中文信息处理发展报告

（2021）

中国中文信息学会

中国·北京

2021.12

前　言

《中文信息处理发展报告》（2021）是中国中文信息学会召集领域专家对中文信息处理学科方向和前沿技术的阶段性梳理。本发展报告的定位是深度科普，旨在向政府、企业、媒体等对中文信息处理感兴趣的社会各界人士简要介绍相关领域的基本概念和应用方向，向高等院校、科研院所和高新技术企业中从事相关工作的专业人士介绍相关领域的前沿技术和发展趋势。

《中文信息处理发展报告》（2021）继续沿用《中文信息处理发展报告》（2016）的编撰思路：对近年来本专业领域内的学科方向进行系统总结梳理，对未来一段时期的前沿技术趋势进行展望。按照各个专业委员会发展历程，结构安排上分为汉字字形信息、速记、计算语言学、少数民族语言文字信息处理、机器翻译、信息检索技术、语音信息技术、社会媒体处理、知识图谱领域、医疗健康与生物信息、网络空间大搜索技术、隐私计算、开源情报技术、自然语言生成与智能写作、情感计算等 15 个专业领域分别进行表述。各个专业领域统一从研究背景与意义、领域发展现状与关键科学问题、领域关键技术进展及趋势、领域产业发展现状及趋势、总结及展望等 5 个部分进行总结梳理和趋势展望。因此，本发展报告既可作为中文信息处理领域的总体发展研究报告使用，亦可作为每个专业领域独立的发展研究报告单独使用。

本发展报告的每个专业领域部分由各个专业技术委员会组织本专业领域内专家和学术团队协同撰写完成，由学会秘书处组织相关专家负责对初稿反馈意见，最后校核、编排、统一成文。

参与本发展报告撰写工作的主要专家如下：

汉字字形信息：张建国等。

速记：廖清等。

计算语言学：车万翔等。

少数民族语言文字信息处理：吐尔根·依布拉音等。

机器翻译：张家俊、黄书剑、李军辉、王瑞、何中军、苏劲松、冯冲、肖桐、史晓东、余正涛、张民等。

信息检索技术：窦志成、范意兴、郭嘉丰、何向南、黄民烈、刘畅、刘奕群、毛佳昕、任昭春、徐君、严睿、殷大伟、张帆、张鹏等。

语音信号技术：郑方、贾珈、王东、徐明星、吴志勇、周强、程星亮等。

社会媒体处理：刘挺、唐杰、林鸿飞、黄萱菁、沈华伟、冯仕政、陈慧敏、刘知远、

丁效、李斌阳、万怀宇、魏忠钰、秦兵、王素格、刘康、夏睿、蔡毅、黄民烈、沈浩、张伦、朱旭琪、孟天广、谢幸、杨洋、杨成、何婷婷、付瑞吉、王明文、彭敏、徐睿峰、邱伟云、左家莉、伍大勇、张洪忠、张伟男、张华平、王彦皓、蔡佳豪、赵鑫、王啸等。

知识图谱领域：陈华钧 、程龚 、韩先培 、侯磊 、胡伟、李涓子、李炜卓、刘康 、刘铭、漆桂林 、秦兵、王昊奋、许斌 、张文、赵军等。

医疗信息处理技术：陈清财、汤步洲、户保田、陈俊杰、闫峻等。

网络空间大搜索技术：贾焰、李爱平、王晔、仇晶等。

隐私计算：李凤华、李晖、邱卫东、牛犇、邹德清等。

开源情报技术：刘科伟、殷复莲、黄永峰、张震、杨震、杨忠良、马谅、文盖雄、夏睿、丁效、齐中祥、管磊、于锐、韩先培等。

自然语言生成与智能写作：黄民烈、万小军、高扬、冯骁骋、严睿、宋睿华、段楠、赵铁军、饶高琦、杨沐昀、肖欣延、吴华、李国东、李丕绩等。

情感计算：秦兵 、徐睿峰、朱廷劭、夏睿 、刘斌 、赵妍妍 、李斌阳等。

由于时间仓促，加之篇幅和视角所限，难免挂一漏万，仅供有志于中文信息处理事业的同仁和青年学者们参考研判，并期待让我们携手同行，再创中文信息处理事业的新辉煌！

中国中文信息学会

2021 年 12 月

目　　录

第一章　汉字字形信息研究进展、现状及趋势

1.1.研究背景与意义

汉字是中华文化的基因和核心，是中华文明得以传承和发展的载体，而计算机中文字体是汉字书写文明在信息化时代的全新表现形式，也是中文信息处理的基础。

在汉字发展历程中，随着不同时代对信息传播的需求变化以及工具的变革，汉字字形也在不断发生变化，并逐渐被赋予了审美功能。从篆、隶、草、楷、行等古老书体类别，到雕版印刷时期的宋体、受西方影响而产生的黑体，再到民国时期的仿宋体以及丰富多样的美术字，都是汉字字形变化的成果。

中文字体最早进入电脑，始于汉字照排系统的研制。1974 年 8 月，国家重点科技攻关项目“汉字信息处理系统工程”（简称“748 工程”）设立，北大教授王选带领科研团队研制出了汉字激光照排系统，并发明了针对汉字的高倍率字形信息压缩技术和高速还原技术，这些成果使汉字排版印刷告别了“铅与火”的历史，开启了汉字进入数字化时代的新篇章。

如今，计算机中文字体已成为中国人进行信息沟通、情感传递、文化表达的必要载体，被广泛应用于出版、印刷、包装、广告、教育、办公、游戏动漫、互联网、移动终端等社会生活的各个领域。不同领域的用户特性、传播媒介的属性，都对汉字字形的创新提出了需求和挑战，从事汉字字形设计与研究的专业队伍不断壮大，中文字体种类日渐丰富；与此同时，在媒体的传播与推动下，热爱汉字、关注汉字字形的群体不断扩大，汉字字形信息已经从一个小众领域逐渐走向大众视野。

1.2.领域发展现状与关键科学问题

中文字库是艺术和技术的完美结合，其中每个汉字是设计师或书写者一笔一画设计或书写出来的，在传情达意的同时，表达视觉审美含义，同时字库内部也包含控制字形还原的代码，中文字库兼具美术作品属性和软件属性。

1.2.1.　字体美术作品著作权得到认可

中国中文信息学会一直多方呼吁加强字库知识产权的保护，2011 年学会发起了“弘扬中华文化，保护计算机中文字体”的倡议， 2012 年、2014 年学会先后给国家相关部

门发函，呼吁对字体加以保护。近年来，在中文字体领域，法律界已经普遍认为：具有艺术美感的独创性的单字，构成了著作权法规定的美术类作品，应当受到法律保护。2014年4月22日，最高人民法院公布了《2013年中国法院50件典型知识产权案例》，将字体单字著作权确权案件列入其中，具有独创性的单字享有美术作品著作权得到了最高人民法院认可。

在各方共同努力下，近年来，法院通过判决、调解、和解等形式支持字体著作权保护的案例越来越多。

1.2.2. 中文字符集标准和汉字字形规范的形成

1.2.2.1. 中文字符集标准

计算机要准确处理各种字符集合，就需要进行字符编码，以便计算机能够识别和存储各种文字。字库中包含的字符是依据不同字符集标准进行收纳。常见的中文字符集标准有 GB2312-80《信息交换用汉字编码字符集 基本集》、GBK《汉字内码扩展规范》、GB18030-2000《信息技术 汉字编码字符集 基本集的扩充》、GB18030-2005《信息技术 中文编码字符集》以及国际标准ISO/IEC 10646等。

1980年，国家标准GB2312-80《信息交换用汉字编码字符集 基本集》发布，由6763个常用汉字和682个全角的非汉字字符组成。

1995年，国家技术监督局为中文Windows 95制定了《汉字内码扩展规范》GBK，共收录包含21003个汉字，涵盖了常用的简体和繁体汉字。

2000年，信息产业部和国家质量技术监督局联合发布GB18030-2000编码标准，全名是《信息技术 信息交换用汉字编码字符集 基本集的扩充》。GB18030-2000规定了常用非汉字符号和27533个汉字。 GB18030-2000是GBK的取代版本，它的主要特点是在GBK基础上增加了CJK统一汉字扩充A的汉字。GB18030-2000是全文强制性标准，市场上销售的产品必须符合。

2005年，GB18030-2005《信息技术 中文编码字符集》发布。在GB18030-2000的基础上增加了CJK统一汉字扩充B的汉字，共42711个，并增加了多种我国少数民族文字的编码。其中GB18030-2000部分为强制性标准。

2010年，GB13000-2010《信息技术 通用多八位编码字符集（UCS）第一部分：体系结构与基本多文种平面》标准发布，该标准等同采用国际标准ISO/IEC 10646：2003，共收录汉字71427个。

最新的国际标准ISO/IEC 10646：2020，收录汉字达93888个，在ISO/IEC 10646：

2003 的基础上增加了 CJK 统一汉字扩充 C、D、E、F、G。

1.2.2.2. 汉字字形规范

为了贯彻《中华人民共和国国家通用语言文字法》，提升国家通用语言文字的规范化、标准化、信息化水平，满足信息时代语言生活和社会发展的需要， 2013 年国务院公布了由教育部、国家语言文字工作委员会组织制定的《通用规范汉字表》。《通用规范汉字表》共收录汉字 8105 个，其中 163 个汉字在 国家标准 GB18030-2005 以外（CJK 统一汉字追加 3 字、扩 C 区 44 字、扩 D 区 8 字、扩 E 区 108 字）。

当前市场上不规范不统一的字形时常显现，如简体字库制作成繁体笔形，少笔画、多笔画等不同程度的字形规范问题。《通用规范汉字表》规定了汉字的写法，是中文字体设计的依据。

1.2.3. 关键科学问题

中文字库是艺术设计与计算机科学相结合的产物,和其他软件一样需要关注开发、分发以及终端应用三个主要阶段。在数字信息时代的今天，稳定高效的开发流程、覆盖广阔的分发以及丰富简明的应用，都是推进我国汉字字形信息行业发展的重要因素。

