头豹 LeadLeo

400-072-5588

头豹研究院 | 人工智能系列深度研究

2019 年
中国自然语言处理行业研究报告

行业走势图



TMT 团队

卢佩珊　　分析师

邮箱：cs@leadleo.com

相关热点报告

- 人工智能系列深度研究——科创板上市，人工智能行业借势迎来发展
- 人工智能系列深度研究——沙利文助力 2019 全球人工智能健康峰会成功举办，共同推动全球人工智能技术和产业发展
- 人工智能系列深度研究——谷歌上线“猜画小歌”，智能技术与人们生活越来越近
- 人工智能系列深度研究——产品智能化浪潮开启，智能家居风口已至

报告摘要

自然语言处理技术是众多人工智能设备（如智能家居设备、智能机器人、智能助手等）不可或缺的核心技术，随着人工智能的深入发展，自然语言处理需求不断提升，智能应用需要自然语言处理技术帮助其实现智能化。

热点一：市场需求增长，自然语言处理市场空间广阔

自然语言处理技术是众多人工智能设备（如智能家居设备、智能机器人、智能助手等）不可或缺的核心技术，随着智能设备数量增长以及行业智能化业务处理水平要求的提高，自然语言处理市场有望得到进一步拓展。

热点二厂商集研发算法、解决方案和应用产品功能于一身

目前中国的自然语言处理厂商较多集研发算法、解决方案以及应用产品功能于一身，厂商自主研发自然语言处理算法，形成一整套自然语言处理关键技术方案，并将自主研发的自然语言处理算法以及技术方案内嵌于自有应用产品体系中。

热点三：自然语言处理应用逐渐成熟，实现广泛落地

得益于自然语言处理技术研究的不断深入，自然语言处理应用逐步得到推广，尤其在知识图谱、机器翻译、阅读理解和智能创作等方面开始产生较成熟的应用。金融、医疗、法律等传统行业的业务处理智能化水平要求上涨，加速自然语言处理技术行业应用落地。

目录

1 方法论 5

1.1 方法论 5

1.2 名词解释 6

2 中国自然语言处理行业市场综述 7

2.1 自然语言处理定义 7

2.2 自然语言处理环节 7

2.3 自然语言处理技术应用 8

2.4 中国自然语言处理产业链分析 9

2.4.1 产业链上游 10

2.4.2 产业链中游 12

2.4.3 产业链下游 13

2.5 中国自然语言处理市场规模 13

3 中国自然语言处理行业驱动因素 15

3.1 自然语言处理要素演进，行业迎来变更式发展 15

3.2 传统行业智能需求增长，带动语言处理需求上涨 16

4 中国自然语言处理行业制约因素 17

4.1 自然语言处理存在技术难题 17

4.2 自然语言处理模型通用性不强 17
4.3 机器理解自然语言技术未有突破 18
5 中国自然语言处理行业相关政策法规 20
6 中国自然语言处理行业发展趋势 22
6.1 多模态语言处理融合 22
6.2 自然语言处理应用逐渐成熟 22
7 中国自然语言处理行业竞争格局 23
7.1 中国自然语言处理行业竞争现状分析 23
7.2 中国自然语言处理行业典型企业分析 24
7.2.1 新译信息科技（深圳）有限公司 24
7.2.2 苏州思必驰信息科技有限公司 26
7.2.3 北京智齿博创科技有限公司 27

图表目录

图 2-1 中国自然语言处理产业链.....10

图 2-2 中国自语言处理市场规模（按营收计），2014-2023 年预测.....14

图 5-1 中国自然语言处理行业相关政策.....21

1 方法论

1.1 方法论

头豹研究院布局中国市场，深入研究 10 大行业，54 个垂直行业的市场变化，已经积累了近 50 万行业研究样本，完成近 10,000 多个独立的研究咨询项目。

- ✓ 研究院依托中国活跃的经济环境，从人工智能、机器学习、深度学习、语音识别、语义理解等领域着手，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

- ✓ 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

- ✓ 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

- ✓ 头豹研究院本次研究于 2019 年 07 月完成。

1.2 名词解释

- **兼类词**：有两种或两种以上词性的词。
- **未登录词**：没有被收录在分词词表中但必须切分出来的词，包括各类专有名词（人名、地名、企业名等)、缩写词、新增词汇等。
- **义项**：每一个不同概念意义事物的叙述内容。
- **GPU**：由大量核心组成的大规模并行计算架构，专为同时处理多重任务而设计的芯片。
- **FPGA**：现场可编程门阵列，是在可编程器件的基础上进一步发展的半定制电路。
- **ASIC**：专用集成电路，是为符合特定用户需求而设计的专用人工智能芯片。
- **自适应**：计算机根据数据特征自动调整处理方法、处理顺序、处理参数、边界条件或约束条件，使其与所处理数据的统计分布特征、结构特征相适应，以取得最佳处理效果的过程。
- **AIUI**：由科大讯飞研发的人机交互新产品，集成了双全工技术、麦克风阵列技术、声纹识别技术、方言识别、语义理解技术和内容服务等。

2 中国自然语言处理行业市场综述

2.1 自然语言处理定义

自然语言处理是通过构建算法使计算机自动分析、表征人类自然语言的学科。自然语言处理是计算机理解和生成自然语言的过程，自然语言处理技术使计算机具有识别、分析、理解和生成自然语言文本（包括字、词、句和篇章）的能力。自然语言处理机制涉及自然语言理解和自然语言生成两个流程：(1) 自然语言理解：计算机理解自然语言文本的思想和意图；（2）自然语言生成：计算机用自然语言文本表述思想和意图。

自然语言理解和分析是一个层次化过程，从词法分析、句法分析、语义分析到语用语境分析层层递进：(1) 词法分析：分析词汇的各个词素，从中获得语言学信息；(2) 句法分析：分析句子和短语的结构，识别各词语、短语在句中的作用以及相互间的关系；(3) 语义分析：找出词义、结构意义及词与结构结合的意义，确定语言所表达的真正含义；（4）语用语境分析：分析语言所存在的外界环境对语言使用者所产生的影响。

