gitbook.cn

GitChat - 《自然语言处理研究报告》的文章

21 ～ 26 分

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。

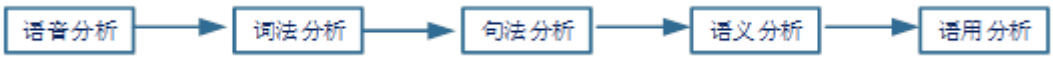


图 1 自然语言理解层次

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。

词法分析的目的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。

语义分析的目的是找出词义、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。

语用分析则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；

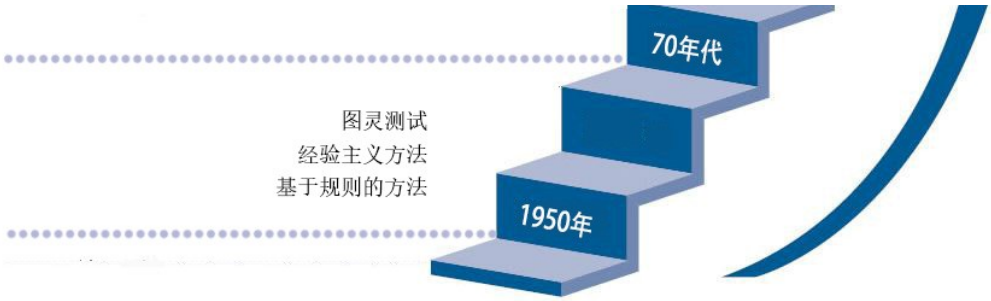
第二，文摘生成，机器有能力生成输入文本的摘要；

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的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到了实质性的突破，从实验室走向实际应用。

从2008年到现在，在不到十年的时间里，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 NLP 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的支持，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 NLP 顶级会议，例如 ACL，COLING 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。

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1.4亿人地用户，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，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。该团队很快就投入对其第一个项目——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成自主开发的了。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9倍于以往循环神经网络的速度实现了目前最高的准确率。

  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（Association

for Computational Linguistics）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、DuerOS 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。