1.2.3.1. 字库开发

字体设计生产是字体开发的重要环节，具体的是将各类字体设计稿、书写字稿，转换成相应的矢量轮廓数据，然后按照 OpenType 规范组织数据（字形和编码），最终生成标准的 TTF/OTF 字库，以满足各种应用软件和操作系统对字体的使用需求。一方面，行业企业自主开发了各自的字体设计工具和平台，基于汉字部件快速检索技术和网络协同技术，为多人协同开发字体开发提供有力的支撑，保证了大字符量中文字体的开发速度。另一方面，随着人工智能技术不断成熟，通过不断的探索研究，目前已经在字体辅助设计方面取得成果，提升了字体开发效率。

1.2.3.2. 字库分发

在一个信息高速发展的时代，人们的生活逐渐向云端转移，中文字库的分发形式也发生了变化。在网络传输环境大幅提升的当下，用户越来越多地使用云服务来获取字体。当前通过云服务来获取字体的方式主要可以分为以下两类：一类是本地使用字体的云服

务，主要服务于需要使用字体作为素材、在各类图形、文本编辑软件中进行本地编辑创作的用户，他们会通过各个公司推出的电脑客户端（如汉仪字库的字由，方正字库的字加等），通过联网获取自己订阅账户中的字体，将其同步至个人设备中进行使用；另一类字体的云服务，需要使用云字库（WebFont）技术，其本质特征是将字体存储于服务器云端，用户在实际需要显示这些字体时向云端即时提出请求。

1.2.3.3. 字体应用

随着移动互联网的发展，为了适应字体在各种场景中的应用，字体厂商研发出了可变字库技术、压缩字库技术、云字库（WebFont）技术以及特效字体技术。可变字库技术是 OpenType1.8 规范的最新字库技术，它允许单个字体文件同时支持多个字体形态,以满足用户场景下不同形态字体的使用需求；通过压缩字库技术，可以随意在受限的移动嵌入式设备中使用大容量中文字体；特效字体技术实现了字形的彩色和动态效果，为字体能够在互动娱乐产品场景中的使用提供基础。云字库（WebFont）技术主要是针对字体的网络使用场景，特别是中文的显示，该技术可以根据网页的显示需求快速动态生成所需的字库数据，并及时在浏览器页面展示。多种新应用技术的不断实现，丰富了字体的应用场景，为字体行业的发展提供技术保障。

1.3.领域关键技术进展及趋势

中文字库兼具美术作品属性和软件属性，除了设计创新，技术实力也是字库企业的重要保障。随着计算机运算能力的提升和人工智能技术的成熟，不少行业企业、院校关注字库相关技术的研发和应用。

1.3.1. 中文字库设计软件

中文字体的设计软件主要包含开放的商业字体设计软件，以及行业企业自主研发、内部使用的中文字体设计软件。业内广泛使用的商业字体软件有 Fontlab、Glyphs，也有设计师使用 Adobe Illustrator 设计字体等，这类软件大多来自国外，开发之初并没有考虑中文字体设计的特点，缺失中文字体开发所需的很多核心功能。西文字体字符量比较少，轮廓相对简单，一般由一个人就可以完成一套字体的开发，但是中文字体字符量较大，最少的也要设计 6000 多汉字，GKB 字符集字库更需要制作 2 万多字。行业企业自

主开发的字体设计软件主要解决了提高庞大中文字符集字形开发的工作效率。

为了满足中文字体开发字体设计软件具有如下功能：

1）贝塞尔曲线和直线的绘制、精调、平滑处理等功能

2）字形轮廓质量检查及自动纠正

3）曲线降阶处理

4）组件快速检索并复用

5）家族化字体生成

6）支持多人协同工作

除上述功能外，行业企业自主开发的软件，还可以根据各自项目需求，快速添加或调整软件功能，以满足自己的个性化需求。大多数行业企业也在生产流程中引入开放的商业字体设计软件，以满足外文字符设计的需要。

1.3.2. AI 辅助字体设计技术

以深度神经网络为代表的的深度学习在计算机视觉、自然语言处理等领域取得了巨大的成功，人工智能技术应用在中文字库领域的主要目标是降低字体设计师的重复劳动，提升中文字体的生产效率。

2011 年由上海印刷技术研究所联合同济大学共同开展了“汉字字库计算机智能制作系统”项目的研发。利用汉字构件及字形数据建立汉字构件字形库，并在此基础上提供汉字自动组合与编辑功能，在规范化和自动化方面为汉字字库的研制开发，提供一个更有效、更合理、更容易控制品质的应用平台。其中“基于神经网络的汉字构字方法”和“字形智能化评价模型与修正方法”是项目的关键技术。该项目经过 2 年多的研发，于 2014 年通过了上海市科委的验收。

2016 年方正手迹公司采用北京大学王选计算机研究所的人工智能辅助字体生成技术，推出手迹造字 APP，只要手写 100 个汉字，约半小时，即可产生包含 6763 汉字的完整个人字库，目前也在华为主题商店、WPS 提供个人造字服务，产生的个人字库，分别可以在华为手机、WPS 文档中使用。2020 年 vivo 手机、2021 年百度输入法也分别推出面向普通个人的个人造字服务。

2018 年阿里巴巴和汉仪合作，推出阿里汉仪智能黑体，由阿里计算平台事业部 PAI 产品线 Deep Learning 团队、阿里人机自然交互实验室以及汉仪字库设计师协作完成。这款字体的生成是人机协同工作的成果，机器学习，人工干预，循环往复，直到最终生成达标字库。

方正、汉仪等行业企业也开始人工智能在精品字库设计方面的研究探索，提升设计效率，已经可以将一套 300—500 字左右的手稿，自动拓展至 GB2312 编码中的全部 6,763 个汉字的字形，由于质量要求高，还需要设计师精修、调整，以便形成面向企业服务的精品字库。

人工智能技术在字体行业的应用是未来不容忽视的发展趋势，一方面极大的提升字体开发效率，另一方面解放了字体设计师的生产力，使未来的字体设计师，能将重心归于创意，更加专注于灵感构思和创意挖掘等更为重要环节。

1.3.3. 可变字体技术

可变字体（Variable Fonts）技术源于 Adobe、Apple、Google、Microsoft 四巨头于 2016 年发布的 OpenType v1.8 字体格式规范，该技术是在已有的 OpenType 字库基础上增加可变特性数据表，数据格式有 OTF 和 TTF 两种。OpenType v1.8 字体格式规范的发布，将允许单个字体文件同时支持多个字体形态，它可以将几个字体紧凑地封装在单个字体文件中，通过定义字体内的变化来实现单轴或者多轴设计空间。目前主流的操作系统、浏览器、设计软件大多在不同程度上支持可变字体。

在网页设计上，只需使用一款可变字体，就可以为网页中各层级的标题以及正文设置不同的字体样式。这不仅能够加快页面的加载速度，也让页面整体排版可以适应不同屏幕尺寸的变化，满足多屏时代的设计需求，提供更好的阅读体验。而在平面设计上，可变字体响应了当下和未来的动态设计需求。它可以应对足够复杂的平面空间和应用场景。

之前大多可变字体以西文为主。中文字库的字符集庞大、字形复杂，实现字体无级可变的难度更高。近年来各个字体厂商研发出越来越多的中文可变字体。

文鼎字库在 2017 年推出了全球首款中文可变字体——“文鼎晶熙黑”，拥有字重、字宽两种可变轴。同年 11 月方正推出全球首款中文三轴可变字体——“方正悠黑”，具有字重、字宽、字高的可变三轴字体设计空间；在 2019 年，方正还为小米品牌打造了“小米兰亭 Pro”可变字体，支持字体粗细的无级调节。汉仪字库在 2021 年 6 月发布了为华为品牌定制的可变字体 HarmonyOS Sans，它是一款多语言的无级可变字体，支持简繁中文、拉丁、西里尔、希腊、阿拉伯等书写系统。

随着人们逐渐了解可变字体，以及更多适配硬件和软件的出现，中文变字体将会有更多的应用形式，在实用功能与视觉设计上，带给我们更多的惊喜。

1.3.4. 压缩字库技术

嵌入式设备，如手机，导航仪，电子书阅读器，要求字体具有体积小，反应速度快，美观多样等特点。但是由于汉字的数量非常多，导致汉字字库的数据量很大，一般包含GB13000基本平面2万7千多汉字的字库，如Windows XP系统中的宋体字库数据量是10M，无法满足嵌入式设备屏幕显示用字的需求。

手机QQ、QQ空间等 移动交互平台希望用户自己可以看到具有特殊效果的字体，与他人聊天时，其他用户无需单独下载安装字库，即时看到其使用的具有特殊效果的字体。

为了解决以上问题，需要研发体量小的压缩字库，同时也可以满足手机QQ希望通过网络快速传输字库的需求。

目前常见的压缩字库有两种，一种是自有格式的压缩字库，一种是标准TrueType格式的压缩字库。自有格式压缩字库压缩率更高，但是依赖自有字体解释引擎，接入应用系统相对复杂；标准TrueType压缩字库格式使用系统自带的字体解释引擎即可，不需要加载额外的引擎，使用标准的接口就可以实现调用，通用性更好。

2015年汉仪研发的FullType 超小字库，其存储容量只有传统TrueType格式的1/5到1/10左右，在手机QQ、QQ 空间上线，显著减少了对手机存储空间的占用，提升了字体的加载速度，优化了用户体验；2016年方正研发的基于标准TrueType格式的压缩字库，存储容量约为标准字库的20%-30%，也在手机 QQ、QQ 空间上线提供字库服务。

1.3.5. 字体特效引擎技术

字体特效引擎是通过字体引擎渲染绘制可以把常规静态文字转化为彩色动态文字，实现字体的多元化应用，在手机QQ平台，2015年汉仪研发了基于Fulltype字库的字体特效渲染引擎，2016年方正推出基于标准的TrueType压缩字库的字体特效引擎。

目前手机QQ提供的互娱式字体主要包括黑白字体、彩色字体、炫动字体、嗨爆字体、艺人手写字体和文娱IP字体等。其中，彩色字体、炫动字体和 嗨爆字体的生成涉及到二维图形渲染画刷引擎技术、矢量彩色字体技术、彩色位图字体自动生成技术，有效解决在移动社交应用场景中，字体千人一面的问题，满足了用户的个性化聊天社交需求。