2.2 自然语言处理环节

（1） 词法分析

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，词性标注是在给定句子中判断并标注各词的词性，而兼类词和未登录词的词性复杂难以确定，标注兼类词与未登录词的词性是词法分析的重要任务。词义标注是在具体语境中明确各词的词义，如多义词拥有多种意义，但在具体语境中表达的意义是可确定的。在不同的具体语境中解决多义词的义项问题是词义标注的重点。

（2） 句法分析

句法分析的基本任务是确定句子的语法结构或句子中词汇间的依存关系，包括确定语言

的语法体系，明确符合语法规则的句子的语法结构以及通过分析语言单位内成分间的依存关系推导句子的句法结构。

（3） 语义分析

语义分析通过建立有效的模型使计算机系统能对各个语言单位（包括词汇、句子和篇章等）进行自动语义分析，从而理解自然语言文本的真实语义。根据理解对象的语言单位不同，可将语义分析分为词汇级语义分析、句子级语义分析以及篇章级语义分析。词汇级语义分析关注如何获取或区别单词的语义，句子级语义分析关注整个句子所表达的语义，篇章级语义分析研究篇章文本的内在结构以及理解篇章文本内语言单元（句子、从句或段落）间的语义关系。

（4） 语用语境分析

语用指人对语言的具体运用，自然语言用语与语境、语言使用者的知识涵养、言语行为、想法和表达意图密切相关。语用分析是计算机在情景语境和文化语境中研究分析语言使用者的表达用意。

2.3 自然语言处理技术应用

随着人工智能的深入发展，自然语言处理需求不断提升，众多类型智能应用需要自然语言处理技术帮助其实现智能化，如（1）文本领域的搜索引擎、信息检索、机器翻译、自动摘要、文本分类、意见挖掘、舆情分析、自动判卷系统、信息过滤和垃圾邮件处理等应用；（2）语音领域的语音助手、智能客服、聊天机器人、自动问答、智能解说和智能远程教学与答疑等应用均需自然语言处理技术理解或生成自然语言。

①自动问答

自动问答应用涉及自然语言的词法、句法、语义等分析问题，是自然语言理解与生成技术应用的集中体现。自动问答系统能自动回答用户提出的问题，反馈给用户基于自然语言表

述的答案，不再是简单的基于关键词匹配排序的文档列表，系统在生成答案的操作中需要正确理解用户所提出的问题，抽取用户问题中的关键信息，进而检索语料库或知识库，将可匹配的最佳答案用自然语言的形式反馈给用户，完成自动问答任务。

②信息检索

信息检索是计算机自主从文档集合中查找用户所需信息的过程。信息检索系统将信息标引、描述以及组织整理后存在于数据库中，将用户输入的检索关键词与数据库中信息的标引词匹配，实现用户的信息检索要求。信息检索要求计算机理解用户输入的自然语言信息，自动将自然语言信息与数据库中的标引信息进行比对，以达成检索任务。如谷歌搜索引擎可通过理解用户输入的自然语言关键词，反馈给用户一个检索目标页面列表，用户可在列表中选择能够满足自己信息需求的页面加以浏览。因为搜索引擎无法通过简单的关键词表达体会用户真正的查询意图，只能将所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。

③情感分析

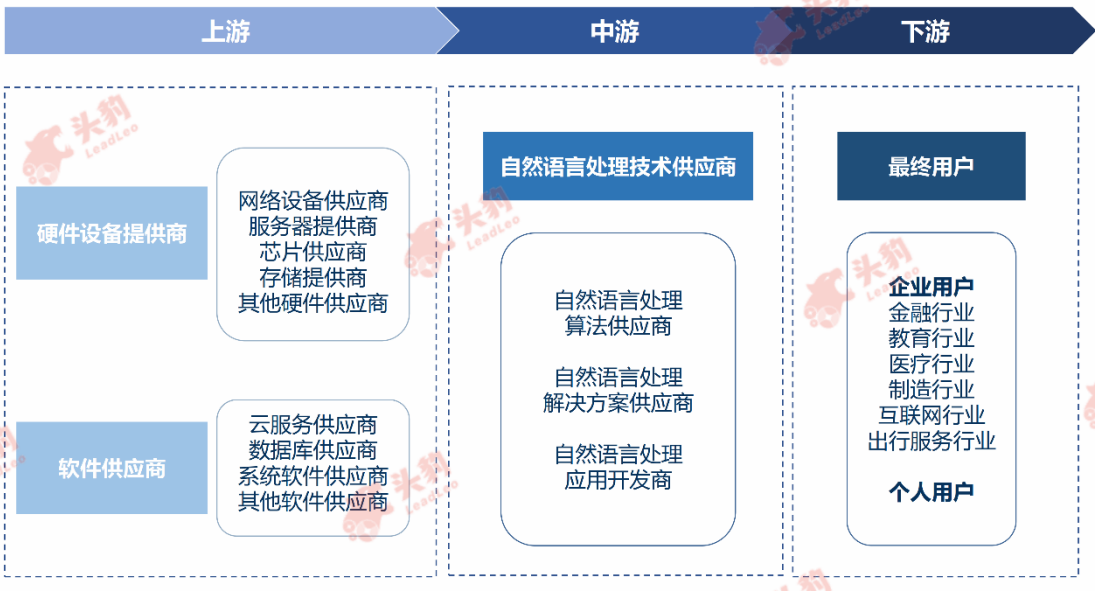
情感分析（意见挖掘），是计算机系统自主对文本的情感倾向（如主观/客观，积极/消极，喜欢/讨厌等）进行挖掘和分析的过程。情感分析能帮助商家自动处理用户评论，将分析过的评论按照排序规则进行展示，帮助商家获得广告营销效果，如淘宝、天猫等电商平台、携程、爱彼迎等旅游住宿平台的商家在评论区可设置自动置顶反映积极情绪的用户评论，置后部分反映负面信息的用户评论，达到吸引用户眼球的效果。商家还可通过分析用户发布的信息了解用户喜好，实现精准营销，如新浪微博上的零售商可根据用户发表的微博，微话题等内容了解用户的个人喜好，为用户定制性的推送优惠及新品信息。

