、Phil Blunsom

收录会议：ACL

Differentiable Scheduled Sampling for Credit Assignment.

作者：Kartik Goyal、Chris Dyer、Taylor Berg-Kirkpatrick

收录会议：ACL

Ontology-Aware Token Embeddings for Prepositional Phrase Attachment.

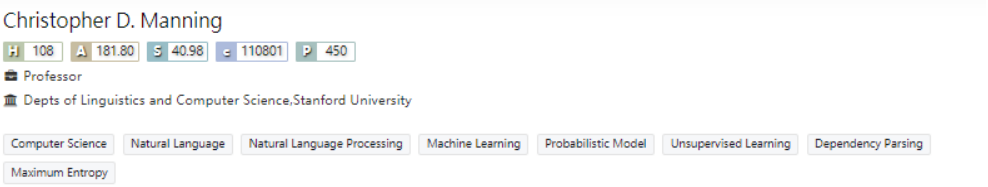
作者：Pradeep Dasigi、Waleed Ammar、Chris Dyer、Eduard H. Hovy

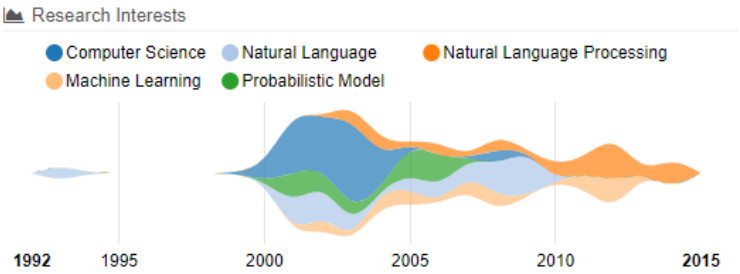
收录会议：ACL

Chris Dyer 所属实验室为 The Language Technologies Institute (LTI) at Carnegie Mellon University

卡内基梅隆大学语言技术研究所主要研究内容包括自然语言处理、计算语言学、信息提取、信息检索、文本挖掘分析、知识表示、机器学习、机器翻译、多通道计算和交互、语音处理、语音界面和对话处理等。

- **Christopher D. Manning**





Christopher D. Manning，斯坦福大学计算机科学与语言学习的教授，1994 年在斯坦福大学获得博士学位。他致力于研究能够智能处理、理解和生成人类语言材料的计算机。Manning 在自然语言处理的深度学习领域有着深入研究，包括递归神经网络、情感分析、神经网络依赖分析等。

Manning 曾获 ACL、CILING、EMNLP 的最佳论文奖。

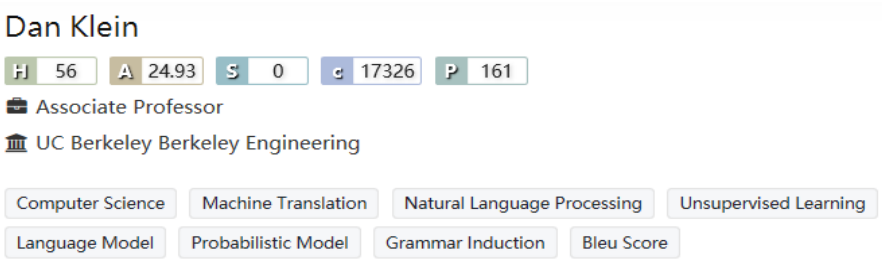
2017 年 Christopher D. Manning 在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

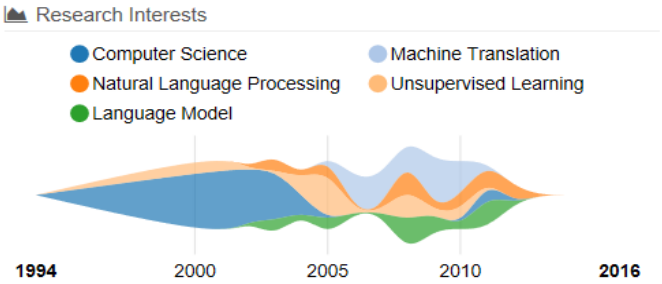
Get To The Point: Summarization with Pointer-Generator Networks.
作者：Abigail See、Peter J.Liu、Christopher D. Manning
收录会议：ACL
Position-aware Attention and Supervised Data Improve Slot Filling.
作者：Yuhao Zhang、Victor Zhong、Danqi Chen、Gabor Angeli、Christopher D. Manning
收录会议：EMNLP
Naturalizing a Programming Language via Interactive Learning.
作者：Sida I. Wang、Samuel Ginn、Percy Liang、Christopher D. Manning
收录会议：ACL
Arc-swift: A Novel Transition System for Dependency Parsing.
作者：Peng Qi、Christopher D. Manning
收录会议：ACL

ChristopherD.Manning 所属研究机构为 The Stanford Natural Language Processing Group

斯坦福大学自然语言处理小组包括了语言学和计算机科学系的成员，是斯坦福人工智能实验室的一部分。主要研究计算机处理和理解人类语言的算法，工作范围从计算语言学的基本研究到语言处理的关键应用技术均有涉猎，涵盖句子理解、自动问答、机器翻译、语法解析和标签、情绪分析和模型的文本和视觉场景等。该小组的一个显著特征是将复杂和深入的语言建模和数据分析与 NLP 的创新概率、机器学习和深度学习方法有效地结合在一起。

- **Dan Klein**





Dan Klein，伯克利大学自然语言处理小组负责人。2004 年在斯坦福大学取得计算机科学的博士学位。主要研究重点是自然语言信息的自组织，兴趣领域包括无监督的语言学习、机器翻译、NLP 的高效算法、信息提取、语言丰富的语言模型、NLP 的符号和统计方法的集成以及历史语言学等。多次在国际顶级会议上发表论文并获奖，如在 2012 年 EMNLP 上获得 Distinguished Paper *"Training Factored PCFGs with Expectation Propagation"*

2017 年 Dan Klein 在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Translating Neuralese.

作者：Jacob Andreas、Anca D. Dragan、Dan Klein

收录会议：ACL

A Minimal Span-Based Neural Constituency Parser.

作者：Mitchell Stern、Jacob Andreas、Dan Klein

收录会议：ACL

Analogs of Linguistic Structure in Deep Representations.

作者：Jacob Andreas、Dan Klein

收录会议：EMNLP

Abstract Syntax Networks for Code Generation and Semantic Parsing.

作者：Maxim Rabinovich、Mitchell Stern、Dan Klein

收录会议：ACL

Fine-Grained Entity Typing with High-Multiplicity Assignments.

作者：Maxim Rabinovich、Dan Klein

收录会议：ACL

Improving Neural Parsing by Disentangling Model Combination and Reranking Effects.

作者：Daniel Fried、Mitchell Stern、Dan Klein

收录会议：ACL

Effective Inference for Generative Neural Parsing.

作者：Mitchell Stern、Daniel Fried、Dan Klein

收录会议：EMNLP

Where is Misty？Interpreting Spatial Descriptors by Modeling Regions in Space.

作者：Nikita Kitaev、Dan Klein

收录会议：EMNLP

Dan Klein 所属实验室为 The Berkeley NLP Group

伯克利大学自然语言处理小组分属于加州大学伯克利分校计算机科学部。主要从事以下几方面的研究工作，语言分析、机器翻译、计算机语言学、基于语义的方法、无监督学习等，多次在顶级国际会议（ACL、EMNLP、AAAI、IJCAI、COLING 等）上发表多篇论文，下表是 2018 年最新被选用的论文。

Constituency Parsing with a Self-Attentive Encoder 作者：Nikita Kitaev、Dan klein

收录会议：ACL 2018
Policy Gradient as a Proxy for Dynamic Oracles in Constiteency Parsing 作者：Daniel Fried、Dan Klein 收录会议：ACL 2018
Learning with Latent Language 作者：Jacob Dan Klein、Sergey Levine 收录会议：NAACL 2018
Unified Pragmatic Models for Generating and Following Instructions 作者：Daniel Fried、Jacob Andreas and Dan Klein 收录会议：NAACL 2018
What's Going On in Neural Constituency Parsers? An Analysis 作者：David Gaddy、Mitchell Stern and Dan Klein 收录会议：NAACL 2018

除了以上提到的，国外还有一些知名自然语言处理实验室，下边做简单介绍。

- **Natural Language Processing Group at University of Notre Dame**

圣母大学自然语言处理小组主要关注机器翻译领域，并有多个项目的研究，如由 DARPA LORELEI 和 Google 赞助的无监督多语言学习模型和算法研究；由亚马逊学术研究奖和谷歌教师研究奖赞助的研究，主要研究课题方向包括基于神经网络的机器翻译模型，以及使用神经网络进行翻译和语言建模的算法等。多次在国际顶级期刊和会议上发表论文，最新的论文有：

Leveraging translations for speech transcription in low-resource settings 作者：Antonis Anastasopoulos、Davis Chiang 收录会议：INTERSPEECH. 2018
Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource 作者：Antonios Anastasopoulos、Justin DeBenedetto、David Chiang. 收录会议：COLING 2018
Composing finite state transducers on GPUs 作者：Arturo Argueta、David Chiang. 收录会议：ACL 2018
Algorithms and training for weighted multiset automata and regular expressions 作者：Justin DeBenedetto、David Chiang. 收录会议：CIAA 2018
Synchronous hyperedge replacement graph grammars 作者：Corey Pennycuff、Satyaki Sikdar、Catalina Vajiac、David Chiang、Tim Weninger. 收录会议 ICGT 2018

目前该小组主要负责人是 David Chiang。

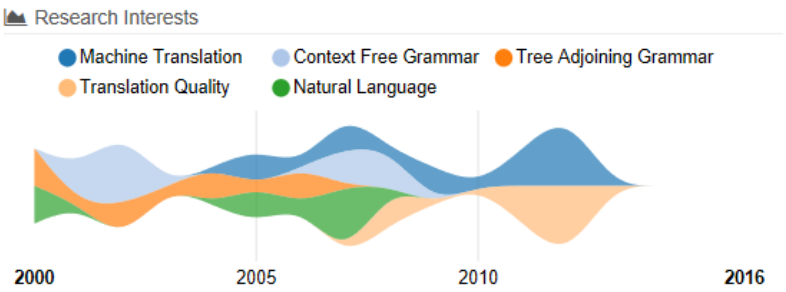
David Chiang (蒋偉)

H 28 | A 9.62 | S 2.73 | c 5692 | P 67

Associate Professor

University of Notre Dame

Machine Translation | Context Free Grammar | Natural Language | Statistical Machine Translation | Translation Quality | Tree Adjoining Grammar | Natural Language Processing | Translation Accuracy



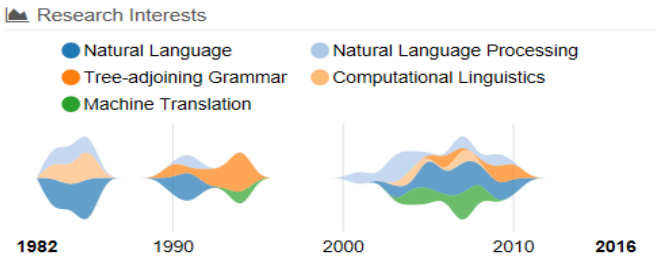
David Chiang，美国圣母大学教授，在宾夕法尼亚大学计算机与信息科学获得博士学位。主要研究领域是自然语言处理，同时在语言翻译、句法分析等方面也有研究。David Chiang 在 2005 年提出的基于短语的翻译模型，对机器翻译来说是一个巨大的进步，他把机器翻译从平面结构建模引向了层次结构建模。

- **The Harvard Natural Language Processing Group**

哈佛自然语言处理小组主要通过机器学习的方法处理人类语言，主要兴趣集中在数列生成的数学模型，以人类语言为基础的人工智能挑战以及用统计工具对语言结构进行探索等方面。该小组的研究出版物和开源项目集中在文本总结、神经机器翻译、反复神经网络的可视化、收缩神经网络的算法、文档中实体跟踪的模型、多模态文本生成、语法错误修正和文本生成的新方法等方面。

Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models 作者：Hendrik Strobelt，Sebastian Gehrmann，Michael Behrisch，Adam Perer，Hanspeter Pfister，Alexander M. Rush
Semi-Amortized Variational Autoencoders 作者：Yoon Kim，Sam Wiseman，Andrew C. Miller，David Sontag，Alexander M. Rush 收录会议：ICML 2018
Compressing Deep Neural Networks with Probabilistic Data Structures 作者：Arturo Argueta，David Chiang. 收录会议：ACL 2018
Algorithms and training for weighted multiset automata and regular expressions 作者：Brandon Reagen，Udit Gupta,Robert Adolf，Michael M. Mitzenmacher，Alexander M. Rush，Gu-Yeon Wei，David Brooks 收录会议：SysML 2018

Stuart Shieber 是该小组的主要负责人。

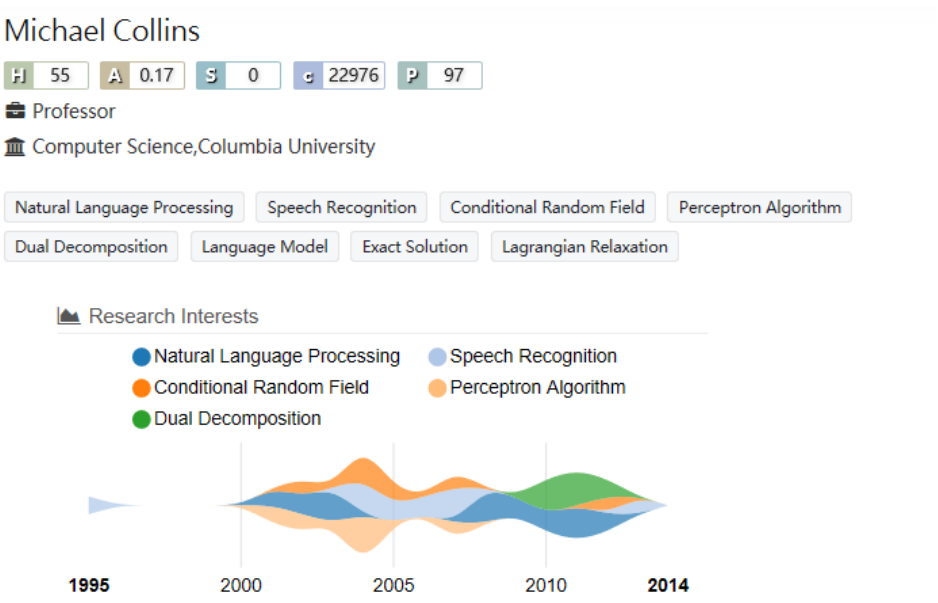


Stuart Shieber，美国计算机协会（Association for Computing Machinery）Fellow 和美国人工智能协会（American Association for Artificial Intelligence）fellow。他综合语言学、理论计算机科学、计算机系统以及人工智能等领域的知识，研究计算机语言学，从计算机科学的角度研究自然语言，在该领域的研究以科学和工程目标，以基础形式和数学工具为基础。具体研究领域包括计算语言学、数学语言学、基于语法的形式、自然语言生成、计算语义、机器翻译以及人机交互等。

- **Natural Language Processing group of Columbia University**

哥伦比亚大学自然语言处理研究室在计算机科学系、计算学习系统中心和生物医学信息系的支持下进行的，将语言洞察力与严谨前沿的机器学习方法和其他计算方法结合起来进行研究。在语言资源创造如语料库、词典等，阿拉伯语 NLP，语言和社交网络，机器翻译，信息提取，数据挖掘，词汇语义、词义消除歧义等方面有着比较深入的研究。

现在该实验室计算机方面的主要负责人为 Michael Collins。



Michael Collins，哥伦比亚大学计算机科学系教授，谷歌 NYC 研究科学家。1998 年在宾夕法尼亚大学获得计算机科学博士学位。主要研究兴趣是自然语言处理和机器翻译。多次在国际顶级会议上发表文章，例如在 EMNLP 2010，CoNLL 2008，UAL 2055 等会议上都获得最佳论文奖，同时还是 ACL 的研究员，获 NSF 生涯奖。

3.2 国内实验室及人才介绍

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。下面基于自然语言处理华人库中的数据对其进行分析。

图 10 AMiner 自然语言处理华人库专家全球分布

图 11 AMiner 自然语言处理华人库专家国内分布

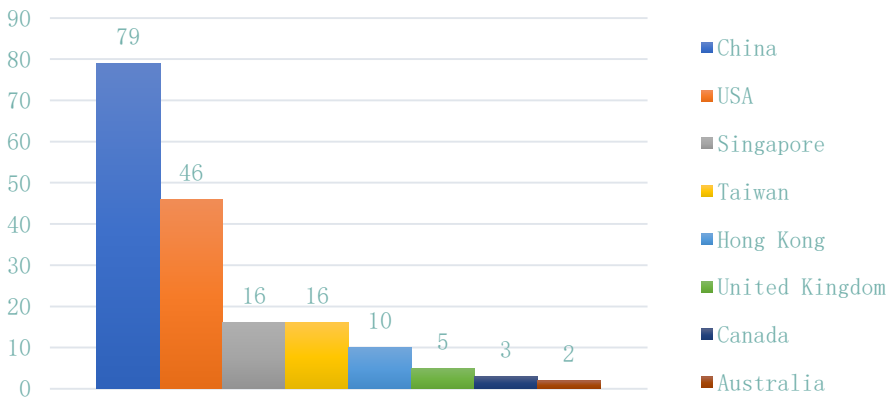


图 12 AMiner 自然语言处理华人库专家地区统计

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。由图 11 可以看出，华人专家在中国流出量大于流入量，美国则正好相反，这也说明就自然领域而言，中国对人才的吸引力要小于美国。

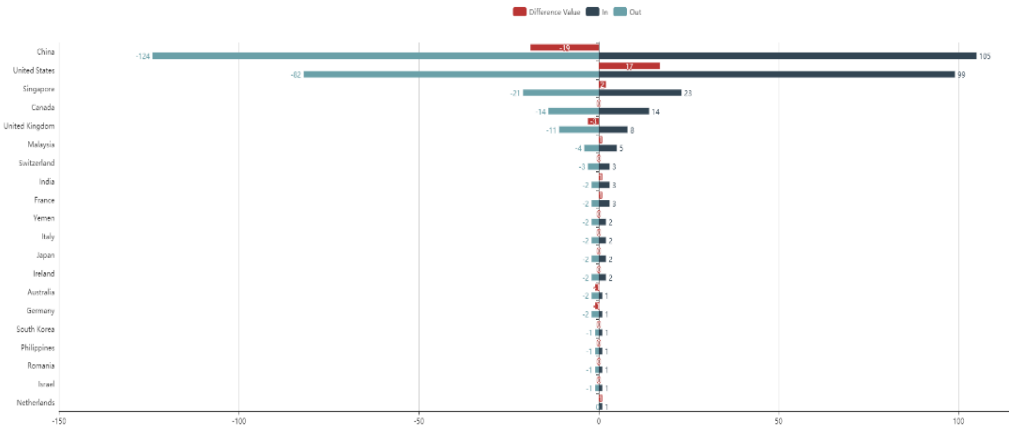


图 13 AMiner 自然语言处理华人库专家流动图

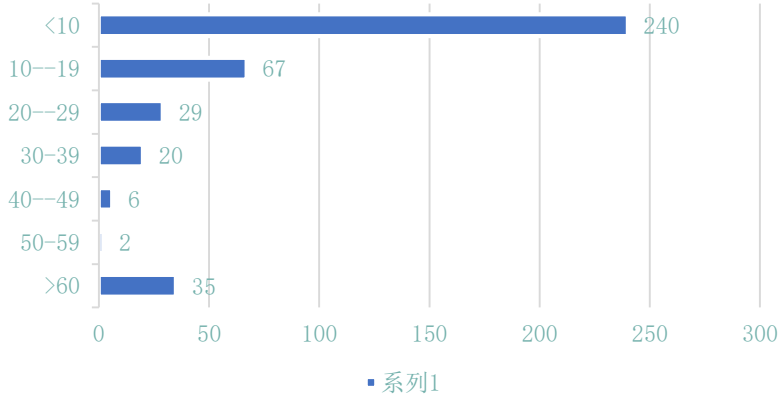


图 14 AMiner 自然语言处理华人库专家 h-index 统计

AMiner 自然语言处理华人库中专家 h-index 指数的平均数为 14，这一数值是远远低于自然语言处理全球 top1000 学者 h-index 指数平均数的。而且，在华人库中，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%。这也说明，自然语言处理的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。

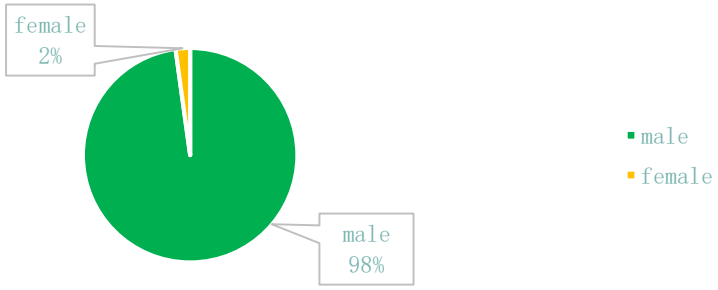


图 15 自然语言处理华人库男女比

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

以下选取在 ACL、EMNLP、NAACL、COLING 等 4 个会议在近 5 年累计发表 10 次以上论文的国内学者包括刘群、刘挺、周明、常宝宝、黄萱菁、刘洋、孙茂松、李素建、万小军、邱锡鹏、穗志方等。以下按照发表论文的多少为序，对这些学者及其所在实验室做简要介绍。

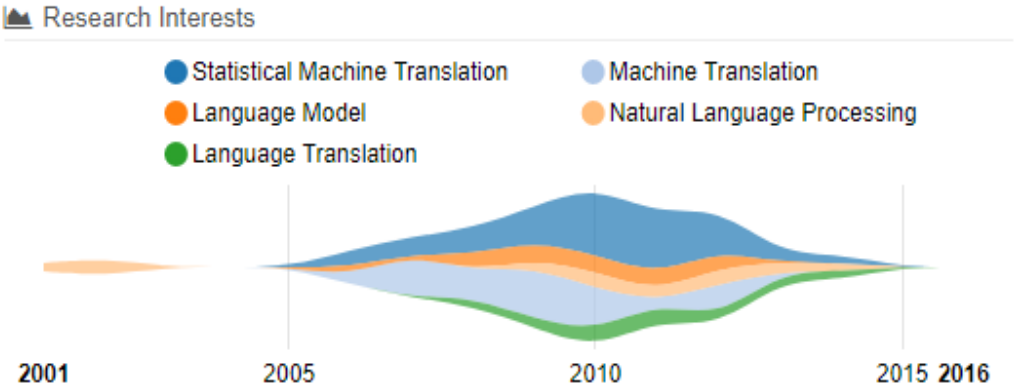
- 刘群

Qun Liu (刘群)

H 33 A 35.49 S 21.13 c 5596 P 194

Institute of Computing Technology,Chinese Academy of Sciences

Statistic Machine Translation | Machine Translation | Language Model | Natural Language Processing | Language Translation | Word Segmentation | Artificial Intelligence | Translation Performance



刘群，中国科学院自然语言处理研究组组长，都柏林大学自然语言处理组组长、项目负责人。主要研究方向是中文自然语言处理，具体包括汉语词法分析、汉语句法分析、语义处理、统计语言模型、辞典和语料库、机器翻译、信息提取、中文信息处理和智能交互中的大规模资源建设、中文信息处理以及智能交互中的评测技术等。曾负责 863 重点项目“机器翻译新方法的研究”和“面向跨语言搜索的机器翻译关键技术研究”等。

2017 年刘群在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Doubly-Attentive Decoder for Multi-modal Neural Machine Translation.

作者：Iacer Calixto、Qun Liu、Nick Campbell

收录会议：ACL

Exploiting Cross-Sentence Context for Neural Machine Translation.

作者：Longyue Wang、Zhaopeng Tu、Andy Way、Qun Liu

收录会议：EMNLP

Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search.

作者：Chris Hokamp、Qun Liu

收录会议：ACL

Deep Neural Machine Translation with Linear Associative Unit.