1.3.6. 云字库（WebFont）技术

随着网站的设计趋于个性化，字体作为网页中最主要的元素，不同风格的字体对于网站的展示愈加重要。网络字体是 CSS3 中的一个模块，主要是把定义的特殊的字体嵌入到网页中，免安装、免下载、在线使用。常用的网络字体格式有 woff、woff2、eot 等，不同格式网络字体适配不同的浏览器。

方正、汉仪等字体厂商都开发并实现了中文字体云字库技术，通过按需截取和高效压缩等技术有效地控制了字体文件的大小，使之和英文字体文件大小相当，提高页面的加载效率，降低对网络带宽的占用，可以兼容市场上大部分浏览器。

云字库技术的实现大大的推动了字体在互联网场景下的应用,云字库优势主要体现在以下几个方面：

极速推送字体： WebFont 极速推送网络字体，使网络字体瞬间加载，速度与效果兼得。

流量分压：为用户量身定制的小字库文件，将被托管在 WebFont 平台上，当用户的页面被浏览时，文件直接从 WebFont 平台推送到客户端终端浏览器，节约服务器流量。

优化搜索引擎排名：虽然图片也可以呈现中文字体，但是 Google、百度等搜索引擎无法辨认出图片的文字内容，无法搜索到网站相关内容。 使用网络字体，则是呈现真实的文字，无论是标题、内容都适合引用。

无级缩放不模糊：图片在放大和缩小的过程中会产生变形或马赛克，网络字体采用的是矢量字体，支持无级缩放，不管放多大或缩再小都不会产生变形或模糊，给用户一致的体验。

改善使用体验：图片在高分辨率的视网膜屏幕中，常遇到分辨率不足的状况；网页字体则以矢量字在网页中真实呈现，根据浏览分辨率做实时的字体描绘，无论放大到任何尺寸都能清晰分明。

1.3.7. AI 字体识别技术

文字识别一直是文档分析中的重要环节，互联网的迅猛发展极大地推动了新字体的传播，字体种类的迅速增长带来了字体识别的新需求。

与常规的光学字符识别不同，字体识别的关键是要区分出不同字体之间的形态差异。字体风格的差异体现在字符形态的多个方面，如部件的空间分布、疏密程度、中宫的聚集程度，以及笔画的粗细、曲直、光滑度，还有笔锋的变化、交叉点的处理等。

目前方正、汉仪等字库公司在字体检测、字体识别等方面都取得了进展。现有的字体识别方法针对常用字体取得不错的识别效果，已经应用在各种消费和商用场景，包括对各种场景下的字体识别、字体风格提取等。

1.3.8. 字库分发技术

在 5G 网络即将走向普及的高速信息时代，中文字体的分发形式也产生了一定变化。无论是之前通过光盘等实体介质发送给用户，还是现在通过邮件、主页、微信公众号自助服务将字体发送给用户，最终的目的都是让用户可以直接获取字体文件在所需设备上进行本地安装。而在网络传输环境大幅提升的当下，用户越来越多的通过联网获取字体，将其同步至个人设备中进行使用。

目前字体市场应用比较广泛的字库分发软件主要是汉仪的“字由”和方正的“字加”。这两大分发软件旨在为独立设计师、广告宣传类企业和其他有用字需求的企业建立一个字体方面相互交流、融通、应用和创新的开放性平台，将传统的字体下载、安装、预览以及在设计软件中的使用字体等功能全部集成到应用中，为众多用户解决“找字体”和“换字体”两大核心问题，提高使用者的工作效率。

1.4.领域产业发展现状及趋势

2020 年 10 月，上海印刷字体展示馆揭牌仪式在现代汉字印刷字体发源地——上海印刷技术研究所隆重举行。新中国成立后，上海印刷技术研究所率先在此从事汉字印刷字体科研攻关，开发了当下广泛使用的宋体、黑体、仿宋体、楷体。2009 年，“汉字印刷字体书写技艺”被列入上海市非物质文化遗产名录。上海印刷字体展示馆面向公众免费开放，让更多人了解现代汉字印刷字体的起源和发展，了解汉字印刷字体的设计规范和创写工艺，更好地宣扬与传承汉字文化、字体文化、非遗文化。

上世纪九十年代，是中国字体行业发展辉煌的时期，全国有十几家从事字库开发的企业，字库数量增长迅猛。进入 2000 年后，由于盗版的日益猖獗，社会版权意识的淡薄，中文字体行业发展遭遇空前困境，稍具规模的字体设计研发企业数量锐减，中国大陆仅剩方正、汉仪、华文、华光与中易，且大多处于勉强维持的状态。

近年来，随着国家对知识产权相关政策法规的不断完善，以及媒体对版权知识的科普与宣传，大众的版权意识显著提升。在字体厂商、设计师群体、设计院校、行业协会

等各方的共同努力和推动下，中文字体产业逐渐呈现健康发展态势。

1.4.1. 中文字体产业发展现状

1.4.1.1. 中文字体行业百花齐放

字体行业在中国消沉了十多年后，中文字体整体的质量和数量都落后于西方、日本和韩国。值得欣慰的是，近几年，设计界掀起了一股字体的热潮，越来越多的设计师意识到了字体对于平面设计的重要性，与此同时，字体设计软件种类越来越多且越来越人性化，设计师坐在电脑前，无需纸笔，就能设计出可以在电脑上使用的字库，大大降低了行业门槛。

设计界的关注和认可带动了整个中文字体行业的发展，目前，国内大大小小的字体厂商、个人字体工作室已有 100 多家，呈现出百花齐放的景象，整个行业处于上升期。

1.4.1.2. 中文字体种类丰富多彩

行业蓬勃发展的直接成果就是中文字体种类的丰富多彩。在创新方面，国内字库公司主要是通过自主创作、外部合作的方式，来不断提高字体产品的创新能力。

外部合作通常是指和社会各界人士的字体合作，包括个人字体设计师、平面设计师、书法家、漫画家、汉字发烧友，甚至民间艺人、影视明星，通过合作的形式将更多优秀创意转化成字体产品。近年来，外部合作的范畴进一步拓宽，多家字体厂商，如方正、上海锐线等，开始尝试与国外字体厂商、设计师合作，将优秀的日文字体开发成中文版，来丰富中文字体的选择。

目前，国内不同厂商推出的中文字体已多达数千款，并且字体的质量也越来越高，涵盖排版正文字、创意美术字、个性手书、传统书法字等不同的风格种类，能够满足社会各领域的不同应用需求。

1.4.1.3. 国家对中文字体行业日益重视

中文字体行业的健康发展离不开国家层面的关注和重视。近年来，国家不断加大对中文字体行业的重视力度，一方面不断完善字库知识产权相关政策法规，另一方面，将中文字库相关内容纳入国家文化发展规划，先后启动了中华字库工程、中华精品字库工程两大重要文化工程。

“中华字库”工程是一项引领中华文化步入信息化、数字化时代的先导性、奠基性工程，目的是要建立全部汉字及少数民族文字的编码和主要字体字符库，工程于 2011 年正式启动。工程共分为 28 包，有近 30 家高校、研究院所和企业参与了工程研发工作，其中包括多家字库企业。中易字库参与承担了第 2 包“数据采集平台研发”、第 22 包“输入法研发”，方正字库承担了第 17 包“当代人名地名用字搜集与整理”、第 20 包“字库制作一：中间字库、宋体楷体等成果字库”的研发，华光字库参与承担了第 18 包“少数民族古文字的搜集与整理”、第 19 包“少数民族现行文字的搜集整理与字库制作”，汉仪字库承担了第 21 包“字库制作二黑体仿宋体及古汉字成果字库”的研发。这些成果将满足中华各民族古今各类文献的出版印刷、数字化处理和传输的需要，全面打通信息化的发展瓶颈，使中华各民族文字的使用，中华文明的普及与传播，更加方便和高效。

“中华精品字库工程”是中华优秀传统文化传承发展工程支持项目，工程由中国文学艺术界联合会、国家语言文字工作委员会共同指导，将精选 100 位中国历代书法名家的代表作品，开发成电脑字库，中国书法家协会负责开发字体的遴选和质量审核，北京北大方正电子有限公司负责字库的开发工作。工程于 2017 年申请立项，截止到 2021 年 6 月底，已对外发布了 35 款精品字库。“中华精品字库工程”是书法艺术和信息技术、汉字应用的高度融合，是推动中华优秀传统文化传承与发展的重要举措，对传承中华文化基因、弘扬中国精神、传播中国价值，都有着重要的作用与意义。工程成果可以满足日益发展的互联网媒体和社会大众多样化汉字字形需要，功在当代，利在千秋。

1.4.1.4. 字体设计相关研究成果不断推出

在字体厂商们不断推出汉字字形创新成果的同时，设计院校的学者们也积极开展汉字字形方向的学术研究，并将研究成果转化成了专业书籍。

2018 年，由中央美术学院设计学院副教授周博博士撰写的《中国现代文字设计图史》，对从晚清、民国到今天一百多年的时间里中国现代文字设计成就做了一番比较全面的梳理，清晰、明了的讲述了中国现代文字的设计历史。

2019 年，湖北美术学院副教授李海平推出的《汉字字形学新论》一书，立足先辈们的累累硕果，结合文字学、书法学和设计学，尝试从一个新的角度探讨汉字字形造字技法，演变过程及相关的影响因素。

2021 年 8 月，澳门理工学院艺术高等学校副教授孙明远撰写的《中国近现代平面设计和文字设计发展历程研究——从 1805 年至 1949 年》付梓，该书以近现代印刷技术在中国的发展为主线，多角度清晰描绘这一时期中国平面设计和文字设计的历史发展进程。

2021 年 10 月，清华大学美术学院陈楠教授的新书《中国汉字设计史》出版上市，

该书以汉字设计传承与创新的发展脉络开篇，结合设计学、传播学和美学的宏观视角，通过研究与分析，挖掘潜藏于汉字艺术审美与信息传播功能背后的思维与方法。

中央美术学院副教授、国际文字设计协会（ATypI）中国国家代表刘钊一直关注中外字体设计交流，组织中国专家在国际文字设计协会 Atypl 论坛演讲，对外传播汉字文化。她统筹引进拉丁文字设计丛书，丛书由中央美术学院、雷丁大学、国际文字设计协会联合推荐，《文本造型》《如何创作字体》《字腔字冲：16 世纪铸字到现代字体设计》分别于 2018 年 5 月、2019 年 2 月、于 2021 年 6 月出版上市。

