大数据与决策研究

（政策与技术跟踪专题）

2022 年第 30 期（总第 140 期）

广西壮族自治区信息中心
广西壮族自治区大数据研究院　　2022 年 8 月 24 日

编者按：近年来，随着自然语言处理技术的迅速发展，出现了一批基于自然语言处理技术的应用系统。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。本期将介绍自然语言处理技术相关情况。

本期要目

广西壮族自治区信息中心
广西壮族自治区大数据研究院

自然语言处理的定义与环节

一、自然语言处理的定义

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。

二、自然语言处理环节

（一）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

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（二）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。

（三）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。

（四）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使用者的表达用意。

（来源:《2019年中国自然语言处理行业研究报告》）

广西壮族自治区信息中心
广西壮族自治区大数据研究院

我国自然语言处理产业链构成

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。

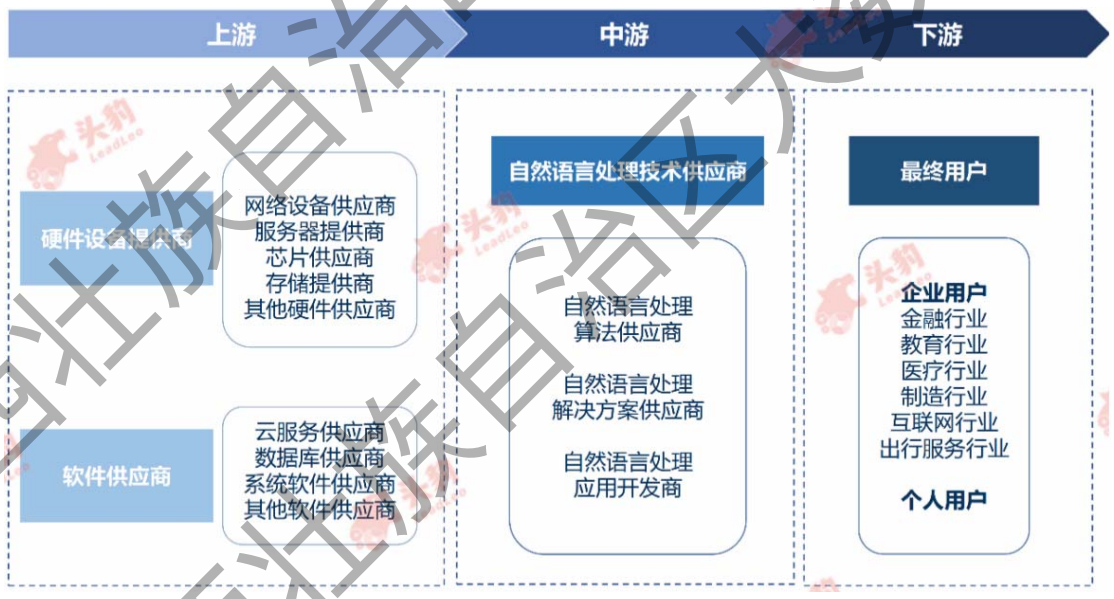

图 1　我国自然语言处理产业链

一、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、

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（一）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。

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（二）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

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（三）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、亚马逊美食评论集、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。

二、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、

广西壮族自治区信息中心 广西壮族自治区大数据研究院

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。

三、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。

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（来源:《2019年中国自然语言处理行业研究报告》）

广西壮族自治区信息中心
广西壮族自治区大数据研究院

自然语言处理的七个典型应用技术

一、机器翻译

机器翻译是指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。机器翻译一般通过以下三种方法实现：一是基于理性的研究方法—基于规则的方法；二是基于经验的研究方法—基于统计的方法；三是与深度学习相结合。

此外，机器翻译的应用场景主要分为五类：（1）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；（2）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；（3）医疗创业公司利用计算机阅览 X 光照片、MRI 和 CT 照片；（4）对机器人、无人机以及无人驾驶汽车的改进至关重要；（5）VR 翻译等。

二、信息检索

信息检索即从相关文档集合中查找用户所需信息的过程。信息检索的工作原理分别为：

- “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；

广西壮族自治区信息中心
广西壮族自治区大数据研究院

. 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。

三、情感分析

情感分析是指通过计算机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。这项技术主要有以下应用场景：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。

四、自动问答

自动问答即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。其工作流程首先要正确理解用户所提出的问题，其次是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，最后是将获取的答案反馈给用户。此外，自动问答技术应用有以下三类：

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。

五、自动文摘

自动文摘是指运用计算机技术，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

广西壮族自治区信息中心
广西壮族自治区大数据研究院

成一个精简版本。该项应用技术具备压缩性、内容完整性以及可读性等特点，且存在两种技术路线：

·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。

六、社会计算

社会计算的定义为在互联网的环境下，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。社会计算的主要应用场景为：

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。

七、信息抽取

信息抽取是指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。其工作原理是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。

广西壮族自治区信息中心
广西壮族自治区大数据研究院

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。

（来源：《自然语言处理研究报告》）

广西壮族自治区信息中心
广西壮族自治区大数据研究院

广西壮族自治区信息中心
广西壮族自治区大数据研究院

编辑部地址：南宁市体强路 18 号广西信息中心 1412 号房
联系电话：0771-6113592
电子邮箱：dsjyjs@gxi. gov. cn
网　　址：http：//gxxxzx. gxzf. gov. cn/

扫描二维码获取
更多决策参考信息