作者：Mingxuan Wang、Zhengdong Lu、Jie Zhou、Qun Liu

收录会议：ACL

Incorporating Global Visual Features into Attention-Based Neural Machine Translation.

作者：Iacer Calixto、Qun Liu、Nick Campbell

收录会议：EMNLP

Incorporating Word Reordering Knowledge into Attention-based Neural Machine Translation.

作者：Jinchao Zhang、Mingxuan Wang、Qun Liu、Jie Zhou

收录会议：ACL

Further Investigation into Reference Bias in Monolingual Evaluation of Machine Translation.

作者：Qingsong Ma、Yvette Graham、Timothy Baldwin、Qun Liu

收录会议：EMNLP

刘群所属实验室为中科院计算所自然语言处理研究组

自然语言处理研究组隶属于中国科学院计算技术研究所智能信息处理重点实验室。研究组教师有刘群、冯洋等人。研究组主要从事自然语言处理和机器翻译相关的研究工作，研究方向包括机器翻译、人机对话、多语言词法分析、句法分析和网络信息挖掘等。研究组已完成和正在承担的国家自然科学基金、863 计划、科技支撑计划、国际合作等课题 40 余项，在自然语言处理和机器翻译领域取得了多项创新性研究成果。研究组自 2004 年重点开展统计机器翻译方面的研究并取得重大突破，并于 2015 年起转向神经机器翻译并取得很大进展。2018 年 7 月，正式加入华为诺亚方舟实验室，任语音语义首席科学家，主导语音和自然语言处理领域的前沿研究和技术创新。

在自然语言处理的顶级国际刊物 CL、AI 和顶级国际学术会议 ACL、IJCAI、AAAI、EMNLP、COLING 上发表高水平论文 70 余篇，取得发明专利 10 余项。研究组已经成功将自主开发的统计机器翻译和神经机器翻译技术推广到汉语、维吾尔语、藏语、蒙古语、英语、韩语、泰语、日语、阿拉伯语等多种语言。部分语种的翻译系统已经在相关领域得到了实际应用，获得用户的好评。

实验室在 2017 年发表论文见下表。

ME-MD: An Effective Framework for Neural Machine Translation with Multiple Encoders and Decoders 作者：Jinchao Zhang、Qun Liu、Jie Zhou 收录会议：IJCAI 2017
Deep Neural Machine Translation With Linear Associative Unit 作者：Mingxuan Wang、Zhengdong Lu、Jie Zhou、Qun Liu 收录会议：ACL 2017
Incorporating Word Reordering Knowledge into Attention-based Neural Machine Translation 作者：inchao Zhang、Mingxuan Wang、Qun Liu、Jie Zhou 收录会议：ACL 2017

Memory-Augmented Neural Machine Translation 作者：Yang Feng、Shiyue Zhang、Andi Zhang、Dong Wang、Andrew Abel 收录会议：EMNLP 2017
Further Investigation into Reference Bias in Monolingual Evaluation of Machine Translation 作者：Qingsong Ma、Yvette Graham、Timothy Baldwin、Qun Liu 收录会议：EMNLP 2017
Blend: a Novel Combined MT Metric Based on Direct Assessment——CASICT-DCU submission to WMT17 Metrics Task 作者：Qingsong Ma、Yvette Graham、Shugen Wang、Qun Liu 收录会议：WMT 2017
CASICT-DCU Neural Machine Translation Systems for WMT17 作者：Jinchao Zhang、Peerachet Porkaew、Jiawei Hu、Qiuye Zhao、Qun Liu 收录会议：WMT 2017

- 刘挺

刘挺，哈尔滨工业大学教授，国家“万人计划”科技创新领军人才。多次担任国家 863 重点项目总体组专家、基金委会评专家。中国计算机学会理事，中国中文信息学会常务理事/社会媒体处理专委会（SMP）主任，曾任国际顶级会议 ACL、EMNLP 领域主席。

主要研究方向为人工智能、自然语言处理和社会计算，是国家 973 课题、国家自然科学基金重点项目负责人。2012-2017 年在自然语言处理领域顶级会议发表的论文数量列世界第 8 位（据剑桥大学统计），主持研制“语言技术平台 LTP”、“大词林”等科研成果被业界广泛使用。曾获国家科技进步二等奖、省科技进步一等奖、钱伟长中文信息处理科学技术一等奖等。

2017 年刘挺在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Discourse Mode Identification in Essays.

作者：Wei Song、Dong Wang、Ruiji Fu、Lizhen Liu、Ting Liu、Guoping Hu

收录会议：ACL

Transition-Based Disfluency Detection using LSTMs.

作者：Shaolei Wang、Yue Zhang、Wanxiang Che、Meishan Zhang、Ting Liu

收录会议：EMNLP

Chinese Zero Pronoun Resolution with Deep Memory Network.

作者：Qingyu Yin、Yu Zhang、Weinan Zhang、Ting Liu

收录会议：EMNLP

刘挺领导的哈工大社会计算与信息检索研究中心

哈工大社会计算与信息检索研究中心（HIT-SCIR）成立于 2000 年 9 月，隶属于计算机科学与技术学院。研究中心成员有主任刘挺教授，副主任秦兵教授，教师包括张宇、车万翔、陈毅恒、张伟男等。研究方向包括语言分析、信息抽取、情感分析、问答系统、社会媒体处理和用户画像 6 个方面。已完成或正在承担的国家 973 课题、国家自然科学基金重点项目、国家 863 重点项目、国际合作、企业合作等课题 60 余项。在这些项目的支持下打造出“语言技术平台 LTP”，提供给百度、腾讯、华为、金山等企业使用，获 2010 年钱伟长中文信息处理科学技术一等奖。

研究中心近年来发表论文 100 余篇，其中在 ACL、SIGIR、IJCAI、EMNLP 等顶级国际学术会议上发表 20 余篇论文，参加国内外技术评测，并在国际 CoNLL’2009 七国语言句法语义分析评测总成绩第一名。研究中心通过与企业合作，已将多项技术嵌入企业产品中，为社会服务。双语例句检索等一批技术嵌入金山词霸产品中，并因此获得 2012 年黑龙江省技术发明二等奖。

- **周明**

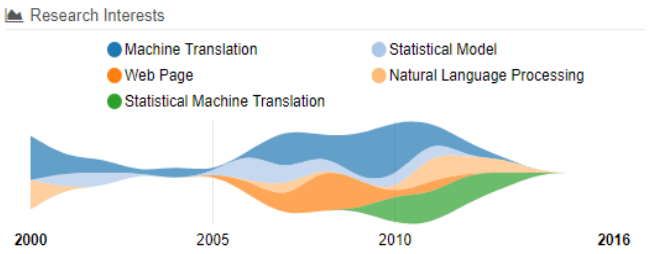
Ming Zhou (周 明)

H 35 | A 53.90 | S 17.07 | C 5406 | P 138

Principal Research Manager

Microsoft Research China

Machine Translation | Web Page | Statistical Model | Statistical Machine Translation | Natural Language Processing | Sentiment Analysis | Semantic Role Labeling | Language Model



周明，微软亚洲研究院自然语言计算组的首席研究员，机器翻译和自然语言处理领域的专家。他的研究兴趣包括搜索引擎、统计和神经机器翻译、问答、聊天机器人、计算机诗歌和文本挖掘等。

1989 年，他设计了“CEMT-I 机器翻译系统”，这是汉英机器翻译的第一个实验，获得了中国大陆政府的科学技术进步奖。1998 年，他设计了著名的中日文机器翻译软件产品 J-Beijing，并获得了日本机械翻译协会 2008 年颁发的机器翻译产品的最高荣誉称号。

周明团队也为 Bing 搜索引擎提供了重要的技术支持，包括单词 breaker、情感分析、speller、解析器和 QnA 等 NLP 技术。他的团队创建了汉英、粤语的机器翻译引擎，为译者

和 Skype 翻译。最近，周明团队与微软产品团队紧密合作，在中国（小冰）、日本（Rinna）和美国（Tay）创建了知名的 chat-bot 产品，拥有 4000 万用户。他在顶级会议（包括 45+ACL 论文）和 NLP 期刊上发表并发表了 100 多篇论文，获得了 38 项国际专利。

2017 年周明在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Gated Self-Matching Networks for Reading Comprehension and Question Answering.

作者：Wenhui Wang、Nan Yang、Furu Wei、Baobao Chang、Ming Zhou

收录会议：EMNLP

Selective Encoding for Abstractive Sentence Summarization.

作者：Qingyu Zhou、Nan Yang、Furu Wei、Ming Zhou

收录会议：ACL

Chunk-based Decoder for Neural Machine Translation.

作者：Shonosuke Ishiwatari、Jingtao Yao、Shujie Liu、Mu Li、Ming Zhou、Naoki Yoshinaga、Masaru、Kitsuregawa、Weijia Jia

收录会议：ACL

Entity Linking for Queries by Searching Wikipedia Sentences.

作者：Chuanqi Tan、Furu Wei、Pengjie Ren、Weifeng Lv、Ming Zhou

收录会议：EMNLP

Stack-based Multi-layer Attention for Transition-based Dependency Parsing.

作者：Zhirui Zhang、Shujie Liu、Mu Li、Ming Zhou、Enhong Chen

收录会议：EMNLP

Sequence-to-Dependency Neural Machine Translation.

作者：Shuangzhi Wu、Dongdong Zhang、Nan Yang、Mu Li、Ming Zhou

收录会议：ACL

Question Generation for Question Answering.

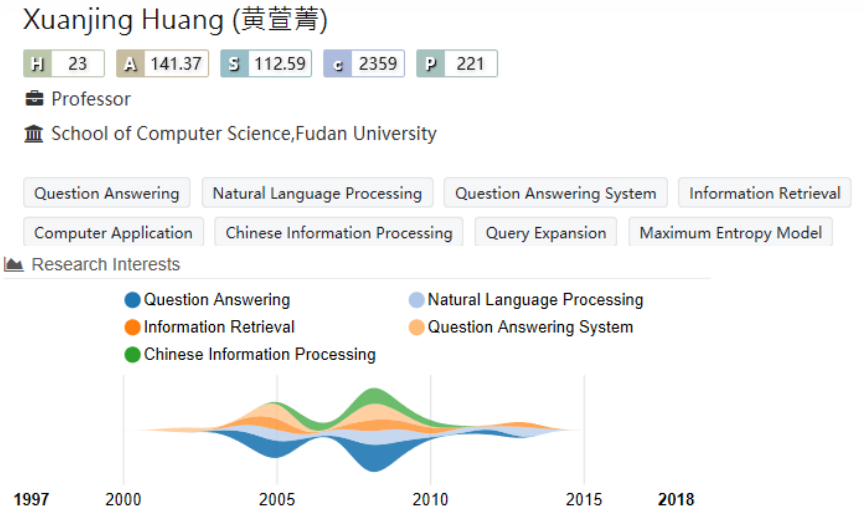
作者：Nan Duan、Duyu Tang、Peng Chen、Ming Zhou

收录会议：EMNLP

周明所属实验室为微软亚洲研究院自然语言计算组。

前文概述篇/自然语言处理业界发展一节中已对其进行介绍，此处不再赘述。

- 黄萱菁



黄萱菁，复旦大学计算机科学技术学院教授、博士生导师。在 SIGIR、ACL、ICML、IJCAI、AAAI、NIPS、CIKM、ISWC、EMNLP、WSDM 和 COLING 等多个国际学术会议上发表论文数十篇。

曾任 2014 年 CIKM 会议竞赛主席，2015 年 WSDM 会议组织者，2015 年全国社会媒体处理大会程序委员会主席，2016 年全国计算语言学会议程序委员会副主席，2017 年自然语言处理与中文计算国际会议程序委员会主席。

多次在人工智能、自然语言处理和信息检索的国际学术会议 IJCAI、ACL、SIGIR、WWW、EMNLP、COLING、CIKM、WSDM 担任程序委员会委员和资深委员。兼任中国中文信息学会常务理事，社会媒体专委会副主任，中国计算机学会中文信息处理专委会委员，中国人工智能学会自然语言理解专委会委员，ACM 和 ACL 会员，《中文信息学报》编委，国家自然科学基金、教育部高校博士点基金和 863 计划同行评议专家。

2017 年黄萱菁在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Adversarial Multi-task Learning for Text Classification.