这些学术研究成果的推出，为汉字字形领域的教育和传播提供了有效工具，也为中文字体行业的创新发展提供了理论支撑。

1.4.1.5. 产学研实现良性循环

近年来，众多字体企业与设计院校之间积极开展产学研合作。一方面，设计高校积极聘请字体企业中有丰富实践经验和扎实理论水平的资深字体设计师，担任导师或客座教授，为字体设计教学提供有力支持；另一方面，字体企业与高校师生紧密合作，积极推动设计成果的创新转化。

如方正与中国美术学院合作推出的首款高校定制字体——方正国美进道体，与湖南师范大学美术学院教授李少波合作推出了方正方俊黑系列字体，与中国美术学院教师孙善春合作推出了首款屏幕手写字体——方正善春屏写；汉仪与上海视觉艺术学院副教授陈嵘合作推出了汉仪新人文宋系列字体，与大连民族大学设计学院教师战国栋合作推出了汉仪瑞虎宋、汉仪瑞意宋，与湖南师大李少波教授合作推出了首款地方文化字体——汉仪霸蛮体。

这些合作字体的推出，是字体设计领域产学研良性循环的体现，既丰富了中文字体的种类，也为更多产学研合作提供了示范。

1.4.1.6. 字体设计力量不断壮大

字体价值的广泛认可，字体企业的健康发展，字体教育的不断普及，行业组织的积极推动，带来了可喜的变化——字体设计力量不断壮大。目前，国内不同规模的字体厂商、字体工作室，有 100 多家，从事专职字体设计的设计师大多有平面设计或书法专业背景，来自八大美院的也不在少数。除了专职字体设计师，还有众多平面设计师、书法

家，积极参与到字体设计创作中来，和字体厂商合作开发字库。在产品化的字体设计之外，越来越多的设计师热衷于运用汉字元素进行设计创作，这类设计师及其作品的传播，也在不知不觉中壮大了字体设计队伍。

行业的发展离不开人才的支撑，除了依靠设计院校的专业人才输出，目前，行业企业、院校也在努力从不同渠道、以不同形式为字体设计师的培育增砖添瓦。『方正奖』设计大赛、「汉仪字体之星设计大赛」、Hiii Typography 中英文字体设计大赛等专业赛事，南京艺术学院和中国文字博物馆举办的“字酷”文字艺术设计展、深圳市平面设计协会举办的 GDC Award（包含字体设计版块）等，吸引了众多设计师和字体爱好者对中文字体设计的关注和参与。与此同时，字体厂商与设计院校合作举办的字体工作坊、设计训练营，则是以短期集中课程的形式，帮助学生提升字体素养、掌握字体设计方法，也颇有成效。

字体设计力量的壮大，必将带来中文字体创新的加速以及产业规模的进一步扩大。

1.4.1.7. 字库 B 端市场呈规模化，运作形式多样化

B 端市场即针对企业或组织的字体授权市场。如今，越来越多的企业开始意识到字体是品牌宣传的重要元素，不管是 logo、广告、海报、包装，都离不开字体，字体应用的好坏直接关系着品牌形象，并且影响企业产品销售。与此同时，字体厂商也在不断完善授权模式，建立简单透明的交易流程，引导企业用户正确购买字体版权。B 端市场作为字体行业的传统市场，近 5 年来发展迅速，逐渐规模化。

除了常规的商业用字授权模式，还出现了许多新的形式，如针对网页的云字库解决方案、针对特殊领域的人口信息字库解决方案，以及企业定制字体、城市定制字体、多文种匹配字体等。B 端市场逐步朝多样化、细分化、定制化、差异化方向发展。

1.4.1.8. 字库 C 端市场发展迅猛，个性化需求激增

C 端市场即针对个人的字体授权市场。当前，我们已经进入移动互联网时代，工具、媒介都变得多元化、自由化，不再是设计的壁垒，与此同时，崇尚个性表达的 90 后、00 后正逐渐成为社会主力军，也是最具消费潜力的群体。伴随国内智能手机市场的快速发展，字库公司逐步与 OPPO、VIVO、华为、小米等手机品牌厂商合作开展非交互类平台授权业务，推出明星手写字体、彩色字体、拼音字体等创意字体，丰富了终端手机用户的系统体验。在各大手机应用市场、手 Q 个性装扮、搜狗输入法、WPS 里，大量个性化字体被年轻消费者购买、使用。

作为近年来新诞生的市场业务类型，C 端市场近几年发展迅猛，并迅速成为红海市场。由于需求激增，大量字体企业、设计工作室及个人设计师参与到 C 端字体的创作中，。随着手机用户的不断增长，C 端市场的销售额也在逐年增长。

1.4.2. 中文字体产业发展趋势

1.4.2.1. 中文字体设计求新求变，不断创新

字体设计是中文字体行业的核心生产力，如何通过设计创新，来满足不断变化的市场需求，是中文字体产业可持续运转的首要问题。在字体价值已得到广泛认可的当下，随着大众审美的不断提高，各行各业都对字体创新提出了需求。

近年来，在广告、包装等应用领域，有时尚感、有创意、风格突出的视觉风格，比较容易受到年轻人的关注，这一市场趋势对时尚创意类字体提出了大量需求；在影视、综艺节目，尤其是新兴的短视频等领域，有文化底蕴、个性鲜明、具有视觉冲击力的字体，往往能带来出其不意的流量效果，这让书法类字体、个性手书类字体成为了市场关注的热点。

中文字体设计只有不断创新、求新求变，顺应市场需求、甚至引领市场需求，才能在竞争中得以生存，为社会创造更多价值。

1.4.2.2. 屏显字体设计大势所趋

如今，屏幕阅读已经成为人们日常获取信息的主要途径。老一代屏显字体为适应较低的像素密度，在中宫、笔画细节、字面率、结构等方面做了很多屈从于技术条件的设计。随着移动互联网的发展，手机高清屏的高像素密度给了屏显字体设计更大的自由度，加之手机阅读的近视距，使得不必再一味追求大字面，从而使字体设计可以回归审美表达，表现空间更为宽广。时代给了字体行业好的机遇，越来越多的字体企业开始了新一代屏显字体的研发。

方正字库先后推出了方正悠黑、方正屏显雅宋、方正悠宋、方正兰亭黑 Pro 等多款阅读舒适、富有人文气息的屏显字体。由仓耳字库推出的仓耳今楷、仓耳云黑、仓耳玄三，也是专门针对屏幕显示而设计。此外， 上海印研所为第七版《辞海》的数字版设计的辞海中黑体、方正为小米手机设计的小米兰亭、为坚果手机设计的 Smartisan T 黑，汉仪为 OPPO 设计的 OPPO Sans、为华为设计的 HarmonyOS Sans，也都是以满足屏幕阅

读的易读性、舒适性为设计出发点，在实际的屏幕应用中都有较好的表现力。

未来，随着屏幕阅读的进一步普及，形式多样化的屏显字体将会成为紧迫的市场需求。

1.4.2.3. 企业定制字体成为关注热点

在竞争激烈的商业环境下，企业之间的差异化竞争，不仅体现在产品功能和外观上，也体现在品牌形象和传播上，而字体作为品牌传播的重要元素，已经得到了越来越多企业的重视。由于字体具有广泛的传播特性，如果一个企业拥有自己的定制字体，当我们看到这个字体时，就有可能联想到这个企业，也就在无形之中对企业形象进行了强化。

如今，随着经济的发展，中国企业不断发展壮大，品牌意识在增强，小米、腾讯、vivo、OPPO、阿里巴巴、京东、可口可乐中国、美团、华为等众多知名企业推出了自己的定制字体，方正、汉仪等字体企业也在不断推广、完善定制字体服务。未来，针对企业的定制字体服务将成为行业的重要发展方向。

1.4.2.4. 字体产业的全球化

全球一体化是当今时代的特征，在中文字体行业主要表现在两个方面：一方面，中文字库里包括外文，如何学习国外优秀的设计方法，并通过全球多语言的字体设计，让中文字体更好地走向国际、满足全世界的使用需求，是摆在中文字体厂商面前的重要课题；另一方面，如何将国外优秀的西文字体引入中国，给中国企业和用户更丰富的用字选择，帮助中国企业更好地走向国际，是行业发展的必然趋势，也是中国字体企业的社会责任。

1.4.2.5. 字体文创迎来热潮

文创产品作为一种兼具审美、功能、内涵的高附加值产品，近年来在全社会的多个领域掀起了热潮。中文字体行业本身有着深厚的文化背景，而中文字体是兼具功能特性和审美特性的视觉化产品，并且种类丰富多样，这些先天条件决定了“字体文创”开发有巨大的潜力和空间。

如今，越来越多的字体厂商和文化企业开始尝试把字体这种文化要素，通过创意与智慧的输入，具象到可以感知、并具有使用功能的各类产品中去，让人们通过这些文创产品，感受字体本身的魅力以及字体背后的文化内涵。未来，围绕字体的文创产品开发，包括与其它知名 IP 的合作文创，将会成为热潮，走进更多大众的日常生活中。

1.4.2.6. 技术驱动行业发展

科技改变世界，创新成就未来。随着全球字体技术的不断发展，越来越多的新技术不断助力字体行业的发展，如字库压缩技术、Webfont 技术、可变字体技术等，这些技术不仅会改变未来中文互联网的视觉面貌，同时也将成为字体行业发展的驱动力量。未来，还将有哪些革命性的字体技术出现，会带来怎样的行业创新，让我们拭目以待。

1.5.总结与展望

近几年汉字在设计中的作用得到了广泛的认可，出现了大家都在说的“汉字热”，越来越多的设计师运用汉字元素进行设计创作，在字体设计方向，众多知名平面设计师和字体厂商合作开发字库，众多字体书法家也将他们的作品写出来做成电脑字库，参与汉字设计、字体设计的人越来越多。这种现象的出现，与设计师、书法家群体自身对汉字文化的热爱与探究分不开，也离不开设计院校在这方面的关注，同时，也得益于像深圳市平面设计协会、中国设计师沙龙等专业设计组织的推动。随着设计圈、文化圈的汉字热，越来越多的人投身中文字体设计，字体设计力量不断壮大，设计创新开始加速。