2.4 中国自然语言处理产业链分析

自然语言处理产业链上游市场主体为基础资源提供商，包括硬件供应商（如芯片供应商、服务器供应商和存储供应商等）和软件供应商（如云服务供应商和数据库供应商等）；中游

市场由自然语言处理算法供应商、自然语言处理解决方案供应商以及自然语言处理应用供应商组成，负责为下游需求端提供服务；下游市场主体为各类型用户，包括企业用户和个人用户，企业用户涉及金融、医疗、教育、出行服务、互联网服务等领域，个人用户则为最终消费者。

图 2-1 中国自然语言处理产业链



来源：头豹研究院编辑整理

2.4.1 产业链上游

自然语言处理产业链上游市场由基础资源供应商组成，涉及网络设备、服务器、芯片、存储、云服务、数据库等软、硬件供应商，负责为自然语言处理技术和产品开发商提供必要的资源支持。

（1） 芯片供应商

现阶段，行业内尚未出现专门用于自然语言处理运算的芯片，核心数据处理芯片 CPU 无法执行自然语言处理结构化运算，目前适用于自然语言处理的芯片类型有 GPU、FPGA、ASIC 和 DSP。

GPU 解决浮点运算、数据并行计算问题优势明显，可提供高密度运算能力，解决大量数据元素并行问题。但 GPU 芯片功耗大，依托于 X86 架构服务器而运行，成本高昂，不适

用于广泛的自然语言处理产品方案的开发，在自然语言处理与传统行业数字化进程结合加深的趋势下，采用 GPU 作为自然语言处理运算芯片的方案不具备成本优势，小型自然语言处理应用项目负担不起高昂成本。

FPGA 具有可编程性，设计者可根据需要的逻辑功能对 FPGA 电路进行快速烧录，从而改变其出厂设计，灵活性强。但 FPGA 的设计布线相对固定，各种型号的 FPGA 芯片逻辑资源相对固定，选定了型号即决定了芯片的逻辑资源上限，无法随意增加运算能力。

ASIC 芯片的运算能力强、规模量产成本低，全定制设计需要设计者完成所有电路的设计，开发周期长，时间成本高昂，主要适用于量大、对运算能力要求较高、开发周期较长的领域。

DSP 内有控制单元、运算单元、各种寄存器以及存储单元，其外围还可以连接若干存储器和一定数量的外部设备，有软、硬件的全面功能，本身是一个微型计算机，运算能力强、速度快、体积小，而且采用软件编程具有高度的灵活性。但目前 DSP 的性能并未通过实践验证，也未生产出可以与 GPU 相匹敌的芯片器件，商业化应用仍在研发过程中。

为满足自然语言处理等人工智能的发展需求，部分针对深度学习的芯片，如 TPU、NPU、DPU 和 BPU 等相继面世，但受场景以及性能限制，专用的人工智能芯片发展尚未成熟。目前自然语言处理运算的最佳芯片方案仍以 GPU 为主导。

（2） 云服务供应商

云服务供应商为自然语言处理研发企业提供基础设施平台，解决自然语言处理技术研发厂商的数据存储、运算以及调用问题。由于性价比、部署方式等因素，自然语言处理研发企业较多选用公有云服务。

目前，公有云服务供应商有：①通过云服务产业链资源优势拓展至公有云服务行业的企业，如电信运营商，网络设备制造商，IDC 厂商等，此类企业拥有较强的资金实力，加

上本身处在公有云产业链上游，基础设施方面优势明显；②大型互联网企业，如亚马逊，腾讯、阿里巴巴等，此类企业资金实力雄厚，客户认可度高，设施齐备、技术成熟，具备发展公有云业务的有利条件；③传统的软件企业，如 Microsoft、Oracle、金蝶等，此类企业的软件产品的市场认可度高，技术积累丰厚，客户资源丰富，有利于向公有云市场拓展。除此之外，行业中存在不少新兴的创业公司，如青云、Ucloud、七牛云等。

（3） 数据

数据是人工智能发展的基石，海量数据为训练人工智能提供原材料。近年来，由学术及研究机构承担建设的公共数据集不断丰富，数据质量不断提高，利于人工智能企业提高智能模型的准确度。例如，可运用于自然语言处理训练的数据集类型不断丰富，维基百科语料库、斯坦福大学问答数据集、亚马孙美食评论集、康奈尔电影对话语料库、经济新闻相关文章等语言集合相继建成，内容覆盖媒体用语、网络用语、电影用语、政府用语等众多自然语言应用场景，有助于自然语言处理研发企业优化用于处理不同领域自然语言的模型的准确度。

2.4.2 产业链中游

自然语言处理产业链中游市场主体主要有自然语言处理算法提供商、解决方案提供商以及应用产品开发商。目前中国的自然语言处理厂商较多集研发算法、解决方案以及应用产品功能于一身，厂商自主研发自然语言处理算法，形成一整套自然语言处理关键技术方案，并将自主研发的自然语言处理算法以及技术方案内嵌于自有应用产品体系中，典型代表有百度、阿里巴巴和腾讯。

百度自然语言处理算法研究覆盖面广，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、知识挖掘等自然语言处理细分领域。百度积累了解决问句理解、答案抽取、观点分析与聚合等环节的一整套深度问答技术方案，目前已将该套技术方案应用于百度搜索引擎、百度手机浏览器、百度翻译、百度语音助手、小度机器人等多个产

品中。百度在自然语言篇章理解方面，形成篇章结构分析、主体分析、内容标签、情感分析等关键技术，且该类关键技术已在百度搜索、百度信息流、糯米等产品中实现应用。

阿里巴巴开展自然语言处理技术研究主要为旗下产品服务，如阿里巴巴在其电商平台中构建知识图谱实现智能导购，对电商用户进行兴趣挖掘实现精准营销，在蚂蚁金融、淘宝卖家等客服场景中实现机器人提供客服服务，在跨境电商业务中采用机器翻译服务进行商家商品信息翻译、广告词翻译以及买家采购需求翻译等。