作者：Pengfei Liu、Xipeng Qiu、Xuanjing Huang

收录会议：ACL

Part-of-Speech Tagging for Twitter with Adversarial Neural Networks.

作者：Tao Gui、Qi Zhang、Haoran Huang、Minlong Peng、Xuanjing Huang

收录会议：EMNLP

Idiom-Aware Compositional Distributed Semantics.

作者：Pengfei Liu、Kaiyu Qian、Xipeng Qiu、Xuanjing Huang

收录会议：EMNLP

黄萱菁领导的复旦大学自然语言处理研究组

复旦大学自然语言与信息检索实验室，致力于社会媒体海量多媒体信息处理的前沿技术研究。主要研究方向包括：自然语言处理、非规范化文本分析、语义计算、信息抽取、倾向性分析、文本挖掘等方面。实验室开发了 NLP 工具包 FudanNLP，FudanNLP 提供了一系列新技术，包括中文分词、词性标注、依赖解析、时间表达式识别和规范化等。

实验室先后承担和参与了国家科技重大专项、国家 973 计划、863 计划、国家自然科学基金课题、上海市科技攻关计划等，并与国内外多所重点大学、公司保持着良好的合作关系。研究成果持续发表在国际权威期刊和一流国际会议（TPAMI、TKDE、ICML、ACL、AAAI、IJCAI、SIGIR、CIKM、EMNLP、COLING 等）。

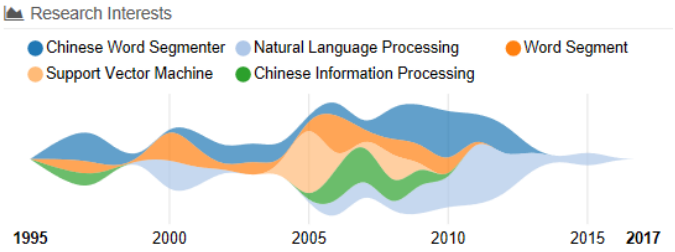
● 孙茂松

Maosong Sun (孙茂松)

H 34 | A 123.30 | S 80.84 | C 4778 | P 248

Department of Computer Science and Technology,Tsinghua University

Chinese Word Segmenter | Natural Language Processing | Word Segment | Machine Learning | Chinese Information Processing | Support Vector Machine | Text Analysis | Natural Languages



孙茂松，清华大学计算机科学与技术系教授。2007-2010 年任该系系主任，主要研究领域为自然语言处理、互联网智能、机器学习、社会计算和计算教育学。国家重点基础研究发展计划（973 计划）项目首席科学家，国家社会科学基金重大项目首席专家。

在国际刊物、国际会议、国内核心刊物上发表论文 160 余篇，主持完成文本信息处理领域 ISO 国际标准 2 项。2007 年获全国语言文字先进工作者，2016 年获全国优秀科技工作者以及首都市民学习之星。多次担任相关领域国际会议和全国性学术会议大会主席或程序委员会主席。

2017 年孙茂松在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

CANE: Context-Aware Network Embedding for Relation Modeling.

作者：Cunchao Tu、Han Liu、Zhiyuan Liu、Maosong Sun

收录会议：ACL

Adversarial Training for Unsupervised Bilingual Lexicon Induction.

作者：Meng Zhang、Yang Liu、Huanbo Luan、Maosong Sun

收录会议：ACL

Visualizing and Understanding Neural Machine Translation.

作者：Yanzhuo Ding、Yang Liu、Huanbo Luan、Maosong Sun

收录会议：ACL

Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction.

作者：Meng Zhang、Yang Liu、Huanbo Luan、Maosong Sun

收录会议：EMNLP

Improved Word Representation Learning with Sememes.

作者：Yilin Niu、Ruobing Xie、Zhiyuan Liu、Maosong Sun

收录会议：ACL

Neural Relation Extraction with Multi-lingual Attention.

作者：Yankai Lin、Zhiyuan Liu、Maosong Sun

收录会议：ACL

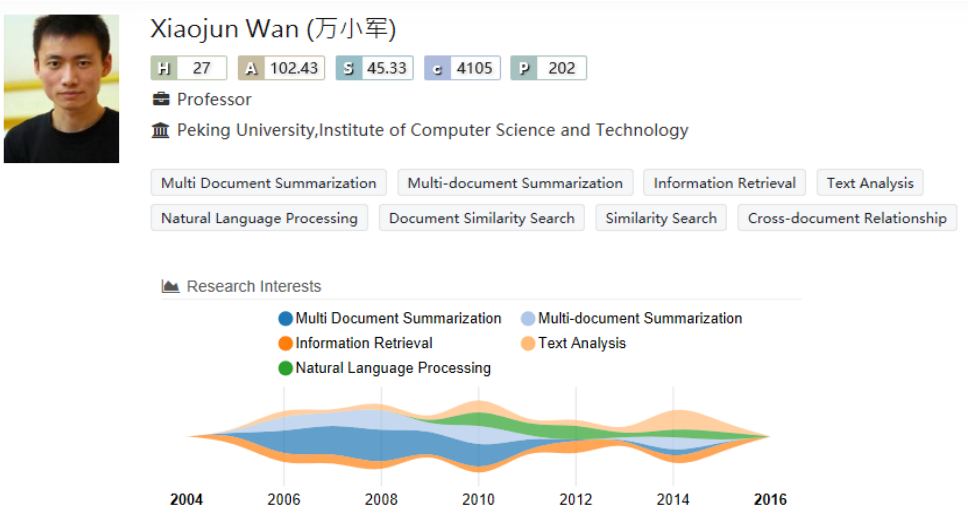
孙茂松领导的清华大学自然语言处理与社会人文计算实验室

清华大学计算机系自然语言处理课题组在 20 世纪 70 年代末，就在黄昌宁教授的带领下从事这方面的研究工作，是国内开展相关研究最早、深具影响力的科研单位，同时也是中国中文信息学会计算语言学专业委员会的挂靠单位。现任学科带头人孙茂松教授任该专业委员会的主任（同时任中国中文信息学会副理事长），其余教师还有刘洋、刘知远等人。目前该课题组对以中文为核心的自然语言处理中的若干前沿课题，进行系统、深入的研究，研究领域的涵盖面正逐步从计算语言学的核心问题扩展到社会计算和人文计算。

该课题组多篇论文被 ACL 2018、IJCAI-ECAI 2018、WWW 2018 录用，内容涉及问答系统、信息检索、机器翻译、诗歌生成、查询推荐等多个领域。具体见下表：

Denoising Distantly Supervised Open-Domain Question Answering 作者：林衍凯、计昊哲、刘知远、孙茂松 收录会议：ACL 2018
Incorporating Chinese Characters of Words for Lexical Sememe Prediction 作者：金晖明*、朱昊*、刘知远、谢若冰、孙茂松、林芬、林乐宇（*同等贡献，本篇文章与腾讯微信合作） 收录会议：ACL 2018
Entity-Duet Neural Ranking:Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval 作者：刘正皓、熊辰炎、孙茂松、刘知远（本篇文章与 CMU 熊辰炎合作） 收录会议：ACL 2018
Towards Robust Neural Machine Translation 作者： 程勇、涂兆鹏、孟凡东、翟俊杰、刘洋（本篇文章与腾讯 AI 实验室合作） 收录会议：ACL 2018
Chinese Poetry Generation with a Working Memory Model 作者：矣晓沅、孙茂松、李若愚、杨宗瀚 收录会议：IJCAI 2018
Query Suggestion with Feedback Memory Network 作者：武彬、熊辰炎、孙茂松、刘知远（本篇文章与 CMU 熊辰炎合作） 收录会议：WWW 2018

- 万小军



万小军，北京大学计算机科学技术研究所教授，博士生导师，语言计算与互联网挖掘实验室负责人。研究方向为自然语言处理与文本挖掘，兴趣领域包括自动文摘与文本生成、情感分析与观点挖掘、语义计算与信息推荐等，在国际重要学术会议与期刊上发表高水平学术论文上百篇。

担任计算语言学顶级国际期刊 Computational Linguistics 编委，TACL 常务评审委员（Standing Reviewing Committee），多次担任自然语言处理领域重要国际会议领域主席或 SPC（包括 ACL、NAACL、IJCAI、IJCNLP 等），以及相关领域多个国际顶级学术会议（ACL、SIGIR、CIKM、EMNLP、NAACL、WWW、AAAI 等）程序委员会委员。研制了自动文摘开源平台 PKUSUMSUM，与今日头条合作推出 AI 写稿机器人小明（Xiaomingbot），与南方

都市报合作推出写稿机器人小南等应用系统。

2017 年万小军在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

Abstractive Document Summarization with a Graph-Based Attentional Neural Model.

作者：Jiwei Tan、Xiaojun Wan、Jianguo Xiao

收录会议：ACL

Parsing to 1-Endpoint-Crossing, Pagenumber-2 Graphs.

作者：Junjie Cao、Sheng Huang、Weiwei Sun、Xiaojun Wan

收录会议：ACL

Semantic Dependency Parsing via Book Embedding.

作者：Weiwei Sun、Junjie Cao、Xiaojun Wan

收录会议：ACL

Quasi-Second-Order Parsing for 1-Endpoint-Crossing, Pagenumber-2 Graphs.

作者：Junjie Cao、Sheng Huang、Weiwei SunXiaojun Wan

收录会议：EMNLP

Towards Automatic Construction of News Overview Articles by News Synthesis.

作者：Jianmin Zhang、Xiaojun Wan

收录会议：EMNLP

Towards a Universal Sentiment Classifier in Multiple languages.

作者：、Kui Xu、Xiaojun Wan

收录会议：EMNLP

万小军所属实验室为北京大学语言计算与互联网挖掘研究组

语言计算与互联网挖掘研究室从属于北京大学计算机科学技术研究所，成立于 2008 年 7 月，负责人为万小军老师。研究室以自然语言处理技术、数据挖掘技术与机器学习技术为基础，对互联网上多源异质的文本大数据进行智能分析与深度挖掘，为互联网搜索、舆情与情报分析、写稿与对话机器人等系统提供关键技术支撑，并从事计算机科学与人文社会科学的交叉科学研究。

研究室当前研究内容包括：1）语义理解：研制全新的语义分析系统实现对人类语言（尤其是汉语）的深层语义理解；2）机器写作：综合利用自动文摘与自然语言生成等技术让机器写出高质量的各类稿件；3）情感计算：针对多语言互联网文本实现高精度情感、立场与幽默分析；4）其他：包括特定情境下的人机对话技术等。

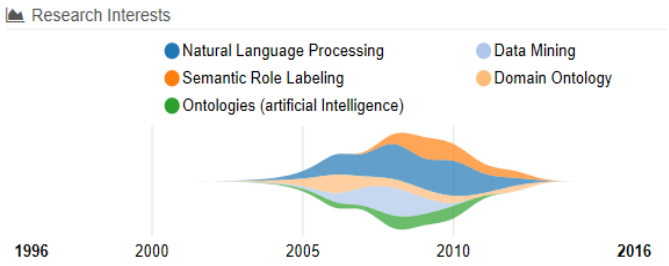
- **穗志方**

Zhifang Sui (穗志方)

H 11 | A 54.69 | S 48.07 | c 365 | P 117

School of Electronics Engineering and Computer Science,Peking University

Natural Language Processing | Data Mining | Semantic Role Labeling | Domain Ontology | Ontologies (artificial Intelligence) | Computational Linguistics | Ontologies | Knowledge Engineering



穗志方，北京大学信息科学技术学院计算语言学实验室主任，教授、博士生导师。2011年度国家科技进步二等奖“综合型语言知识库”项目第二完成人。长期从事自然语言处理方面的研究。

在计算语言学国际顶级会议 ACL 2000、COLING 2008、CONLL 2008、ACL 2009、EMNLP 2009、AIRS 2008 上发表多篇学术论文。作为课题负责人主持的科研项目有：国家自然科学基金项目“汉语动词子语类框架自动获取技术研究”、“基于结构化学习的语义角色标注研究”、“基于 Web 的概念实例及其属性值提取方法研究”，国家社科基金项目“面向文本内容提取的生成性组件库研究及建设”等。

2017 年穗志方在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

A Progressive Learning Approach to Chinese SRL Using Heterogeneous Data.