中国字体行业在 30 余年的发展历程中，经历过辉煌，步入过低谷，如今又再次回归大众视野。30 余年来，中文字库从数量到质量都有了巨大的提升，应用范围也从传统的出版、印刷领域逐步拓展到包装、广告、广电、游戏动漫、网站、终端设备等众多领域，移动互联网的迅猛发展，也对字体应用提出了更新、更高的需求，字体已在不知不觉中渗入到了现代人生活的方方面面。如今，中文字体行业已经逐步摸索出了一条从学术研究、字体设计、技术开发，到商业授权、用字服务、字体文创的成熟产业链条，找到了适合自己的盈利模式，产学研也成果斐然，已经从不被大众了解的小众行业，逐步发展成了有广泛大众基础、形成一定规模、有着巨大潜力的重要文化产业。

中文字体行业还面临很多问题和困难，针对大众的字体版权宣传仍需加强，专业字体设计人才培养机制不够健全，字体设计软件不够高效等等。这些问题仍然制约着中国字体行业的发展。未来，汉字字形信息专业委员会将团结领域内的从业人员和企事业单位，在荆棘中继续前行，为繁荣和发展中文字体事业而努力！

第二章　速记研究进展、现状及趋势

2.1.研究背景与意义

速记的目标是使用简便的符号快速记录语言。中文手写速记至今已百余年，曾出现了很多速记方案，吸引了大批人员学习与应用。随着录音录像、个人电脑等的普及，包括中文速录机的发明与推广应用，以及语音识别技术的快速发展，手写速记的学习与使用人群逐渐缩小。手写速记使用方便、快捷的优势，促使一些速记家们仍在不断创新手写速记理论，努力探索更加易于理解和学练的体系，在整体记录速度等各项指标上寻求突破，并寻求更加广泛的应用推广途径。

亚伟中文速录机的发明，奠定了速录产业的基础，形成了速录机、速录服务及速录培训三者支撑的专业速录产业。速录机从第一代 A 型机，到可以独立储存信息的二型机和具备无线连接能力的 RF 型机。后根据市场需求研制了机械拉杆式无声速录机，一体化速录机，以及亚伟速录语音伴侣、移动伴侣、专用支架等周边产品，已形成产品系列。速录服务也从商务应用扩展到学术及政务；速录培训形成了社会专门培训与职业教育共同育人的体系；市场也从现场速录拓展到远程速录，从单一速录拓展到速录翻译等，但总体上呈现供不应求的局面。

语音识别技术是与中文速录机同时商用的。最近 5 年，语音识别领域取得重要突破，以迅飞为代表的语音识别产品得以广泛应用，让很多有速录需求的单位和个人得以享受到语音识别技术带来的类似速录服务的体验。随之而来的就是语音识别的结果如何更好地为人们所使用，有一些是但凭现有技术暂时无法解决的，需要人来配合。速录科技企业开始谋求与语音识别产品的融合，积极探索如何在“人机耦合”状态下充分发挥各自优势。

2.2.领域发展现状与关键科学问题

2.2.1.　如何满足多样化的速录需求

中文速记服务是一个具有广泛需求，并且正在高速增长的市场，主要客户群包括会议主办方、公检法系统及律师事务所、政府机关、新闻传媒、网络直播机构、大型企业、文秘及文学创作等。但由于新闻采访、律师访谈、专家讲座等场合专业速记服务无法“随叫随到”；除一线和部分中心城市外，专业速记人员的服务“一时难求”；以及普通人群

偶尔需要交互体验良好的口述文字记录，目前专业速记服务“高不可攀”等等局限，限制了速录需求的发展。需要从服务模式、软硬件研发等方面进行理论与技术研发以满足多样化的速录需求。

2.2.2. 如何适应速录岗位普及化趋势

亚伟中文速录问世以来，不断为社会创造适应各种应用场合需要的速录岗位。从法院系统到人大、外交部和中央机关，从电视台到网站，从各种会议到企事业单位等，到处都活跃着亚伟中文速录师的身影，在国家机关、事业单位、外资企业、上市公司及国有大中型企业，以及电视、网络、广播等媒体的字幕制作、新闻采访及社会团体等单位的中、高层担任速录师或速录秘书等职。需要不断研究各个岗位的速录需求、工作能力和素质要求，改进人才培养模式，增强速录技术的适应性。

2.2.3. 如何将人工智能与速录系统相融合

采集语言信息是人类一直要面对的一项挑战，不仅劳动强度大，往往还力不从心。从古代速记的起源，到近代打字机的制造和录音机的使用，然后是近现代计算机相关文字录入技术的普及，再到速录技术的诞生，人们的劳动强度逐渐减轻，劳动的效率突飞猛进。当近年来“人工智能”技术应用闪亮登场，给大家带来了希望，希望能够依赖“语音识别系统”自动完成所有的工作。急需进行人工智能与速录技术的比较研究，找到两者的契合点，完美地发挥两者的优势，应对新技术的挑战。

2.2.4. 如何培养专业化的速录人才

院校亚伟中文速录人才的培养已经从单一的“书记员方向”或“书记官”扩展到了文秘、会展等相关专业，这些专业还将进一步与各个学科领域的不同专业相互衔接，培养专业细分的速录人才，应对方方面面对亚伟速录人才的特殊需求。目前，除法院、检察院外，还有就职于医疗机构、互联网企业等不同行业不同专业领域的专职速录师，更有擅长于服务 IT、金融等专门领域学术会议的职业速录师，而且经常无法满足需求。

2013 年到 2017 年，我会连续 5 年成功申办并组织了“全国职业院校技能大赛”高职文秘速录赛项，参赛选手覆盖全国，有力地推动了中文信息处理领域信息采集与处理人才的培养和就业，为社会培养了数以万计的速录师。

2019 年起，工业和信息化部人才交流中心主办“信息处理大赛”，由北京市速记协会承办，北京速录科技有限公司提供亚伟中文速录技术及竞赛系统支持，促进了以信息

速录为前提的中文信息处理相关专业技能培训的融合。

2020 年，“中文速录”成为教育部第四批“1+X”试点职业技能等级证书，对中文速录技能的工作领域、工作任务和要求进行了重新组合与认定，结合新时代信息处理的需要，为人才培养质量和数量的提升提供了新的可能。

2.3. 领域关键技术进展及趋势

2.3.1. 速录系统国产化适配

速录系统是在 DOS 系统环境下研制的，随着 WINDOWS 系统的出现及升级，速录系统也进行了升级。目前，我国政府正在推行电脑软硬件国产化，这要让速录系统在国产系统下运行，需要进行“适配”工作。速录机属硬件产品，使用通用 HID 设备接口与计算机相连接。以往的产品适合 WINDOWS 环境，为提高效率有所简化，并未严格执行有关技术规范，不能直接被国产操作系统识别，也需要进行“适配”工作。该任务目标是完全适配所有的国产操作系统，保证在国产操作系统环境下能够正常使用。重点是开发统信 UOS、麒麟等国产操作系统环境下的速录系统版本。相应地生产改进的新型速录机硬件，使其兼容所有的操作系统环境。目前软件已实现基本功能，正在进行适配测试，通过认证后即可上线下载。硬件也已经具备批量生产能力。该任务主要影响全国法院系统办公及庭审应用，政府部门如人大、政协、各大部委的办公及会议记录等。

2.3.2. 人工智能语音识别与操作速录软硬件的速录师融合

语音技术发展到今天已经达到了实用门槛，加上云计算、大数据等技术的配合，语音技术正在深刻改变世界。在专业速录领域，人工智能语音识别在一些场合已经能够准确地进行同步“逐字”记录。语音识别产品已经部署到了法院庭审等场合，还提供“录音转写”的服务。但以商业服务的要求，语音识别仍难以替代专业速录服务，其原因在于语音识别的结果还不能满足客户稿件的最终要求，与速录师的稿件相比，依然是“半成品”，后期再整理的工作依然不小。目前速录公司与语音识别企业正在努力将两者融合，充分发挥各自的长处，减轻速录师的劳动强度，提供高质量的记录稿。此外，在机器翻译领域，也在与人工速录进行融合，提高机器翻译结果的可读性。人工智能与中文速录的融合，可能改变专业速录服务的工作模式，人工智能有可能以主、从两种方式融合到速录服务过程中。

2.3.3. 推广中文信息处理应用技能职业教育，推行中文速录职业技能等级证书

速记速录（含人工智能语音识别）都属于信息的采集。在大数据时代，信息的采集、整理与应用技能直接影响岗位工作质量与效率。目前，110 余所合作职业院校设立了中文速录人才培养基地，校企共同制定专业人才培养方案、教学计划，组织行业企业的专家、技术能手进校园、进课堂担任导师，定期组织院校教师、专业负责人到企业进修，定期组织教师开展技术培训及教学研讨，推进了人工智能背景下中文信息处理领域相关从业人员及职业院校相关专业学生的信息素养教育与训练，增强了相关人员的信息处理与运用能力和实际操作技能，促进了信息和大数据产业应用技术的普及与推广，为信息处理类相关专业群提供了实训建议及实施方案。在此基础上，“中文速录”被教育部列入 1+X 职业技能等级证书第四批推广计划。

2.4.领域产业发展现状及趋势

2.4.1. 全面适应办公需求

中文速录的高速度，能够保证思维的连贯性，抓住灵感，保证任务及时完成，节约时间，提高效率；中文速录的过程，运动频率低，能够有效减轻人们信息采集的劳动强度，降低紧张感，减少疲劳，轻松工作；中文速录以词为单位录入原理，使工作过程对脑力的消耗低，不容易干扰头脑的创造性思考，能够轻松实现边听、边想、边采集、边整理，使头脑集中思考，减轻思维负担，提高文稿的逻辑性，激发创作的灵活性，保证材料的严谨、通顺、精彩。