2.4.3 产业链下游

自然语言处理产业链下游市场主体为各类型用户，包括企业用户和个人用户。企业用户主要购买行业应用，如智能客服产品、舆情分析产品、文本分类产品等，帮助企业用户提升业务处理的智能化水平。目前的 B 端市场是自然语言处理厂商竞争的焦点，部分应用产品（如智能客服、舆情分析产品等）尝试了商业化运作，市场反馈良好，但众多细分领域市场发展并未成熟，市场空间仍待挖掘。

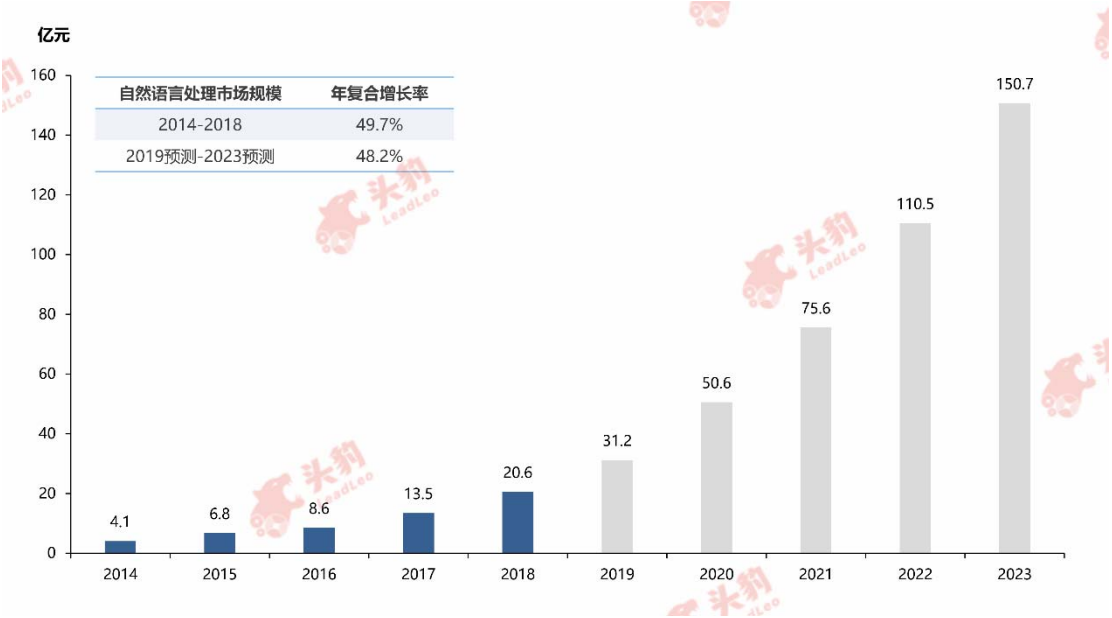
个人用户主要使用手机语音助手、机器翻译软件、信息检索以及互联网搜索等服务。个人用户使用的自然语言处理技术应用产品较多是自然语言处理厂商免费提供的，自然语言处理厂商普遍未在 C 端市场开发清晰的商业模式。

2.5 中国自然语言处理市场规模

现阶段，自然语言处理技术商业化并不成熟，部分已实现商业化应用的自然语言处理技术相关产品（如智能客服、搜索引擎等）均无法将收益直接归因于自然语言处理技术，单独运用自然语言处理技术的产品应用（文档分类、舆情分析等）尚未产生明显受益，因此自然语言处理技术产生的市场营收规模仍然较小，截至 2018 年，自然语言处理市场营收规模仅为 20.6 亿元人民币。但自然语言处理技术是众多人工智能设备（如智能家居设备、智能机器人、智能助手等）不可或缺的核心技术，随着智能设备数量增长以及行业智能化业务处理

水平要求的提高，自然语言处理市场有望得到进一步拓展，自然语言处理市场营收规模有望在 2019-2023 年间以 48.2%的年复合增速实现快速增长。

图 2-2 中国自语言处理市场规模（按营收计），2014-2023 年预测



来源：头豹研究院编辑整理

3 中国自然语言处理行业驱动因素

3.1 自然语言处理要素演进，行业迎来变更式发展

数据量、运算力和算法模型是影响自然语言处理行业发展的三大要素。2012 年以来，数据量的上涨、运算力的提升和深度学习算法的出现促进了自然语言处理行业的快速发展。互联网、社交媒体、移动设备的普及，使产生并存储的**数据量**急剧增加，2020 年全球将有超 500 亿的终端与设备联网，产生的数据总量将大于 40 泽字节，人均每天产生的数据量预计达 1.5GB。中国 2020 年的数据总量将达到全球数据总量的 20.0%，联网设备预计从 2016 年的 8.4 亿个增长至 35.0 亿个，连接设备数的快速增长产生了大量的应用数据。数据的爆发式增长有助于优化自然语言处理算法，海量优质的场景数据能够帮助算法模型完成高效精准的识别训练。

算法是计算机基于其所训练的数据集归纳出的识别逻辑，算法模型的优化可使自然语言处理技术更精准地理解与生成自然语言文本。深度学习算法通过使用复杂结构设计和各种梯度技术，将多重非线性结构变换成多个处理层，实现对大量数据样本的抽象计算，拟合出一个可处理新输入信息的函数模型，解决数据分类或预测问题。基于深层神经网络的深度学习方法从根本上改变了自然语言处理技术的面貌，改变了自然语言处理问题的定义，变更了自然语言处理所使用的数学工具，深度学习的出现突破了传统浅层学习算法的局限，重塑了自然语言处理算法的设计思路，极大地促进自然语言处理研究的发展。