作者：Qiaolin Xia、Baobao Chang、Zhifang Sui

收录会议：ACL

Affinity-Preserving Random Walk for Multi-Document Summarization.

作者：Zhe Zhao、Tao Liu、Shen Li、Bofang Li、Xiaoyong Du

收录会议：EMNLP

A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction.

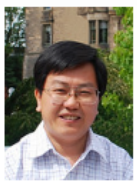
作者：Tianyu Liu、Kexiang Wang、Baobao Chang、Zhifang Sui

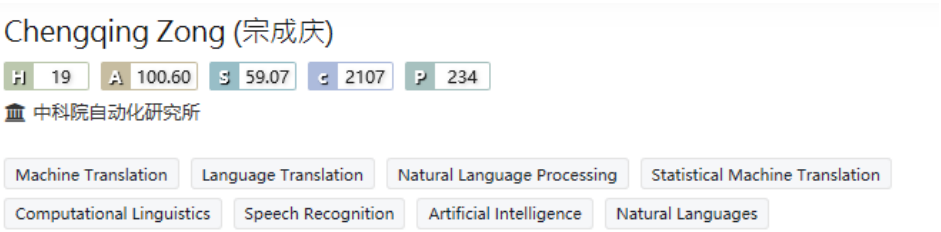
收录会议：EMNLP

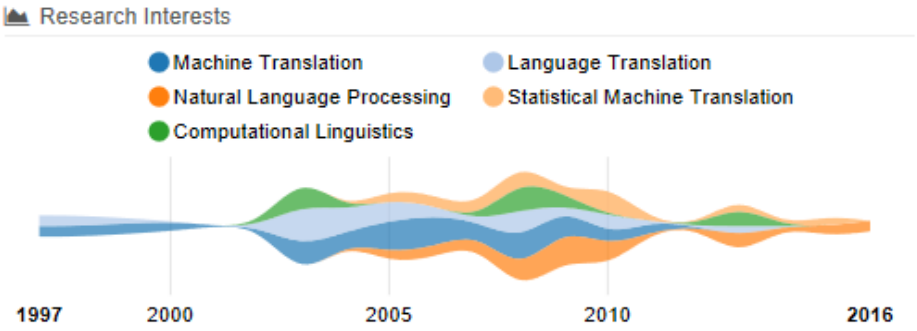
穗志方所属实验室为北京大学计算语言学教育部重点实验室

计算语言学教育部重点实验室依托北京大学建设。实验室研究人员由北京大学信息科学技术学院计算语言学研究所、中文系、软件与微电子学院语言信息工程系、计算机技术研究所、心理系和外语学院的相关研究人员构成。主要研究方向包括：中文计算的基础理论与模型；大规模多层次语言知识库构建的方法；国家语言资源整理与语音数据库建设；海量文本内容分析与动态监控；多语言信息处理和机器翻译。

● 宗成庆







宗成庆，模式识别国家重点实验室研究员、博士生导师。主要从事自然语言处理、机器翻译和文本数据挖掘等相关领域的研究。主持国家自然科学基金项目、863 计划项目和重点研发计划重点专项等 10 余项，发表论文 150 余篇，出版专著和译著各一部。

2013 年当选国际计算语言学委员会（ICCL）委员。目前担任亚洲自然语言处理学会（AFNLP）候任主席、中国中文信息学会副理事长、学术期刊 ACM TALLIP 副主编（Associate Editor）、《自动化学报》副主编、IEEE Intelligent Systems 编委、MachineTranslation 编委和 JCST 编委。2013 年获国务院颁发的政府特殊津贴，2014 年获“钱伟长中文信息处理科学技术奖”一等奖，2015 年获国家科技进步奖二等奖，2017 年获北京市优秀教师荣誉称号。

2017 年宗成庆在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有

Neural System Combination for Machine Translation.

作者：Long Zhou、Wenpeng Hu、Jiajun Zhang、Chengqing Zong

收录会议：ACL

Exploiting Word Internal Structures for Generic Chinese Sentence Representation.

作者：Shaonan Wang、Jiajun Zhang、Chengqing Zong

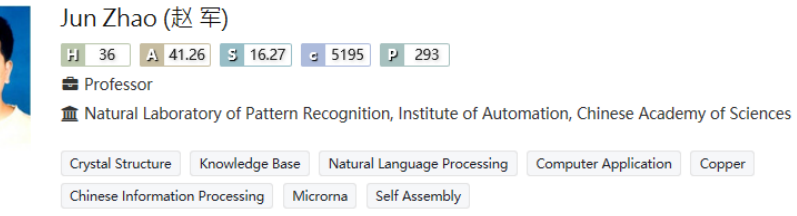
收录会议：EMNLP

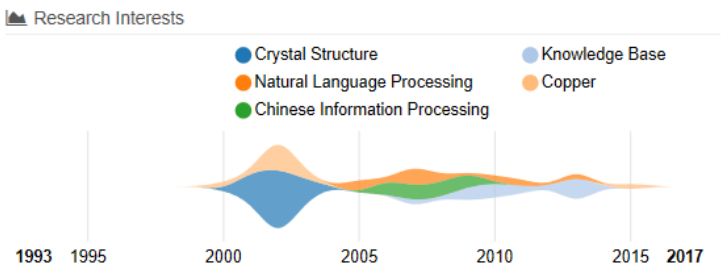
Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video.

作者：Haoran Li、Junnan Zhu、Cong Ma、Jiajun Zhang、Chengqing Zong

收录会议：EMNLP

- 赵军





赵军，中科院研究员，博士生导师。1998 年在清华大学计算机科学与技术系获得博士学位。1998 年-2002 年在香港科技大学计算机科学系做博士后、访问学者。2002 年 5 月至今在中科院自动化所模式识别国家重点实验室工作。

主持国家自然科学基金重点项目、973 计划等国家级项目。研究方向为信息提取和问答系统等。在 IEEE TKDE、JMLR 等顶级国际期刊和 ACL、SIGIR、EMNLP、COLING 等顶级国际会议上发表论文六十余篇，获 COLING-2014 最佳论文奖，获 KDD-CUP2011 亚军（2/1297）。研发了汉语文本分析、信息抽取和知识工程、百科问答等软件工具和平台，在中国大百科全书出版社、华为公司、讯飞公司等得到应用。

2017 年赵军在 ACL、EMNLP、NAACL、COLING 等会议发表的论文有：

An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge.

作者：Yanchao Hao、Yuanzhe Zhang、Kang Liu、Shizhu He、Zhanyi Liu、Hua Wu、Jun Zhao

收录会议：ACL

Generating Natural Answers by Incorporating Copying and Retrieving Mechanisms in Sequence-to-Sequence Learning.

作者：Shizhu He、Cao Liu、Kang Liu、Jun Zhao

收录会议：ACL

Handling Cold-Start Problem in Review Spam Detection by Jointly Embedding Texts and Behaviors.

作者：Xuepeng Wang、Kang Liu、Jun Zhao

收录会议：ACL

Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms.

作者：Shulin Liu、Yubo Chen、Kang Liu、Jun Zhao

收录会议：ACL

Automatically Labeled Data Generation for Large Scale Event Extraction.

作者：Kang Liu、Yubo Chen、Shulin Liu、Xiang Zhang 、Jun Zhao

收录会议：ACL

宗成庆和赵军所属实验室为中科院模式识别国家重点实验室

中科院模式识别国家重点实验室自然语言处理组主要成员有宗成庆、赵军、周玉、刘康、张家俊、汪昆、陆征等。该小组主要从事自然语言处理基础、机器翻译、信息抽取和问答系统等相关研究工作，力图在自然语言处理的理论模型和应用系统开发方面做出创新成果。目前研究组的主要方向包括：自然语言处理基础技术（汉语词语切分、句法分析、语义分析和篇章分析等）、多语言机器翻译、信息抽取（实体识别、实体关系抽取、观点挖掘等）和智能问答系统（基于知识库的问答系统、知识推理、社区问答等）。

近年来，研究组注重于自然语言处理基础理论和应用基础的相关研究，承担了一系列包括国家自然科学基金项目、973 计划课题、863 计划项目和支撑计划项目等在内的基础研究和应用基础研究类项目，以及一批企业应用合作项目。在自然语言处理及相关领域顶级国际期刊（CL、TASLP、TKDE、JMLR、TACL、Information Sciences、Intelligent Systems 等）和学术会议（AAAI、IJCAI、ACL、SIGIR、WWW 等）上发表了一系列论文。2009 年获得第 23 届亚太语言、信息与计算国际会议（PACLIC）最佳论文奖，2012 年获得第一届自然语言处理与中文计算会议（NLPCC）最佳论文奖，2014 年获得第 25 届国际计算语言学大会（COLING）最佳论文奖。获得了 10 余项国家发明专利。

国内学者在国际会议获得 Best paper 的有以下两个：

Pengcheng Yang、Xu Sun、Wei Li、Shuming Ma、Wei Wu、Houfeng Wang 的 *SGM: Sequence Generation Model for Multi-label Classification* 在 2018 COLING 会议中被评为 Best error analysis 和 Best evaluation。

- 王厚峰

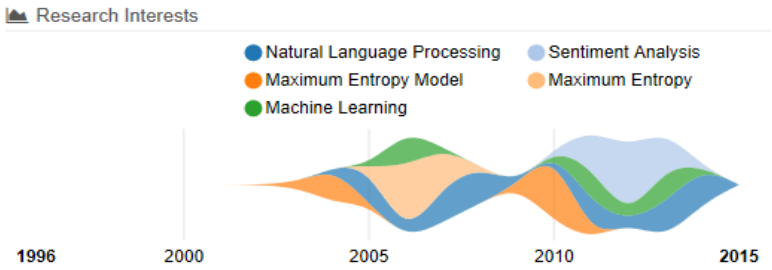
Houfeng Wang (王厚峰)

H 14 | A 24.98 | S 0 | c 725 | P 80

Professor

北京大学信息科学技术学院

Natural Language Processing | Maximum Entropy Model | Machine Learning | Sentiment Analysis | Computational Linguistics | Feature Selection | Semi Supervised Learning | Chinese Abbreviation



王厚峰，北京大学信息科学技术学院教授，北京大学计算语言学研究所所长。主要研究兴趣包括情感分析、问答与会话、自然语言语言语篇分析等，曾作为首席专家主持过国家 863 项目、国家社科基金重大项目，负责国家自然科学基金重大研究计划等。在 ACL、EMNLP、COLING、AAAI、IJCAI、ICML 等会议以及 Computational Linguistics 等期刊发表论文 70 余篇。

Fan Bu、Xiaoyan Zhu、Ming Li 等的 *Measuring the Non-compositionality of Multiword Expressions* 在 2010 年 COLING 会议上被评为 best paper。

- 朱小燕

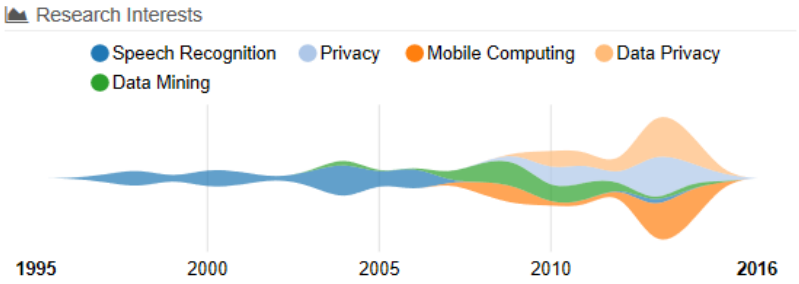
Xiaoyan Zhu (朱小燕)

H 36 | A 136.20 | S 31.09 | c 5859 | P 323

Professor

Department of Computer Science and Technology,Tsinghua University

Speech Recognition | Privacy | Data Privacy | Mobile Computing | Data Mining | Protein–protein Interaction | Ad Hoc Network | Information Retrieval