根据北京速录科技有限公司人才中心提供的信息，在各行各业各类机构的行政办公岗位上，累计有约 15 万速录文秘从业人员。他们主要分布在科技、银行金融、文化、传媒、房地产、生物、医药等行业，其中包括各类央企、上市公司、大企业集团等。

据国统计局《中国统计年鉴 2020》公布的数据，全国规模以上工业企业数量为 377815 家，若每一家企业都拥有一名速录文秘人员，就是近 40 万人，人才缺口很大，这还没有包含非工业企业、社会组织、慈善机构等。

2.4.2. 全面覆盖各行各业

中文速录技术实现庭审记录计算机化。从 1997 年起，我国法院系统开始推广应用中文速录技术实现庭审记录计算机化。全国各个法院纷纷组织书记员速录技能培训，大

幅提升书记员庭审记录的专业能力，缩短开庭时间，提高办案效率。

截止 2018 年上半年，全国法院、检察院共有员额法官 21 万多名。按 1：1：1，约 38％的法官、检察官没有书记员！书记员缺额约 8 万人。2018-2019 年，全国 10 个省、市、自治区发布法院、检察院书记员招聘公告，累计招聘岗位 17548 人。而当年，全国近 80 所开设法律事务、法律文秘等与书记员岗位相关专业的职业院校，年培养人数仅约 5000 人，供不应求。

中文速录应用全面提升速记工作专业能力。全国人大信息中心从大专毕业生中定向招聘、集训、优选，组建专业速录队伍，承担全国两会及每两个月一次的人大常委会会议记录任务，会议简报等相关文件的速度和质量全面提升，得到领导高度认可，代表委员们的满意度提升。

2.4.3. 全面满足职业发展

中文速录发展迅猛。速录市场是随着“会展经济”“论坛经济”的发展而带动起来的。根据中国旅游饭店业协会、中国旅行社协会、中国会议酒店联盟联合对国内 25 个省市自治区 89 个城市的会议情况统计表明，每年召开超过 15000 场大型会议，近 70%为企业主办。这还只是全部会议的一个缩影。如今，网络媒体的迅猛发展再次将速录人才的紧缺现象加速凸现，重大事件、重大新闻、各种论坛、明星访谈等都要通过媒体进行文字直播，速录秘书已成为大型企事业单位的重要人才，都有巨大的市场需求。

中文速录 1+X 职业技能等级证书全面启动。1+X 职业技能等级证书制度的施行是深入贯彻落实全国教育大会、全国职教大会精神，完善职业教育和培训体系，深化复合型技术技能人才培养培训模式和评价机制改革，提高人才培养质量的重要举措。中文速录职业技能等级证书结合了现代办公和专业速录服务的社会需求，体现了相关岗位的技能要求，涵盖了现代文秘等专业的相关培养目标，适合职业院校在现代文秘等专业开展培训试点工作，有利于提升现代文秘类等专业学生的综合素质和就业能力。

2.5.总结及展望

2.5.1. 速录需求多样化

智能手机和移动互联网的普及，对人们的工作和生活状态产生了深远的影响，大大方便了信息的传递与获取。未来，人们势必越来越离不开网络，也越来越依赖网络带来的便利与快捷，直接引发了随时、随地的速录应用需求。同时，这些变化打破了传统速

录服务的地域局限、时间局限和消费局限，使中文速录能够满足消费者的多样化速录应用需求。

2.5.2. 速录岗位普及化

提高信息处理的效率正在成为人们的普遍共识，同一个现场不同主体分别购买速录服务以及同一个活动要求多个现场同时提供速录服务的现象屡见不鲜；招聘速录岗位的单位也越来越多。未来，速录岗位普及化的趋势不可避免。中文速录将渗透到各行各业、各个领域，走近大多数人的身边。

2.5.3. 速录人才复合化

对于速录人才，速录技能是一项核心的职业能力，但在不同的就业领域，速录人才应具备的速录速度以及相关的能力模块，也有较大差异。速录秘书要能胜任企业办公行政的各种任务；书记员要完成司法文书的使用与制作、法院笔录的制作、诉讼文书档案管理；职业速录师，要在工作中不断学习，扩充专业知识。总之，速录任务趋向复杂，速录人才走向复合，更加强调对速录师综合能力和素质的要求。

2.5.4. 速录系统智能化

信息采集的目的是为了应用，只有对所采集到的素材进行精确合理地处理，才能“为我所用”。现在客户对高级职业速录师的要求很苛刻，一定要做到“语言毕、文稿成”，“成”的是“文稿”，具足“信”“达”“雅”，可以直接拿来使用。她们在速录的同时就在对所听到的信息实时进行必要的处理，才能够直接成稿。而这个要求目前是单纯使用人工智能系统所不能完成的。当然，对速录师的体力和脑力消耗极大。“人工智能+速录”的智能化系统，将帮助速录人员轻松完成信息处理前端任务，大大减轻速录师的劳动强度，把主要精力放在对信息的处理上，这是未来智能化中文速录系统发展的主要方向。

2.5.5. 人才培养专业化

中文速录从单纯的中文速录技术技能的熟练掌握和使用，逐渐融入行政办公领域的相关技能，又“回归”到中文信息处理领域的规范要求，还将进一步整合网络多媒体、人工智能等新技术领域对中文速录的新要求，不断丰富中文速录人才培养的技能规范。未来应用需求的增长，将带动人才培养数量的增加。以法院系统公开的数据测算，仅书

记员的需求量就是 2017 年培训数量的 10 倍之多。由此可以预见未来社会对中文速录人才数量的需求。且未来人才专业化培养的程度越来越高，能适应各个细分的专业领域。

第三章　计算语言学研究进展、现状及趋势

3.1.研究背景与意义

自然语言，通常指的是人类语言（本文特指文本符号，而非语音信号），是人类思维的载体和交流的基本工具，也是人类区别于动物的根本标志，更是人类智能发展的重要外在体现形式之一。

计算语言学（Computational Linguistics，CL）主要从语言学的角度出发，是语言学的一个分支，也是语言学、心理学、数学和计算机科学相互渗透的一门交叉学科。它既要利用计算机对各种语言现象进行定量化、精密化的统计研究，又要对语言规律作出形式化的描述，为计算机的信息处理提供理论依据。

自然语言处理（Natural Language Processing，NLP）主要从计算机科学的角度出发，研究用计算机来理解和生成自然语言的各种理论和方法，属于人工智能领域的一个重要甚至核心分支。计算语言学与自然语言处理的界限早已变得模糊，因此本文将不对这两个概念加以区分。

随着互联网的快速发展，网络文本呈爆炸性增长，为自然语言处理提出了巨大的**应用需求**。同时，自然语言处理研究也为人们更深刻地理解语言的机理和社会的机制提供了一条重要的途径，因此具有重要的**科学意义**。

目前，人们普遍认为人工智能的发展经历了从运算智能到感知智能，再到认知智能三个发展阶段。其中，**运算智能**关注的是机器的基础运算和存储能力，在这方面，机器已经完胜人类。**感知智能**则强调机器的模式识别能力，如语音的识别以及图像的识别，目前机器在感知智能上的水平基本达到甚至超过了人类的水平。然而，在涉及自然语言处理以及常识建模和推理等研究的**认知智能上**，机器与人类还有很大的差距。是否具有认知智能，也被认为是人类和动物的主要区别之一。

为什么计算机在处理自然语言时会如此困难呢？这主要是因为自然语言具有高度的**抽象性**、近乎无穷变化的语义**组合性**、无处不在的**歧义性**、标注规范的**主观性**、语言表达的**非规范性**和持续的**进化性**等，理解语言通常更是需要**背景知识**和**推理能力**。

综上所述，由于自然语言处理所面临的众多问题，使其成为目前制约人工智能取得更大突破和更广泛应用的瓶颈之一。因此自然语言处理又被誉为**“人工智能皇冠上的明珠”**，并吸引了越来越多的人工智能研究者加入该研究之中。

3.2.领域发展现状与关键科学问题

3.2.1. 自然语言处理发展历史

自然语言处理自诞生之日起经历了两大研究范式的转换，即**理性主义**和**经验主义**。受到语料规模以及计算能力的限制，早期的自然语言处理主要采用基于理性主义的规则方法，通过专家总结的符号逻辑知识处理通用的自然语言现象。然而，由于自然语言的复杂性，基于理性主义的规则方法在面对实际应用场景中的问题时显得力不从心。

20 世纪 90 年代开始，随着计算机运算速度和存储容量的快速增加，以及统计学习方法的愈发成熟，使得以语料库为核心的**统计学习方法**在自然语言处理领域得以大规模应用。由于大规模的语料库中包含了大量关于语言的知识，使得基于语料库的统计自然语言处理方法能够更加客观、准确、细致地捕获语言规律。这一时期，词法分析、句法分析、信息抽取、机器翻译、自动问答等领域的研究均取得了一定程度的成功。

尽管基于统计学习的自然语言处理取得了一定程度的成功，但它也有明显的局限性，也就是需要事先利用经验性规则将原始的自然语言输入转化为机器能够处理的向量形式。这一转化过程（也称为特征提取）需要细致的人工和一定的专业知识，因此也被称为**特征工程**。

2010 年之后，基于深度神经网络的**表示学习**方法（也称**深度学习**）逐渐兴起，可以直接端到端地学习各种自然语言处理任务，不再依赖人工设计的特征。所谓表示学习，是指机器能根据输入自动发现可用于识别或分类等任务的表示。具体地，深度学习模型在结构上通常包含多层的处理层。最底层的处理层接收原始输入，然后对其进行抽象处理，其后的每一层都在前一层的结果上进行更深层次的抽象，最后一层的抽象结果即为输入的一个表示，用于最终的目标任务。其中的抽象处理，是由模型内部的参数来进行控制的，而参数的值则是根据训练数据上模型的表现，使用反向传播算法学习得到的。由此可以看出，深度学习可以有效地避免统计学习方法中的人工特征提取操作，自动地发现对于目标任务有效的表示。在语音识别、计算机视觉等领域，深度学习已经取得了目前最好的效果，在自然语言处理领域，深度学习同样引发了一系列的变革。