运算力方面，GPU、FPGA，ASIC 等专用芯片的出现，缓解了自然语言数据处理速度难题，解决传统的 CPU 芯片算力不足问题。新兴的 GPU、FPGA、ASIC 等具有良好的并行计算能力，芯片性能大幅优于 CPU，算力比 CPU 芯片强数十倍甚至百倍，缩短了模型运算时间，使得自然语言运算模型的进步速度加快。

3.2 传统行业智能需求增长，带动语言处理需求上涨

金融、医疗、法律等传统行业的业务处理智能化水平要求上涨，加速自然语言处理技术在行业应用中的落地进程。例如，智能问答、资讯舆情分析、文档信息抽取、文档自动生成等应用逐渐**在金融领域**的智能投研、智能投顾、智能客服和智能运营等场景实现应用。在智能投研场景中，投研人员每天需要通过多渠道搜索大量金融资讯信息以掌握金融市场动态，而金融资讯信息极为丰富（如涉及标的公司重大新闻、公告、财务状况，金融产品信息、宏观经济状况、政策法规变化、社交媒体评论等），数据量庞大，单靠投研人员从浩如烟海的资讯库中搜索并阅读分析相关内容，时间精力耗费严重，工作效率难以提高，在此种情况下，具备资讯分类（按公司、产品、行业范围分类）、情感分析（分析新闻、公司或产品的正负面信息）、自动文摘（分析文章的主要内容）及资讯个性化推荐功能的资讯舆情分析应用逐步受到投研人士和金融机构的推崇，资讯舆情分析应用对海量定性数据自动摘要、归纳、缩简和抽取，为投研人员提供高效的信息展示，大幅缩短投研人员搜索以及摘取资讯的时间，提高其工作效率。

在法律领域，自然语言处理应用可帮助法律从业人员进行案例搜索、判决预测、法律文书自动生成、法律文本翻译等，实现事件预处理，减少从业人员相关案件处理耗费时间。**在医疗领域**，病历的辅助录入、医学资料的检索与分析、辅助诊断等医疗步骤可由相关自然语言处理应用辅助进行。现代医学资料浩如烟海，新的医学手段、方法发展迅猛，医生和专家无法及时完全掌握所有的医学发展动态，相关自然语言处理应用可帮助医生快速准确地寻找各种疑难病症的最新研究进展，为医生的诊断提供及时有效的参考。

随着自然语言处理技术的深入发展，自然语言处理应用有望与更多传统行业实现更深层次的结合，为人类带来更多人工智能效应。

4 中国自然语言处理行业制约因素

4.1 自然语言处理存在技术难题

自然语言处理与语音识别以及图像识别不同，自然语言处理不是单轮处理操作过程，算法模型对单一输入进行单轮分析无法输出结果，自然语言的语义与上下文或前后轮对话语境有关，需要对输入的自然语言进行多轮分析方可得到结果。而目前的深度学习技术，多轮建模发展并不成熟，与语音识别以及图像识别的单轮分析模型的成熟度无法相提并论。

此外，人类在语言交流过程中，使用的话语表达内容通常暗含常识性表示（如婴儿、小孩以及成人代表不同年龄段人群），话语表达风格反应用语者的个性化特征，不同的人对相同的内容可有不同的表达方式。现阶段的自然语言处理模型未能解决常识问题和个性化问题，如手机语音助手听到“查找附近的餐馆”指令时，可在地图上执行查找附近餐馆的任务，并显示附近餐馆信息，但若发号指令者说“我饿了”，手机语音助手则不会有任何反应，因为语音助手缺乏“饿了需要进食”的常识，而人类听了“我饿了”的话语，反应一定是“需要进食”。众多潜藏在人类意识中的常识，计算机并不具备且无法学习，模型开发者亦不知道将常识知识用在机器问答和机器搜索中的效果如何，业内尚待建立常识知识库，测试训练模型。

4.2 自然语言处理模型通用性不强

自然语言使用情况复杂，不同场景、不同语种、不同专业的自然用语差异较大，所需要的自然语言处理层次不同，基于某一领域语料库建立的自然语言处理模型不具通用性，应用于其他领域时处理结果明显较差，系统性能显著下降。自然语言模型不具通用性限制了自然语言产品在不同应用领域的推广，大幅减缓了自然语言处理行业的发展进程。

虽然近年来研究者们针对语言模型、语序模型、语调模型等问题对自然语言处理进行了

众多领域自适应的调试，但每种调试方法只能解决单方面的自适应问题，通用性问题仍然无法解决。自然语言处理厂商无法用单一模型解决自然语言处理方面的所有问题，只能根据不同领域，开发相应的自然语言模型，运行相关领域的知识库，处理特定领域的自然语言问题。但在目前行业应用尚未普及，产品商业模式并不清晰的情况下，自然语言处理厂商缺乏动力为少数应用以及各种领域开发专用算法，导致自然语言处理技术难以大范围推广落地。

4.3 机器理解自然语言技术未有突破

自然语言理解结果不准确，限制自然语言处理应用推广。自然语言理解与生成是机器在人工智能、大数据等技术指引下产生的机器行为。机器由于难以拥有生活常识、文化背景和人类风俗习惯知识储备，对于作为文化载体的自然语言，无法结合具体的语言环境、风俗习惯等因素分析语言内容，只能通过系统设定的规则机械地分析源语的语法结构，语言逻辑无法处理，对于复杂句子结构或层层句式嵌套的情况，机器的自然语言理解准确度不高。机器对自然语言长句的理解能力以及全方位体会语用者的真实表达意图的能力无法在短期内得到明显提高，与人类的语言理解能力相比，差距仍然存在较大。自然语言处理应用暂时只能通过完善部分领域的知识库和语料库，优化部分领域内的自然语言处理应用的用户体验。