朱小燕，清华计算机系教授，博士生导师，智能技术与系统国家重点实验室主要负责人。

主要研究领域为智能信息处理，其中包括：模式识别、神经元网络、机器学习、自然语言处理、信息提取和智能问答系统等。近年研究工作主要集中于生物领域文本信息处理和新一代智能信息获取的研究。作为项目负责人先后承担国家 863，973 项目，自然科学基金项目、国际合作项目多项。

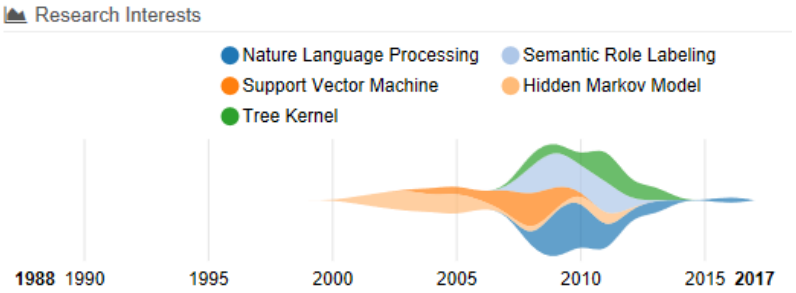
1997 年获国家教委科技进步二等奖，2003 年获北京市科技进步二等奖。获得国家发明专利 3 项。在各种国际刊物和会议上发表论文近 100 篇。其中包括国际刊物 Genome Biology，Bioinformatics、BMC Bioinformatics、Medical informatics、IEEE Transactions.on SMC、IEE Electronics Letters、Neural Parallel & Science Computations、Document Analysis and Recognition，以及国际会议 SIG KDD、ACL、COLING、CIKM 等。

朱小燕所属实验室为清华大学智能技术与系统国家重点实验室

智能技术与系统国家重点实验室依托在清华大学，1987 年 7 月开始筹建。1990 年 2 月通过国家验收，并正式对外开放运行。从 1990 年至 2003 年这十三年间，实验室顺利通过国家自然科学基金委受科技部委托组织的全部三次专家组评估，并被评估为 A（优秀实验室）。1994 年 10 月在庆祝国家重点实验室建设十周年表彰大会上，智能技术与系统国家重点实验室获集体“金牛奖”。1997 年被科技部列为试点实验室。2004 年庆祝国家重点实验室建设二十周年表彰大会上，本实验室再次荣获集体“金牛奖”。从 2004 年开始，实验室参与筹建清华信息科学与技术国家实验室。实验室学术委员会由 17 名国内外著名专家组成。实验室学术委员会名誉主任为中科院院士张钹教授，主任为应明生教授、副主任为邓志东教授。

除此之外，活跃在自然语言领域的中国学者还有：

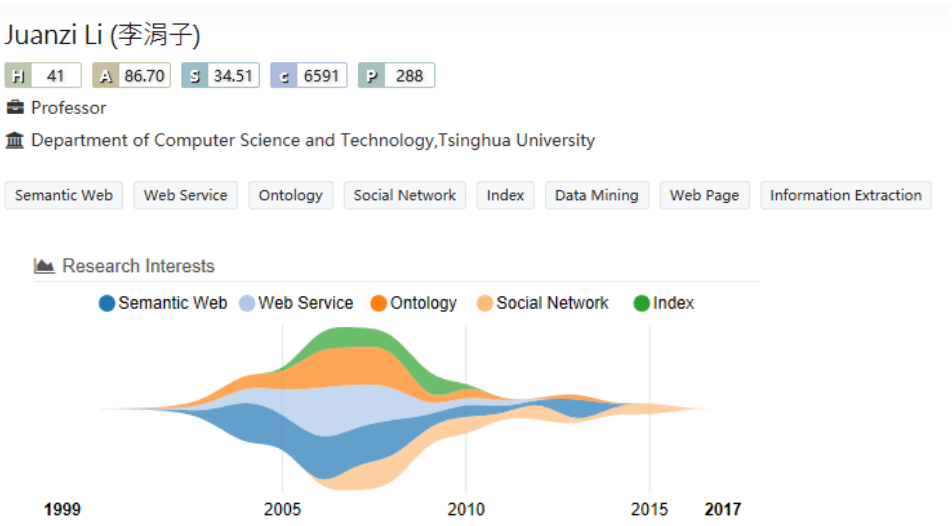
- 周国栋



周国栋，苏州大学计算机科学与技术学院教授，苏州大学 NLP 实验室负责人。主要研究兴趣是自然语言处理、中文计算、信息抽取和自然语言认知等。自 1999 年起，一直是 ACM、ACL、IEEE computer society 的会员，负责了多项国家 863 项目、国家重点研究项目等。

近 5 年来发表国际著名 SCI 期刊论文 20 多篇和国际顶级会议论文 80 多篇，主持 NSFC 项目 4 个(包括重点项目 1 个)。曾担任国际自然语言理解领域顶级 SCI 期刊 Computational Linguistics 编委，目前担任 ACM TALLIP 副主编、《软件学报》责任编委、CCF 中文信息技术专委会副主任委员、苏州大学校学术委员会委员。

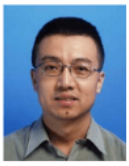
- 李涓子



李涓子，清华大学教授，博士生导师。中国中文信息学会语言与知识计算专委会主任、中国计算机学会术语委员会执行委员。

研究兴趣是语义 Web，新闻挖掘与跨语言知识图谱构建。多篇论文发表在重要国际会议（WWW、IJCAI、SIGIR、SIGKDD）和学术期刊（TKDE、TKDD）。主持多项国家级、部委级和国际合作项目研究，包括国家自然科学项目重点，欧盟第七合作框架、新华社等项目。获得 2013 年人工智能学会科技进步一等奖，2013 年电子学会自然科学二等奖。

- 张民



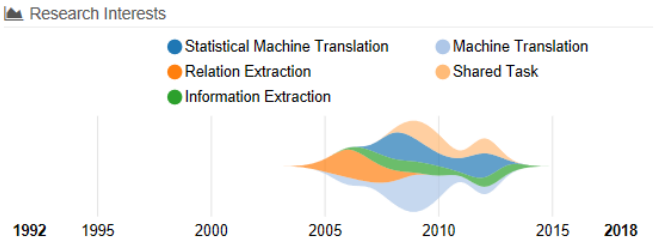
Min Zhang (张 民)

H 31 A 28.05 S 12.68 C 3942 P 137

Professor

苏州大学计算机科学与技术学院

Statistical Machine Translation | Machine Translation | Relation Extraction | Information Extraction | Shared Task | Machine Transliteration | Common Benchmarking Platform | Language Model



张民，苏州大学计算机科学与技术学院副院长。2003 年 12 月，他加入新加坡信息通信研究所并于 2007 年在研究所建立了统计机器翻译团队。2012 年加入苏州大学，并于 2013 年在该大学成立智能计算研究所。

目前的研究兴趣包括机器翻译、自然语言处理、信息提取、社交网络计算、互联网智能、智能计算和机器学习。近年来在国际顶级学报和顶级会议发表学术论文 150 余篇，Springer 出版英文专著两部，主编 Springer 和 IEEE CPS 出版英文书籍十本。他一直积极地为研究界做贡献，组织多会议并在许多会议和讲座中进行演讲。

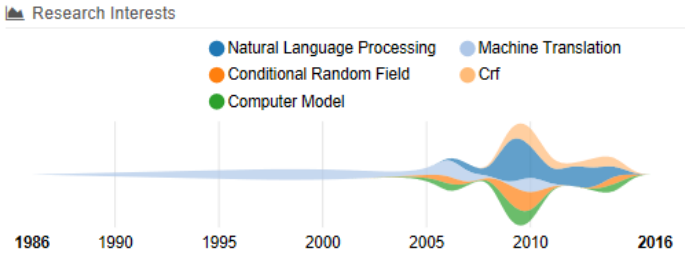
- 黄河燕

Heyan Huang (黄河燕)

H 11 A 20.98 S 5.27 C 353 P 87

北京理工大学计算机学院

Natural Language Processing | Machine Translation | Conditional Random Field | Crf | Computer Model | Hidden Markov Model | Machine Learning | Word Sense Disambiguation



黄河燕，语言智能处理与机器翻译领域专家，北京理工大学计算机学院院长、教授，北京市海量语言信息处理与云计算应用工程技术研究中心主任。长期从事语言智能处理的理论及应用研究，主持多项国家自然科学基金重点项目、国家重点研发计划项目、973 计划课题等重要科研项目，曾获国家科技进步一等奖、二等奖等奖项，被授予“全国优秀科技工作者”称号。

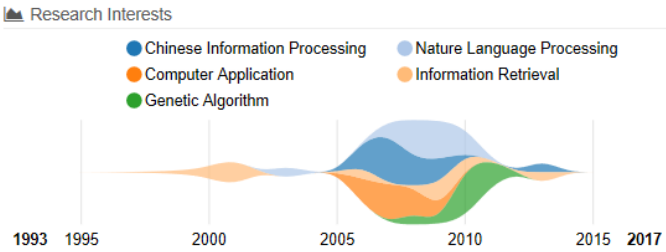
- 孙乐

Le Sun (孙乐)

H 18 | A 4.03 | S 0 | c 1650 | P 198

Institute of Software,Chinese Academy of Sciences, Beijing, China

Chinese Information Processing | Nature Language Processing | Computer Application | Information Retrieval | Genetic Algorithm | Machine Learning | Churn | Language Model



孙乐，中国科学院软件研究所，研究员，博士生导师。中国中文信息学会副理事长兼秘书长。《中文信息学报》副主编。2003 至 2005 年，先后在英国 Birmingham 大学、加拿大 Montreal 大学做访问学者，从事语料库和信息检索研究。

目前主要研究兴趣：基于知识的语言理解、信息抽取、问答系统、信息检索等。在国内外主要刊物和会议上共发表论文 80 多篇。曾任 2008 和 2009 国际测评 NTCIR MOAT 中文简体任务的组织者、国际计算语言学大会（COLING 2010）组织委员会联席主席、机器翻译峰会（MT Summit 2011）组织委员会联席主席、中文语言评测国际会议（CLP2010、2012、2014）大会主席、国际计算语言学年会（ACL 2015）组织委员会联席主席等。

3.3 ACL2018 奖项介绍

2018 年 7 月 15 在墨尔本开幕的 ACL 公布了其最佳论文名单，包括 3 篇最佳长论文和 2 篇最佳短论文以及 1 篇最佳 demo 论文，值得一提的是 Amazon Door Prize 中北京大学和哈尔滨大学上榜，ACL2018 终身成就奖为爱丁堡大学 Mark Steedman 获得。

- 最佳长论文：

Finding syntax in human encephalography with beam search

用波束搜索在人脑成像中寻找句法

作者：John Hale，Chris Dyer，Adhiguna Kuncoro，Jonathan R.Brennan

论文摘要：循环神经网络文法（RNNGs）是对于树-字符串对的生成式模型，它们依靠神经网络来评价派生的选择。用束搜索对它们进行解析可以得到各种不同复杂度的评价指标，比如单词惊异数（word surprisal count）和解析器动作数（parser action count）。当把它们用作回归因子，解析人类大脑成像图像中对于自然语言文本的电生理学响应时，它们可以带来两个增幅效果：一个早期的峰值以及一个类似 P600 的稍迟的峰值。相比之下，一个不具有句法结构的神经语言模型无法达到任何可靠的增幅效果。通过对不同模型的对比，早期峰值的出现可以归功于 RNNG 中的句法组合。结果中体现出的这种模式表明 RNNG+束搜索的组合可以作为正常人类语言处理中的语法处理的一个不错的机理解释模型。

论文地址：https://arxiv.org/abs/1806.04127

Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information

学习如何问好的问题：通过完全信息下的期待值为追问问题排序

作者：Sudha Rao，Hal Daumé III

论文摘要：在沟通中，提问是一大基本要素：如果机器不知道如何问问题，那它们也就无法高效地与人类合作。在这项研究中，作者们构建了一个神经网络用于给追问的问题做排名。作者们模型设计的启发来源于完全信息情况下的期待值：一个可以期待获得有用的答案的问题就是一个好问题。作者们根据 StackExchange 上抓取的数据研究了这个问题；StackExchange 是一个内容丰富的在线咨询平台，其中有人发帖咨询以后，别的用户会在下面追问起到解释澄清作用的问题，以便更好地了解状况、帮助到发帖人。论文作者们创建了一个由这样的追问问题组成的数据集，其中包含了 StackExchange 上 askubuntu、unix、superuser 这三个领域的约 77k 组发帖+追问问题+问题的回答。作者们在其中的 500 组样本上评估了自己的模型，相比其他基准模型有显著的提高；同时他们也与人类专家的判断进行了对比。