除了可以自动发现有效特征外，表示学习方法的另一个好处是打通了不同任务之间的壁垒。传统统计学习方法需要针对不同的任务设计不同的特征，这些特征往往是无法通用的。而表示学习能够将不同任务在**相同的向量空间**内进行表示，从而具备跨任务迁移的能力。除了可以跨任务外，还可以实现跨语言甚至跨模态的迁移。从而可以综合利用多项任务、多种语言、多个模态的数据，使得人工智能向更通用的方向迈进了一步。

同样，得益于深度学习技术的快速发展，自然语言处理的另一个主要研究方向—**自然语言生成**也取得了长足进步。长期以来，自然语言生成的研究几乎处于停滞状态，除了使用模板生成一些简单的语句外，并没有什么太有效的解决办法。随着基于深度学习的序列到序列生成框架的提出，这种逐词的文本生成方法全面提升了生成技术的灵活性和实用性，完全革新了机器翻译、文本摘要、人机对话等任务的技术范式。

虽然深度学习技术大幅提高了自然语言处理系统的准确率，但是基于深度学习的算法有一个致命的缺点，就是过度**依赖于大规模有标注数据**。对于语音识别、图像处理等感知类任务，标注数据相对容易获得，如在图像处理领域，人们已经为上百万幅的图像标注了相应的类别（如 ImageNet 数据集）；用于语音识别的“语音—文本”平行语料库也有几十万小时。然而，由于自然语言处理这一认知类任务所具有的“主观性”特点，以及其所面对的任务和领域众多，使得大规模语料库标注的时间和人力成本都过于高昂，因此自然语言处理的标注数据往往不够充足，很难满足深度学习模型训练的需要。

3.2.2. 自然语言处理发展现状

早期的静态词向量预训练模型，以及后来的动态词向量预训练模型，特别 2018 年以来，以 BERT、GPT 为代表的**超大规模预训练语言模型**恰好弥补了自然语言处理标注数据不足的缺点，帮助自然语言处理取得了一系列的突破，使得包括阅读理解在内的几乎所有自然语言处理任务性能都得到了大幅提高，在有些数据集上甚至达到或超过了人类水平。

所谓**预训练模型**（Pre-trained Models），即首先在一个原任务上预先训练一个初始模型，然后在下游任务（也称目标任务）上继续对该模型进行**精调**（Fine-tune），从而达到提高下游任务准确率的目的。本质上，这也是迁移学习（Transfer Learning）思想的一种应用。然而，由于同样需要人工标注，导致原任务标注数据的规模往往也是非常有限的。那么，如何获得更大规模的标注数据呢？

其实文本自身的顺序性就是一种天然的标注数据，通过若干连续出现的词语预测下一个词语（又称**语言模型**）就可以构成一项源任务。由于图书、网页等文本数据规模近乎无限，这样就可以非常容易地获得超大规模的预训练数据。有人将这种不需要人工标注数据的预训练学习方法称为无监督学习（Unsupervised Learning），其实这并不准确，因为学习的过程仍然是有监督的（Supervised），更准确的叫法应该是**自监督学习**（Self-supervised Learning）。

为了能够刻画大规模数据中复杂的语言现象，还要求所使用的深度学习模型容量足够大。基于自注意力的 **Transformer** 模型显著地提升了对于自然语言的建模能力，是近

年来具有里程碑意义的进展之一。要想在可容忍的时间内，在如此大规模的数据上训练一个超大规模的 Transformer 模型，也离不开以 GPU、TPU 为代表的现代并行计算硬件。可以说，超大规模预训练语言模型完全依赖“蛮力”，在**大数据**、**大模型**和**大计算资源**的加持下，使自然语言处理取得了长足的进步。如 OpenAI 推出的 GPT-3，是一个具有 1,750 亿参数的巨大规模，无需接受任何特定任务的训练，便可以通过小样本学习完成十余种文本生成任务（如问答、风格迁移、网页生成、自动编曲等）。目前，**预训练模型已经开启了自然语言处理的新时代**。

综上，可以看出自然语言处理的发展历史呈现了一种明显的**“同质化”**趋势。早期的自然语言处理算法需要根据不同的任务编写特定的逻辑将输入文本转换为更高级别的特征，然后使用相对同质化的机器学习算法（如支持向量机）进行结果预测；此后，深度学习技术能够使用更加同质化的模型架构（如卷积神经网络），在输入文本上直接进行学习，并在学习的过程中自动**“涌现”**出用于预测的更高级别的特征；而预训练模型同质化的特性更加明显，目前几乎所有最新的自然语言处理模型都源自少数大规模预训练模型（如 BERT、RoBERTa、BART、T5 等）。GPT-3 模型更是能够做到一次预训练，即可直接（或仅使用极少量训练样本）完成特定的下游任务。

接下来，本文将从自然语言处理的范式迁移、词法句法分析、语义分析、信息抽取和基于知识的自然语言处理等五方面对自然语言处理领域近五年（2017-2021）的发展加以总结和展望。

3.3. 领域关键技术进展及趋势

3.3.1. 自然语言处理的范式迁移

范式是建模一类任务的通用框架。过去几年随着神经网络架构逐渐向 Transformer[1] 统一以及大规模预训练模型[2]的普及，大多数自然语言处理（NLP）任务的建模已经收敛到几种主流的范式，本节将梳理过去几年中 NLP 的范式迁移现象和趋势，更全面的介绍可以参考文献[3]。

3.3.1.1. 任务定义和目标

本文将 NLP 任务中广泛使用的范式归为以下 7 类，即分类（Class）、匹配（Matching）、序列标注（SeqLab）、阅读理解（MRC）、序列到序列（Seq2Seq）、序列到动作序列（Seq2ASeq）和语言模型（(M)LM），如图 1 所示。

具体的范式描述如下：

分类范式（Class）为文本指定预定义的标签。文本分类通常将文本输入一个基于深度神经网络的编码器来提取特征，然后将其输入一个浅层分类器来预测标签，如$\mathcal{Y}$=CLS(ENC($\mathcal{X}$))。$\mathcal{Y}$可以是独热编码，ENC(·)通常是卷积网络、循环网络或 Transformers，CLS(·)常由一个简单的多层感知器和汇聚层实现。

匹配范式（Matching）是预测两个文本语义相关性的一种范式。Matching 范式可以简单地表述为$\mathcal{Y}$=CLS(ENC($\mathcal{X}_a$,$\mathcal{X}_b$))，$\mathcal{X}_a$和$\mathcal{X}_b$是被预测的两段文本，$\mathcal{Y}$可以是离散或连续的。

序列标注范式（SeqLab）可用于模拟各种任务，如词性标注（POS）、命名实体识别（NER）和组块分析。传统的基于神经网络的序列标注模型由编码器和解码器组成，如$y_1,\cdots,y_n$=DEC(ENC($x_1,\cdots,x_n$))。$y_1,\cdots,y_n$是$x_1,\cdots,x_n$对应的标签。

机器阅读理解范式（MRC）从输入序列中提取连续词元序列（span）来回答给定的问题。MRC 范式可以描述为$y_k\cdots y_{k+l}$=DEC(ENC($\mathcal{X}_p$,$\mathcal{X}_q$))，$\mathcal{X}_p$和$\mathcal{X}_q$表示篇章和问题，$y_k\cdots y_{k+l}$是从$\mathcal{X}_p$或$\mathcal{X}_q$中获得 span。

序列到序列范式（Seq2Seq）是一种通用且功能强大的范式，可以处理各种 NLP 任务。Seq2Seq 范式通常由编码器—解码器框架实现，如$y_1,\cdots,y_m$=DEC(ENC($x_1,\cdots,x_n$))。与 SeqLab 不同，这里输入和输出的长度不需要相同。

序列到动作序列范式（Seq2ASeq）是一种广泛使用的结构化预测范式。Seq2ASeq 范式的例子通常被称为基于转移的模型，可规范为$\mathcal{A}$=CLS(ENC($\mathcal{X}$),$\mathcal{C}$)，$\mathcal{A}=a_1,\cdots,a_n$是动作序列，$\mathcal{C}=c_1,\cdots,c_{m-1}$是状态序列。

语言模型范式（LM）估计给定单词序列出现在句子中的概率。它可以被简单表示为x_k=DEC($x_1,\cdots,x_{k-1}$)，DEC 可以是任何自回归的模型。一种 LM 的变体 ML 可以被规范为：$\bar{x}$=DEC(ENC($\tilde{x}$))，$\tilde{x}$由将x的一些词元（token）替换为特殊词元[MASK]得到，x表示待预测的词元。

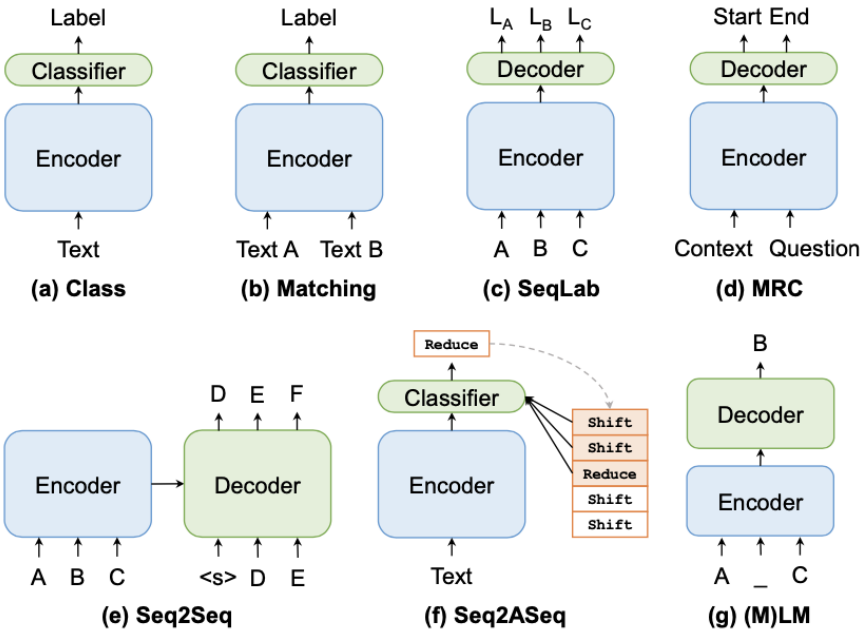


图 1 自然语言处理中的七种主流范式

3.3.1.2. 技术方法与研究现状

本节回顾在不同 NLP 任务中发生的范式转移：文本分类、自然语言推理、命名实体识别、方面级情感分析、关系抽取、文本摘要和语法分析。

传统的文本分类任务可以通过 Class 范式很好地解决。但其变体（如多标签分类）可能具有挑战性。为此，Yang et al.[4]采用 Seq2Seq 范式，以更好地捕捉多标签分类任务中标签之间的相互作用。Sun et al.[5]采用 Matching 范式预测输入对$(\mathcal{X}, L_y)$是否匹配，$\mathcal{X}$是原文本，L_y是类y的描述。