innovation 创新地图 map

推广

前哨2020 科技特训营

掌握创新武器　抓住科技红利

Insights into Tech and the Future

直播时间
每周四20:00-21:00

全年50次直播课程
+私享群互动

随报随听



王煜全

海银资本创始合伙人
得到《全球创新260讲》主理人

扫码报名

微信咨询：InnovationmapSM
电话咨询：157-1284-6605

5 中国自然语言处理行业相关政策法规

2015 年以来，中央及各地方政府陆续出台相关政策，积极支持人工智能产业的发展，并推进人工智能在各个细分领域的渗透和壮大。2015 年 7 月，国务院发布《关于积极推进“互联网+”行动的指导意见》，将“互联网+人工智能”列为其中 11 项重点行动之一。2017 年 7 月，国务院印发《新一代人工智能发展规划》，将人工智能上升至国家战略，并确立了人工智能发展三步走的目标，为中国人工智能产业发力奠定制度基础。2017 年 10 月，十九大将“人工智能”写进党代会报告，并提出加快建设制造强国，发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合等意见。截至 2018 年，中央及各地方政府在短短 3 年内出台了 60 余项与人工智能产业发展相关的政策，足以看出国家对人工智能产业化的重视。自然语言处理是人工智能的应用体现，同样受到政策的鼓励发展。

2016 年 5 月，发改委、科技部、工信部及中央网信办联合发布《“互联网”+人工智能三年行动实施方案》要求加强产学研用合作，明确支持国家工程实验室、国家工程（技术）研究中心等创新平台建设，布局国家级创新中心，共同推动人工智能基础理论、共性技术、应用技术研究。加快基于人工智能的计算机视听觉、生物特征识别、复杂环境识别、新型人机交互、自然语言理解、机器翻译、智能决策控制、网络安全等应用技术研发和产业化，加强前沿技术布局，构造未来融合创新技术基础。

2017 年 7 月，国务院发布《新一代人工智能发展规划》，明确提出建立新一代人工智能关键共性技术体系，重点突破自然语言的语法逻辑、字符概念表征和深度语义分析的核心技术，推进人类与机器的有效沟通和自由交互，实现多风格多语言多领域的自然语言智能理解和自动生成。着力研究短文本的计算与分析技术，跨语言文本挖掘技术和面向机器认知智能的语义理解技术，多媒体信息理解的人机对话系统。

2017 年 12 月，工信部印发《促进新一代人工智能产业发展三年行动计划 (2018-2020 年)》，提出面向语音识别、视觉识别、自然语言处理等基础领域及工业、医疗、金融、交通等行业领域，支持建设高质量人工智能训练资源库、标准测试数据集并推动共享，鼓励建设提供知识图谱、算法训练、产品优化等共性服务的开放性云平台。

2018 年 11 月，工信部发布《新一代人工智能产业创新重点任务揭榜工作方案》，提出大力构建行业训练资源库，到 2020 年，基础语音、视频图像、自然语言文本对话等公共训练数据量大幅提升，在工业、医疗、金融、交通等领域汇集一定规模的行业应用数据，用于支持创业创新。

图 5-1 中国自然语言处理行业相关政策

政策名称	颁布日期	颁布主体	主要内容及影响
《新一代人工智能产业创新重点任务揭榜工作方案》	2018-11	工信部	提出大力构建行业训练资源库，到2020年，基础语音、视频图像、自然语言文本对话等公共训练数据量大幅提升，在工业、医疗、金融、交通等领域汇集一定规模的行业应用数据，用于支持创业创新
《促进新一代人工智能产业发展三年行动计划（2018-2020年）》	2017-12	工信部	提出面向语音识别、视觉识别、自然语言处理等基础领域及工业、医疗、金融、交通等行业领域，支持建设高质量人工智能训练资源库、标准测试数据集并推动共享，鼓励建设提供知识图谱、算法训练、产品优化等共性服务的开放性云平台
《新一代人工智能发展规划》	2017-07	国务院	提出建立新一代人工智能关键共性技术体系，重点突破自然语言的语法逻辑、字符概念表征和深度语义分析的核心技术，推进人类与机器的有效沟通和自由交互，实现多风格多语言多领域的自然语言智能理解和自动生成，着力研究短文本的计算与分析技术，跨语言文本挖掘技术和面向机器认知智能的语义理解技术
《“互联网”+人工智能三年行动实施方案》	2016-05	发改委 科技部 工信部 中央网信办	加快基于人工智能的计算机视听觉、生物特征识别、复杂环境识别、新型人机交互、自然语言理解、机器翻译、智能决策控制、网络安全等应用技术研发和产业化，加强前沿技术布局，构造未来融合创新技术基础

来源：头豹研究院编辑整理

6 中国自然语言处理行业发展趋势

6.1 多模态语言处理融合

深度学习神经网络的引进使得语言模态、文字模态、图像模态、视频模态的编码和解码可在同一个深度学习框架下统一运行。不同模态的对象可被同一模式编码与解码，同一模式的编码与解码即可使不同模态对象随意融合，各种语言分析的结果可与语音分析、图像分析结果结合应用，产生更多的产品应用模式，如科大讯飞推出的语音交互系统 AIUI，融合了语音技术和语义理解技术，使智能助手如同人类大脑能多功能区域协同工作，提升了应用的智能化水平以及人机交互效果。

机器的各方面智能不应是相互独立的，应像人类大脑般同时处理听觉、视觉以及触觉认知与感知信息，未来的自然语言处理技术必定与语音处理技术、图像处理技术等人工智能技术相互融合，赋予 AI 高度智能。

6.2 自然语言处理应用逐渐成熟

随着自然语言处理技术研究的不断深入，自然语言处理应用逐步得到推广，尤其在知识图谱、机器翻译、阅读理解和智能创作等方面开始产生较成熟的应用。

知识图谱已在科研、金融、医疗、互联网服务、汽车等领域得到广泛应用。2018 年以来，百度搜索引擎构建知识图谱，为用户提供全面的智能搜索服务。阿里健康与国家级医疗健康大数据平台启动建立医学知识图谱“医知鹿”，并在医疗领域逐步推广。腾讯推出医疗 AI 引擎“腾讯睿知”，发力智能导诊。美团通过构建生活知识图谱为用户提供智能化的生活服务。汽车之家构建汽车领域的知识图谱帮助旗下 APP“家家小秘”实现图谱问答。