论文地址：https://arxiv.org/abs/1805.04655

Let's do it "again": A First Computational Approach to Detecting Adverbial Presupposition Triggers

让我们「再」做一次：首个检测假定状态触发副词的计算性方法

作者：Andre Cianflone，Yulan Feng，Jad Kabbara，Jackie Chi Kit Cheung

论文摘要：这篇论文中，作者们介绍了一种新的研究课题——预测副词词性的假定状态触发语（adverbial presupposition triggers），比如 also 和 again。完成这样的任务需要在对话上下文里寻找重复出现的或者相似的内容；这项任务的研究成果则可以在文本总结或者对话系统这样的自然语言生成任务中起到帮助。作者们为这项任务创造了两个新的数据集，分别由 Penn Treebank 和 AnnotatedEnglish Gigaword 生成，而且也专为这项任务设计了一种新的注意力机制。作者们设计的注意力机制无需额外的可训练网络参数就可以增强基准 RNN 模型的表现，这最小化了这一注意力机制带来的额外计算开销。作者们在文中表明，他们的模型相比多个基准模型都有统计显著的更高表现，其中包括基于 LSTM 的语言模型。

论文地址：https://www.cs.mcgill.ca/~jkabba/acl2018paper.pdf

- 最佳短论文

Know What You Don't Know: Unanswerable Questions for SQuAD

知道你不知道的：SQuAD 中无法回答的问题

作者：Pranav Rajpurkar，Robin Jia，Percy Liang

论文摘要：提取式的阅读理解系统一般都能够在给定的文档内容中找到正确的内容来回答问题。不过对于正确答案没有明示在阅读文本中的问题，它们就经常会做出不可靠的猜测。目前现有的阅读理解问答数据集，要么只关注了可回答的问题，要么使用自动生成的无法回答的问题，很容易识别出来。为了改善这些问题，作者们提出了 SQuAD2.0 数据集，这是斯坦福问答数据集 SQuAD 的最新版本。SQuAD2.0 在现有的十万个问题-答案对的基础上增加了超过五万个无法回答的问题，它们由人类众包者对抗性地生成，看起来很像可以回答的问题。一个问答系统如果想要在 SQuAD2.0 上获得好的表现，它不仅需要在问题能够回答时给出正确的答案，还要在给定的阅读材料中不包含答案时做出决定、拒绝回答这个问题。SQuAD2.0 也设立了新的人类表现基准线，EM86.831，F189.452。对于现有模型来说 SQuAD2.0 是一个具有挑战性的自然语言理解任务，一个强有力的基于神经网络的系统可以在 SQuAD1.1 上得到 86%的 F1 分数，但在 SQuAD2.0 上只能得到 66%。

论文地址：https://arxiv.org/abs/1806.03822

'Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions

“更浅的颜色”也可能仍然是黑暗的：建模比较性的颜色描述

作者：Olivia Winn，Smaranda Muresan

论文摘要：我们提出了一种在颜色描述领域内建立基准比较性形容词的新范式。给定一个参考 RGB 色和一个比较项（例如更亮、更暗），我们的模型会学习建立比较项的基准，将其作为 RGB 空间中的一个方向，这样颜色就会沿着向量植根于比较色中。我们的模型产生了比较形容词的基本表示形式，在期望的改变方向上，平均精确度为 0.65 余弦相似性。与目标颜色相比，依据向量的颜色描述方法 Delta-E 值小于 7，这表明这种方法与人类感知的差异非常小。这一方法使用了一个新创建的数据集，该数据集来自现有的标记好的颜色数据。

论文地址：http://aclweb.org/anthology/P18-2125

- **最佳 demo 论文**

Out-of-the-box Universal Romanization Tool

开箱即用的通用罗马化工具

作者：Ulf Hermjakob，Jonathan May，Kevin Knight

论文摘要：我们想介绍 uroman，这个工具可以把五花八门的语言和文字（如中文、阿拉伯语、西里尔文）转换为普通拉丁文。该工具基于 Unicode 数据以及其他表，可以处理几乎所有的字符集（包括一些晦涩难懂的语言比如藏文和提非纳文）。uroman 还可以将不同文本中的数字转换为阿拉伯数字。罗马化让比较不同文本的字符串相似性变得更加容易，因为

不再需要将两种文字翻译成中间文字再比较。本工具作为一个 Perl 脚本，可以免费提供，可用于数据处理管道和交互式演示网页。

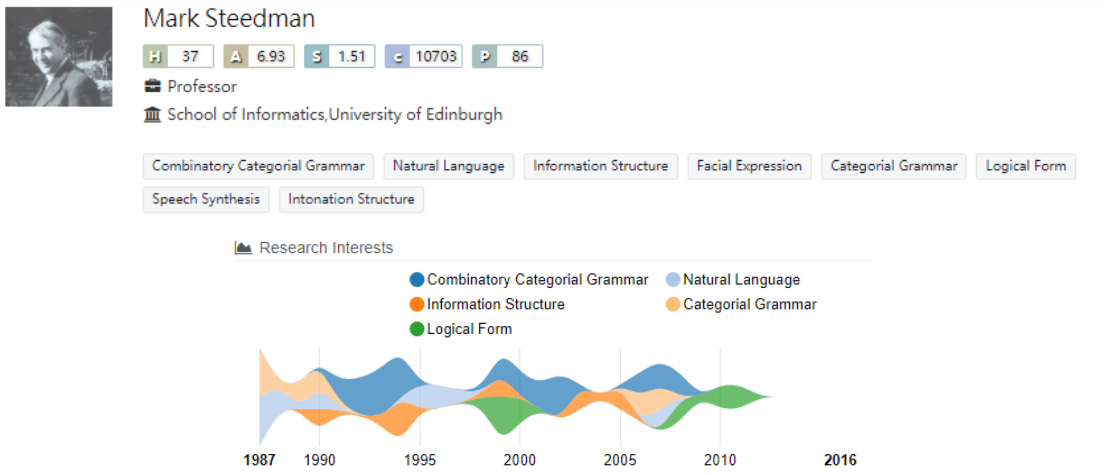
论文地址：http://aclweb.org/anthology/P18-4003

Amazon Door Prize

Wanxiang Che	哈尔滨工业大学	Wei Xue	佛罗里达国际大学
Prachi Manchanda	德里理工学院	Sam Wei	悉尼大学
Fuli Luo	北京大学	Samir Kumar	M12
Nikhilesh Bhatnagar	海得拉巴国际信息技术研究所	Jin-ge Yao	北京大学&微软

ACL 终身成就奖

ACL 终身成就奖由 Mark Steedman 获得。

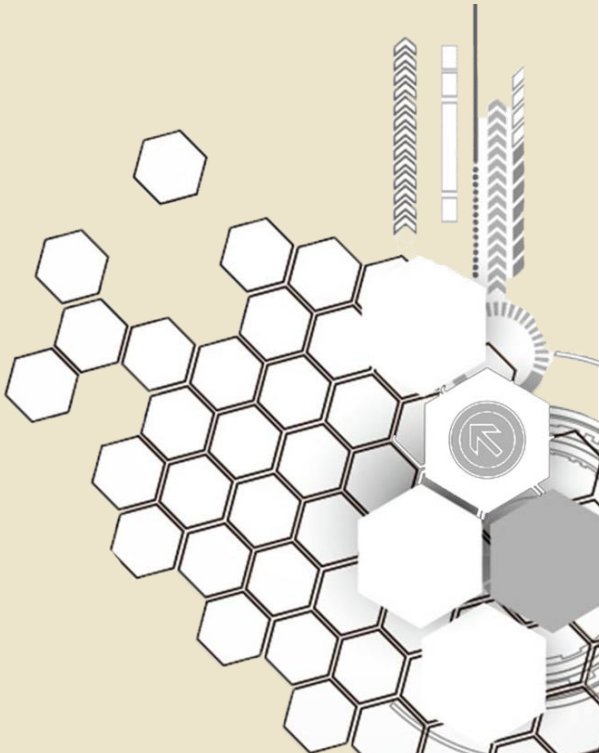


Mark Steedman 出生于 1946 年，1968 年毕业于苏塞克斯大学（University of Sussex），1973 年，获得爱丁堡大学人工智能博士学位（论文：《音乐知觉的形式化描述》）。此后，他曾担任华威大学心理学讲师，爱丁堡大学计算机语言学讲师，宾夕法尼亚大学计算机与信息科学学院副教授，也曾在德克萨斯大学奥斯汀分校，奈梅亨马克斯普朗克心理语言研究所和费城宾夕法尼亚大学担任过访问学者。

目前他任爱丁堡大学信息学院认知科学系主任，主要研究领域有计算语言学、人工智能和认知科学、AI 会话的有意义语调生成、动画对话、手势交流以及组合范畴语法（Combinatory categorial grammar，CCG）等。此外，他对计算音乐分析和组合逻辑等领域也很感兴趣。

4 trend

应用篇



4 应用篇

从知识产业角度来看，自然语言处理软件占有重要的地位，专家系统、数据库、知识库，计算机辅助设计系统（CAD）、计算机辅助教学系统（Cal）、计算机辅助决策系统、办公室自动化管理系统、智能机器人等，全都需要自然语言做人机界面。长远看来，具有篇章理解能力的自然语言理解系统可用于机器自动翻译、情报检索、自动标引及自动文摘等领域，有着广阔的应用前景。

随着自然语言处理研究的不断深入和发展，应用领域越来越广。

文本方面的应用主要有：基于自然语言理解的智能搜索引擎和智能检索、智能机器翻译、自动摘要与文本综合、文本分类与文件整理、智能自动作文系统、自动判卷系统、信息过滤与垃圾邮件处理、文学研究与古文研究、语法校对、文本数据挖掘与智能决策以及基于自然语言的计算机程序设计等。

语音方面的应用主要有：机器同声传译、智能远程教学与答疑、语音控制、智能客户服务、机器聊天与智能参谋、智能交通信息服务（ATIS）、智能解说与体育新闻实时解说、语音挖掘与多媒体挖掘、多媒体信息提取与文本转化以及对残疾人智能帮助系统等。

此外，建立在自然语言处理技术基础之上的心理学、认知学、哲学、混沌学说的共同发展，将使人们对智能的起源问题有新的认识。如果把计算机网络和未来的网格看作是由机器组成的机器社会，那么一种属于机器的智能可能会因为人类的参与以及机器社会中各元素的相互作用而自然诞生。这样，机器必将能够通过“图灵测试”，达到“会思考”的层次。而有关智能机器的研究也会诞生一系列新的领域，比如，机器心理学和机器认知学等。

其中，机器心理学主要研究机器的心理反应和意图。美国圣迭戈神经科学研究所研制的机器人 DarwinV II，能够根据其感知对外部事物进行分类，并根据经验和知识采取相应的对策。然而，机器心理学的研究不能局限于此，人们还需要对机器的意识、知觉、思想、情感、情绪、创造力、机器社会、机器交流等方面进行研究，而这一切还需要计算机科学、心理学、神经科学的同步发展。

我们选取一些自然语言处理应用较为频繁的场景进行介绍。

- **知识图谱**

知识图谱能够描述复杂的关联关系，它的应用极为广泛，最为人所知的就是被用在搜索引擎中丰富搜索结果，并为搜索结果提供结构化结果体现关联，这也是 google 提出知识图谱的初衷。同时微软小冰、苹果 siri 等聊天机器人中也加入了知识图谱的应用，IBM Watson 是问答系统中应用知识图谱较为典型的例子。按照应用方式，可以将知识图谱的应用分为语义搜索、知识问答、以及基于知识的大数据分析和决策等。

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。知识问答是基于知识库的问答，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。在大数据的分析和决策方面，知识图谱起到了辅助作用，典型应用是美国 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反映出英剧版《纸牌屋》很受欢迎，于是拍摄了美剧《纸