智能创作应用方面，百度推出人工智能写作辅助平台“创作大脑”，其语义智能纠错功能识别准确率超 95%，达到大学生写作纠错能力的平均水平，能为人类作者提供良好的纠

错、提取信息等辅助写作服务，提高写作者的写作效率。微软的智能助手小冰开始向作词、作曲、演唱方面发展，能辅助音乐创作者进行音乐创作。IBM 推出的智能辩手曾与以色列国际辩论协会主席以及以色列国家辩论冠军进行辩论比赛，IBM 智能辩手通过处理大量文本，就特定主题构建出有良好结构的演讲，清晰明了地反驳了这两位辩论对手。智能创作应用需要运用到目前业内研发的各种自然语言处理技术，是业界自然语言处理技术水平的良好体现，智能创作应用的成熟表现，有效地反映了业内自然语言处理技术良好的发展成果。



7 中国自然语言处理行业竞争格局

7.1 中国自然语言处理行业竞争现状分析

中国自然语言处理市场参与者可分为互联网巨头企业、自然语言处理技术研发企业以及自然语言处理创业企业，其中互联网巨头企业在自然语言处理行业占据约 80%的市场份额，自然语言处理技术研发企业以及创业企业合计共占 20%的市场份额。

互联网巨头企业通过资金、技术、数据资源等实力，实现自然语言处理的基础软、硬件

层、技术层以及应用层全产业链布局，如百度开发了自然语言处理应用平台，为众多自然语言研发企业开源提供底层研发架构支持，同时百度自主研发自然语言处理基础算法、问答系统、阅读理解等核心技术，并将核心技术应用于百度搜索、百度新闻、百度翻译、百度助手等 C 端产品中，提升了自有产品的智能水平。

自然语言处理技术研发企业在自然语言处理技术方面积累深厚，拓展了稳定的 B 端业务市场，典型代表有搜狗和科大讯飞。例如，科大讯飞推出完善智能硬件语音交互解决方案，帮助企业用户解决语音交互、语义理解等技术研发难题。B 端市场的定制化需求高，对自然语言处理厂商的研发能力以及资金投入要求高，导致缺乏资金、技术积累的自然语言处理创业企业难以大范围拓展 B 端市场。

自然语言创业企业多以具体的自然语言处理场景、垂直细分的应用场景切入自然语言处理市场，为下游用户提供单一类型的行业应用解决方案或消费级产品，如智齿科技为企业用户提供专门的智能客服解决方案。

根据从事自然语言处理研究时间长达 12 年的专家分析，互联网巨头有深厚的资源积累，技术研发实力雄厚，自然语言处理技术的迭代将长期由互联网巨头主导。然互联网巨头的技术研发多以业务布局为导向，旗下产品生态不包含的领域，互联网巨头一般不会涉足，因此为自然语言处理技术研发企业以及创业企业留下市场空间，自然语言处理技术研发企业以及创业企业能从细分领域开拓市场，避开与互联网巨头企业的直接竞争。

7.2 中国自然语言处理行业典型企业分析

7.2.1 新译信息科技（深圳）有限公司

7.2.1.1 企业简介

新译科技成立于 2014 年 12 月，是一家研发人工智能技术的科技公司，以机器翻译和语音识别等技术为基础，向用户提供在线机器翻译、在线辅助翻译平台和互联网语言服务平

台等一系列关于翻译领域的产品和服务。新译科技以软、硬件智能翻译为承载体，服务于国家“一带一路”多语传播平台建设，为金融、专利、法律、医学等垂直领域提供全球化基础性多语沟通服务。

7.2.1.2 产品研发

(1) B 端和 G 端产品：新译科技的 B 端和 G 端产品包括软件和硬件产品，主要面向大型央企、军工企业、互联网媒体及政府机关等企业和政府机构提供服务。

①定制化机器翻译是一个混合神经网络翻译引擎，该服务能统一管理企业语料，给企业特定的数据和管理账号，满足企业个性化、精细化、专业性的需求，达到保护企业数据隐私和数据的效果。

②智能翻译插件包括办公翻译插件和网页翻译插件，办公翻译插件可以在 WPS 和 Office 两大办公系统上使用，满足多语言文档的快速转换和阅读；网页翻译插件支持各大浏览器网页翻译，一键转化母语阅读。

(2) C 端产品：新译科技主要向个人消费者提供在线文档翻译、PC 端翻译、智译 APP 服务以及可穿戴式翻译产品。智译 APP 可实现语音和文本的多种语言智能翻译，适用于旅游、社交、住宿、商务等领域。在线文档翻译支持 10 种常用文档格式输入，可应用于通用或专业领域的文档翻译，翻译质量较高。

7.2.1.3 竞争优势

新译科技拥有较强的技术优势，同时拥有澳门大学自然语言处理与葡中智能翻译（NLP2CT）实验室、葡萄牙里斯本 L2F 语音实验室、美国卡耐基梅隆大学 LTI 实验室以及新译-澳大-清华人工智能研究院四所优质自然语言研究实验室提供的技术人才和原创技术支持，公司所研发的智能翻译产品于 2017 年通过国家科技成果认定，并获得深圳青年技术奖以及吴文俊人工智能科学技术奖。新译科技目前拥有近百人的团队，核心技术研发人员主

要在清华大学自然语言处理研究中心，澳门团队主要开发机器翻译的各个衍生品，深圳团队做自然语言处理底层设计，北京团队负责前端、产品和销售。新译科技未来将会着力提升限定性神经网络翻译技术、增强交互式辅助翻译水平、提高语义翻译的技术、改进自动修正技术提高后编译能力水平。