牌屋》，大受追捧。

- **机器翻译**

机器翻译是自然语言处理最为人知的应用场景，一般是将机器翻译作为某个应用的组成部分，例如跨语言的搜索引流等。目前以 IBM、谷歌、微软为代表的国外科研机构和企业均相继成立机器翻译团队，专门从事智能翻译研究。如 IBM 于 2009 年 9 月推出 ViaVoice Translator 机器翻译软件，为自动化翻译奠定了基础；2011 年开始，伴随着语音识别、机器翻译技术、DNN（深度神经网络）技术的快速发展和经济全球化的需求，口语自动翻译研究成为当今信息处理领域新的研究热点；Google 于 2011 年 1 月正式在其 Android 系统上推出了升级版的机器翻译服务；微软的 Skype 于 2014 年 12 月宣布推出实时机器翻译的预览版、支持英语和西班牙语的实时翻译，并宣布支持 40 多种语言的文本实时翻译功能。

尤其值得之注意的是，在“一带一路”这一发展背景下，合作沟通会涉及 60 多个国家、53 种语言，此时机器翻译的技术应用显得尤为重要，语言的畅通是“一带一路”战略得以实施的重要基础。而机器翻译涉及到语义分析、上下文环境等诸多挑战，其发展道路还有很长一段路要走。

- **聊天机器人**

聊天机器人是指能通过聊天 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，其特点是成本低、高效且持续工作。例如 siri，小娜等对话机器人是一个应用场景。除此之外，聊天机器人在一些电商网站有着很实用的价值，可以充当客服角色，例如京东客服 jimi，有很多基本的问题，其实并不需要真的联系人工客服来解决。通过应用智能问答系统，可以排除掉大量的用户问题，比如商品的质量投诉、商品的基本信息查询等程式化问题，在这些特定的场景中，特别是会被问到高度可预测的问题中，利用聊天机器人可以节省大量的人工成本。

- **文本分类**

文本分类是指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。这一技术的关键问题是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。

按照其领域分类不同的期刊、新闻报道，甚至多文档分类也是可能的。文本分类的一个重要应用之处是垃圾电子邮件检测，除此之外，腾讯、新浪、搜狐之类的门户网站每天产生的信息分繁杂多，依靠人工整理分类是一项耗时巨大的工作且很不现实，此时文本分类技术的应用就显得极为重要。

- **搜索引擎**

自然语言处理技术例如词义消歧、句法分析、指代消解等技术在搜索引擎中常常被使用。搜索引擎的职责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。搜索引擎最基本的模式是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。每一个环节，都需要用到自然语言处理。用百度举例，比如用户可以搜“天气”、“日历”、“机票”及“汇率”这样的模糊需求，会直接在搜索结果呈现结果。用户还可以搜索“范冰冰演过的电视剧”这样的复杂问题，百度都可以准确地回答。

一方面，有了自然语言处理技术才使得搜索引擎能够快速精准的返回用户的搜索结果，几乎所有的自然语言处理技术都在搜索引擎中有应用的影子；另一方面，搜索引擎（例如谷歌商业帝国和百度巨头）在商业上的成功，也促进了自然语言处理技术的进步。

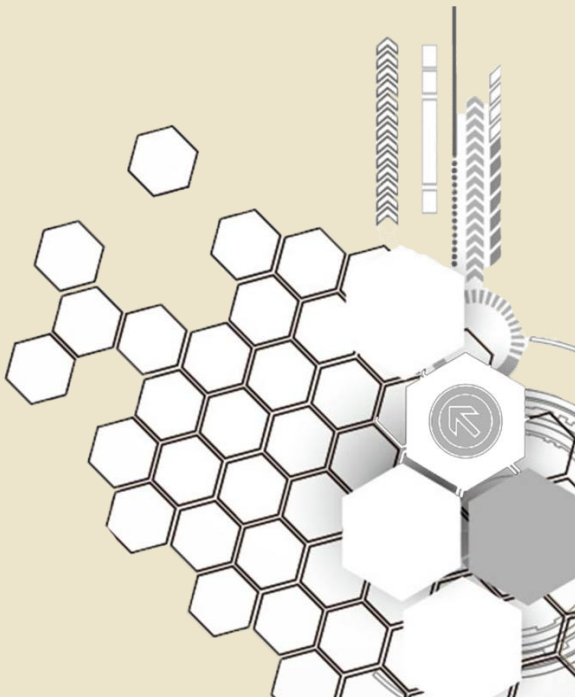
- **推荐系统**

第一个推荐系统是 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，利用用户的标注和行为信息对邮件进行重排序。推荐系统依赖数据、算法、人机交互等环节的相互配合，应用了数据挖掘技术、信息检索技术以及计算统计学等技术使用推荐系统的目的是联系用户和信息，帮助用户发现对自己有价值的信息，同时让信息能够展示在对它感兴趣的用户面前，精准推荐，用来解决信息过载和用户无明确需求的问题。

推荐系统在音乐电影的推荐、电子商务产品推荐、个性化阅读、社交网络好友推荐等场景发挥着重要的作用，美国 Netflix 2/3 的电影是因为被推荐而观看，Google news 利用推荐系统提升了 38%的点击率，Amazon 的销售中推荐占比高达 35%。

5 application

趋势篇



5 趋势篇

随着深度学习时代的来临，神经网络成为一种强大的机器学习工具，自然语言处理取得了许多突破性发展，情绪分析、自动问答、机器翻译等领域都飞速发展。

下图分别是 AMiner 计算出的自然语言处理近期热点和全球热点。通过对 1994-2017 年间自然语言处理领域有关论文的挖掘，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。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。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。

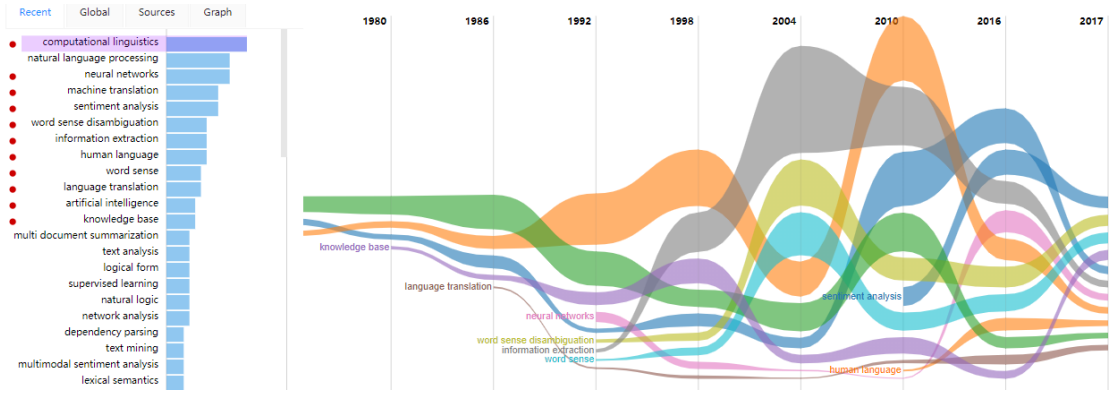


图 16 自然语言处理近期热点图

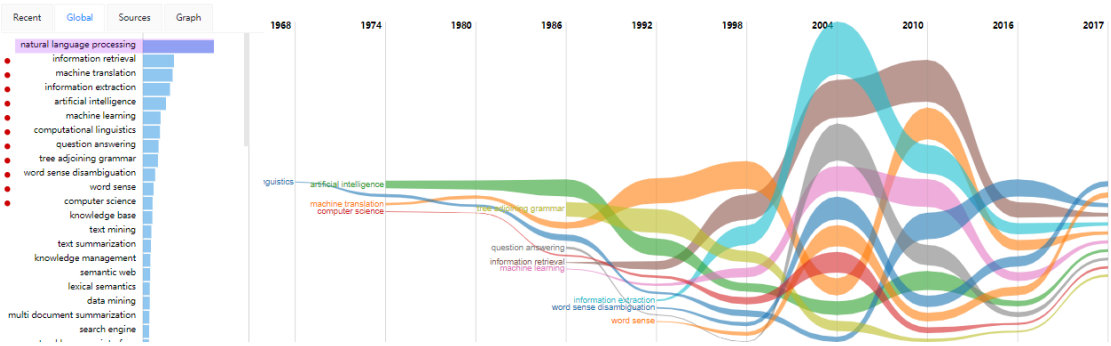


图 17 自然语言处理全球热点图

图 14 显示，情绪分析、词义消歧、知识库和计算机语言学将是最近的热点发展趋势，图 15 显示词义消歧、词义理解、计算机语言学、信息检索和信息提取将是自然语言处理全球热点。

我们同时在微博@ArnetMiner 中发起了关于自然语言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%。我们依据排列由高到低选取其中几项展开介绍。

- **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。文本理解和推理是自然语言处理的重要部分，现在的机器软件已经可以根据文本的语境上下文分辨代词等指示词，这是文本理解与推理从浅层分析向深度理解迈进的重要一步。

- **对话机器人：实用化、场景化**

从最初 2012 年到 2014 年的语音助手，到 2014 年起逐渐出现的聊天机器人微软小冰、百度小度，再到 2016 年哈工 SCIR-笨笨，对话机器人越来越智能。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，即做特定场景时有用的人机对话。

- **NLP+行业：与专业领域深度结合**

银行、电器、医药、司法、教育等领域对自然语言处理的需求都非常多。自然语言处理与各行各业的结合越来越紧密，专业化的服务趋势逐渐增强。刘挺教授预测，自然语言处理首先会在信息准备充分，并且服务方式本身就是知识和信息的领域产生突破，例如医疗、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**

自然语言处理中学习模式有一个较为明显的变化。在浅层到深层的学习模式中，浅层学习是分步骤的，深度学习的方法贯穿在浅层分析的每个步骤中，由各个步骤连接而成。而直接的深度学习则是直接从端到端，人为贡献的知识在深度学习中所占的比重大幅度减小。但如何将深度学习应用于自然语言处理需要进行更多的研究和探索，针对不同任务的不同字词表示，将先验知识和深度学习相结合是未来的一个发展趋势。

- **文本情感分析：事实性文本到情感文本**

之前的研究主要是新闻领域的事实性文本，现在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。如 2017 年新浪微舆情和哈工大推出“情绪地图”，网民可以登录新浪舆情官方网站查询任何关键词的“情绪地图”，这是语义情绪分析在舆情分析产业的首次正式应用。

参考文献

[1] 中文信息处理发展报告 2016

[2] 李涓子，侯磊 知识图谱研究综述.[J]山西大学学报 2017

[3] 冯志伟.机器翻译研究.[M].北京：中国对外翻译出版社.2004

[4] 冯志伟.自然语言处理的形式模型[M].北京：中国科学技术大学出版社 2010

[5] 吴军，数学之美[M].北京：人民邮电出版社 2012

[6] 2006-2020 年国家信息化发展战略[Z] 中共中央办公厅、国务院办公厅 2006

[7] 刘奕群，马少平，洪涛等 搜索引擎技术基础[M] 北京：清华大学出版社 2010

[8] 韩家炜等，数据挖掘：概念与技术[M] 北京：机械工业出版社 2012

版权声明

AMiner

科技情报大数据挖掘服务平台

AMiner 平台由清华大学计算机系研发，拥有我国完全自主知识产权。系统 2006 年上线，吸引了全球 220 个国家和地区 800 多万独立 IP 访问，数据下载量 230 万次，年度访问量 1000 万，成为学术搜索和社会网络挖掘研究的重要数据和实验平台。

项目团队与中国工程科技知识中心、微软学术搜索、ACM、IEEE、DBLP、美国艾伦研究所、圣母大学、英国南安普顿大学等国际知名机构建立了良好的合作关系，共享数据及技术资源。系统相关核心技术申请专利 20 余项，发表论文 300 余篇，其中顶级期刊和会议（CCFA 类）60 篇，编著英文论著两部，Google 引用超过 8000 次，SCI 他引超过 1000 次。

项目成果及核心技术应用于工程院、科技部、自然基金委、华为、腾讯、阿里巴巴、百度等国内外 20 多家企事业单位，为各单位的系统建设及产品升级提供了重要数据及技术支撑。