7.2.2 苏州思必驰信息科技有限公司

7.2.2.1 企业简介

思必驰成立于 2007 年，是研发智能语音交互技术的智能语音服务商，为企业和开发者提供自然语言交互解决方案，包括 DUI 开放平台、企业级智能服务、人机对话操作系统、人工智能芯片模组等。思必驰已于 2018 年 5 月完成由元禾控股、中民投领投，深创投、富士康、联发科跟投的 5 亿元人民币 D 轮融资。

7.2.2.2 产品研发

（1） 语音交互

思必驰为用户提供移动端的语音交互技术支持，适用于 Android、iOS 等系统，可应用于语音助手、智能客服、地图导航等领域，具备语音唤醒、识别、合成以及语义理解等多种能力。在智能车载场景中，思必驰为用户提供“云+端”解决方案，利用智能语音操作系统控制智能后视镜、智能车机、便携式导航仪等设施，实现双手解放。

（2） 智能家居

思必驰拥有软硬一体化的智能家居解决方案，为智能家居产品提供高性能、低成本的声源定位、个性唤醒、语音识别、语义理解、对话交互等功能，使家居产品可适应近场、远场的语音交互，方便用户控制，提高家居产品的智能化水平。

7.2.2.3 竞争优势

思必驰拥有自主知识产权的人机对话、语音识别、语义理解、语音合成、声纹识别等综

合语音技术，其智能人机对话技术突破了传统语音技术不能很好支持复杂语音交互功能的难题，使语音输入不再局限于呆板简单的句式，在复杂环境和自然口语交流的情况下，能保证优异的语音分析精度和稳健的人机对话性能。其智能人机对话技术能够提供基于不完整或不准确的语音识别结果进行智能语义推理，通过针对特定领域特定应用需求、对话目标、对话行为、对话状态和对话上下文的统计建模，解决传统系统设计中规则无法覆盖实际对话状态的问题，同时增强系统对于错误识别结果和错误推理结果的自适应性，大幅度提升用户的语音交互体验。

思必驰语音识别技术解决了大量的词典和语言模型的识别问题，采用了鲁棒特征提取的方法，能够有效降低噪声环境下的错误率，并用动态噪声自适应来使统一的模型适用于各种噪声环境，支持全国各地方不同口音普通话识别，在用户实际使用环境下，引擎的识别率能保持较高水准。

7.2.3 北京智齿博创科技有限公司

7.2.3.1 企业简介

智齿科技成立于 2014 年 5 月，是一家提供智能客服服务的互联网创业企业。智齿科技将自然语言理解、机器学习及大数据技术进行有效整合，构建了包括云呼叫中心、智能机器人客服、人工在线客服、工单系统、智齿 360、智能外呼机器人在内的智能全客服平台和智能外呼平台，为企业用户提供多样化的智能客服服务，为用户解决移动时代、体验经济环境下传统客服所不能解决的企业客服痛点。智齿科技于 2018 年 4 月获得由云启资本领投，耀途资本、上海原龙投资和博雅盛景等机构跟投的 1.5 亿元人民币 B+轮融资。

7.2.3.2 产品研发

智齿科技推出云呼叫中心、客服机器人、在线客服、工单系统、外呼机器人等 SaaS 客服产品，为金融、教育、电商、企服、生活消费、互娱、政企等领域的客户提供一站式客服

解决方案，同时构建了多元丰富的统计分析报表，帮助企业用户分析顾客关注焦点、营销活动结果验收、客服人工工作量及效率，为客服主管提供数据支撑，辅助企业科学经营决策。

智齿客服产品支持桌面网站、移动网站、微信、微博、APP 等接入方式，用户只需把代码复制到网站上即可将多平台顾客汇集于同一平台管理，操作简便。

7.2.3.3 竞争优势

智齿科技的智能客服业务行业服务经验丰富。智齿科技在 26 个行业构建知识图谱，其中 16 个行业构建了 40 个细分场景知识图谱，提供行业知识库及行业客服机器人问答，问答语料库内容覆盖面广，能为众多领域行业用户提供智能客服问答服务，目前智齿科技服务过的企业数量超 5 万家，覆盖电商、互联网金融、生活服务、企业服务、在线教育、互动娱乐等多个细分行业，包括用友、富士康、滴滴出行、趣分期、新东方、搜狐、销售易等知名企业用户。

智齿客服率先将自然语言处理技术及机器学习引入客服产品，实现精准理解用户问题并匹配最佳答案，反馈答案准确率高达 98%。智齿客服支持的机器人客服与人工客服自由切换模式，可解决 80%人工客服的同质化回复问题，同时机器人客服可自动组织标准答案，帮助客服人员大幅提升客服回复质量和效率。

头豹研究院简介

- 头豹研究院是中国大陆地区首家 B2B 模式人工智能技术的互联网商业咨询平台，已形成集行业研究、政企咨询、产业规划、会展会议行业服务等业务为一体的一站式行业服务体系，整合多方资源，致力于为用户提供最专业、最完整、最省时的行业和企业数据库服务，帮助用户实现知识共建，产权共享
- 公司致力于以优质商业资源共享为基础，利用大数据、区块链和人工智能等技术，围绕产业焦点、热点问题，基于丰富案例和海量数据，通过开放合作的研究平台，汇集各界智慧，推动产业健康、有序、可持续发展



四大核心服务：

企业服务

为企业提供**定制化报告**服务、**管理咨询**、**战略**调整等服务

云研究院服务

提供行业分析师**外派驻场**服务，平台数据库、报告库及内部研究团队提供技术支持服务

行业排名、展会宣传

行业峰会策划、**奖项**评选、行业**白皮书**等服务

园区规划、产业规划

地方**产业规划**，**园区**企业孵化服务

报告阅读渠道

头豹科技创新网 —— www.leadleo.com PC端阅读全行业、千本研报



头豹小程序 —— 微信小程序搜索“头豹”、手机扫右侧二维码阅读研报


图说


表说


专家说


数说

详情请咨询

客服电话

400-072-5588

上海

王先生： 13611634866
李女士： 13061967127

南京

杨先生： 13120628075
唐先生： 18014813521

深圳

李先生： 18916233114
李女士： 18049912451