自然语言推理（NLI）通常在 Matching 范式中进行建模，两个输入文本$(\mathcal{X}_a, \mathcal{X}_b)$被编码并互相作用，再连接分类器预测它们的关系。随着 BERT 等功能强大的编码器出现，NLI 任务可以通过将两个文本连接为一个文本在 Class 范式中解决。

命名实体识别（NER）可以被分为 3 类：常规 NER、嵌套 NER 和非连续 NER。传统的方法基于 SeqLab、Class 和 Seq2ASeq 来分别解决 3 个任务。Li et al.[6]提出将常规 NER 和嵌套 NER 规范为 MRC 任务。Yan et al.[7]使用一种基于 Seq2Seq 范式的统一模型来解决所有 3 种子任务。

方面级情感分析（ABSB）是一种细粒度的情感分析，可以分为 7 种子任务以被不同的范式处理。Mao et al.[8]采用 MRC 范式处理所有的 ABSB 子任务。Yan et al.[9]通过将任务的标签转化为词元序列，再使用 Seq2Seq 范式来处理。

关系抽取（RE）主要有两个子任务：关系预测和三元组抽取。前者主要通过 Class 范式解决，而后者常以流水线方式处理：首先使用 SeqLab 范式提取实体，再使用 Class 范式预测实体间关系。Zeng et al.[10]使用 Seq2Seq 范式处理三元组抽取任务，

Levy et al.[11]使用 MRC 范式处理 RE 任务。此外，三元组抽取也可以通过转化为多轮对话后用 MRC 范式处理。

解决文本摘要任务有两种不同的方法：抽取式摘要和生成式摘要。前者通常使用 SeqLab 范式，而后者常通过 Seq2Seq 范式直接生成。McCann et al.[12]将其规范为一个问答任务，并使用 Seq2Seq 模型解决；Zhong et al.[13]提出用 Matching 范式处理抽取式摘要。

语法分析在机器翻译和问答等应用中有重要作用。基于转移和基于图的方法是两种常用的手段。前者通常使用 Seq2ASeq 范式，而后者使用 Class 范式解决。通过将目标树结构线性化为一个序列，该任务可以通过 Seq2Seq 范式解决。此外，Gan et al.[14]使用 MRC 范式来解决依存分析任务。

3.3.1.3. 技术展望与发展趋势

一些范式已经显示出将各种 NLP 任务规范为统一框架的潜在能力，提供了将单个模型作为不同 NLP 任务的统一解决方案的可能性。单个统一模型的优势可以概括为：不再需要大量标注数据、泛化能力强以及部署便捷。

主要探讨以下 4 种可能统一不同 NLP 任务的范式：(M)LM、Matching、MRC 和 Seq2Seq。将下游任务规范为(M)LM 任务是利用预训练语言模型的自然方式。(M)LM 可使用无监督数据处理理解和生成任务。另一个可能的统一范式是 Matching。Matching 的优势在于只需要设计标签描述，工程量较小。但 Matching 需要大量 NLI 数据进一步训练，领域迁移受限，且无法做生成任务。MRC 范式通过生成任务特定的问题并训练 MRC 模型，从输入文本中根据问题选择正确的 span。MRC 的框架模型十分通用，但难以发挥已有训练模型的能力。Seq2Seq 是一个通用且灵活的范式，非常适用于复杂任务，但也受限于自回归生成导致较慢的推理速度。

最近，基于提示的微调[15]（prompt-based tuning）迅速流行起来。相比之下，其他潜在的统一范式没有得到充分的探索。通过预训练或其他技术探索更强大的 Matching、MRC 或 Seq2Seq 模型或许应受到更多的重视。

3.3.2. 词法、句法分析

3.3.2.1. 任务简介、目标及意义

词法分析和句法分析是自然语言处理的基础任务，可以被应用到许多自然语言处理下游任务中去，例如机器翻译[16]和文本摘要[17]。

词法分析主要包括词性标注这一任务。词性标注指基于词性含义以及词的上下文来为输入文本中的每个词进行词性标注的过程，常见的词性标签有名词、动词、形容词等。词性标注一般没有直接应用场景，但它却能为许多下游任务提供帮助，例如，在词义消歧任务当中，词义和词性常常是相关联的，比如“翻译”一词既可指职业也可指行为，这两个词义的一大区别即为其词性不同：前者为名词而后者为动词。

句法分析旨在对输入的文本句子进行分析以得到句子的句法结构。常见的句法分析有依存句法分析和成分句法分析。依存句法分析识别句子中词与词之间的相互依存关系，而成分句法分析识别句子中的层次化短语语法结构。句法分析在诸多自然语言处理下游任务中都有应用，例如在嵌套命名实体识别任务中，由于实体间存在相互嵌套现象，因此非常适合和成分句法分析中的层次化短语语法结构共同建模。

3.3.2.2. 技术方法和研究现状

词法分析 最简单的词性标注器是使用字典中最常见的词性作为当前词的词性，但这种简单的规则只可以解决大约 85%的词性标注问题。为了解决词性歧义的问题，研究者们使用机器学习算法进行词性预测。在基于统计方法的时代，研究人员手动提取字词特征，例如字母大小写、前缀、后缀等特征，并使用隐马尔可夫、条件随机场等模型计算可能的标签序列的概率分布，并选择最佳标签序列作为输出[18,19]。进入神经网络时代后，常见的做法是使用 LSTM[20]、Transformer[21]等编码器对输入文本进行编码，并使用 Softmax 或者 CRF 进行解码预测，这种方法在基于《华尔街日报》的 WSJ 数据集上取得了超过 97%的准确率[22]。近几年以来，为了进一步提升性能和鲁棒性，研究人员尝试在词性标注模型上展开编码长距离标签依赖关系[23]等工作。

句法分析 主流的句法分析方法主要分为两种：基于转移的方法[24]和基于图的方法[25–27]。基于转移的方法通过预测一系列转移操作来构建合法的句法树结构，这种方法需要同时建模缓存区（已经生成的部分树结构）、堆栈区（等待输入的文本序列）和已经预测出来的转移操作序列，其中常见的缓存区和堆栈区的建模方法为 stack-LSTM，转移操作序列的建模方法常用 LSTM；基于图的方法首先编码输入、给文本局部打分，而后采

用动态规划等算法来恢复句法树结构，该种方法采用的主流编码器包括 LSTM 和 Transformer，解码器一般基于最大生成树算法（依存句法分析）或 CKY 算法（成分句法分析）。近几年来，随着大规模预训练语言模型的出现，BERT、XLNET 等预训练语言模型也常被用作句法分析器的编码器。当前最佳的依存句法分析器是基于图的方法[28,29]，使用 BERT 后可以在基于《华尔街日报》来标注的宾夕法尼亚大学树库数据集上取得了超过 96%的有标签 F-1 分数；最佳的成分句法分析器亦采用了基于图的方法[27]，在使用 BERT 的情况下在宾大树库上取得了接近 96%的 F-1 值。与此同时，句法分析领域也有新的模型架构、转移范式不断涌现，例如，Zhang et al.[30]提出了一种可以批处理的基于 CRF 的成分句法分析器，Yang et al.[31]提出一种基于连结（attach）和并列（juxtapose）的新转移范式。

联合建模 为了解决错误传播问题、进一步提高词法分析和句法分析模型的表现，一个常见方法是将词性标注和句法分析进行联合建模[32]。具体来说，词性标注、依存句法分析和成分句法分析这三个任务中，任意两个任务或者全部三个任务均可组合起来进行联合建模。研究人员发现，联合建模可以有效提升参与建模的各个任务的准确率，例如，Zhou et al.[28]在宾大树库上进行依存句法分析和成分句法分析的联合建模，在两个任务上的错误率分别比单独建模减少了 16%和 3%。

3.3.2.3. 发展趋势

在词法和句法分析任务上，随着在新闻领域（宾大树库所基于的领域）内模型的表现接近理论上限，研究人员们将视线转向了更加具有实用性、同时也富有挑战性的跨领域和多语言场景中去，具体来说，研究人员们试图探究在低资源、零资源的情景下如何使得词法、句法分析器仍旧得以应用，沿着这个研究方向，近期工作包括了跨语言、跨领域词法分析器的设计[33,34]、新领域树库的构建[35]和跨领域、跨语言句法分析器的构建[36,37]等工作。

3.3.3. 语义分析

3.3.3.1. 任务简介、目标及研究意义

语义分析（semantic analysis）是生成意义表示并将这些意义指派给语言输入的过程[38]。根据语言输入的粒度不同，语义分析又可进一步分为词汇级语义分析、句子级语义

分析和篇章级语义分析。通常，词汇级语义分析主要关注如何区分和获取单个词语的语义，经典任务是词义消歧（Word Sense Disambiguation, WSD）[39]，即在特定的语境中，识别出某个歧义词的正确词义；句子级语义分析主要关注解析由词语所组成的句子的语义，根据分析的深浅程度又分为浅层语义分析和深层语义分析，其中浅层语义分析的经典任务是语义角色标注（Semantic Role Labeling, SRL）[40]，即识别出给定句子的谓词及谓词的相应语义角色成分。深层语义分析，又称为语义解析，即将输入的句子转换为计算机可识别、可计算的语义表示，语义解析又根据应用情境的不同，可分为自然语言到结构化查询（language to query）、语言到代码（language to code）和语言到机器操作指令（language to instruction）；篇章级语义分析主要关注由句子组成的篇章的内在结构并理解各个句子的语义以及句子与句子之间的语义关系，进而理解整个篇章的语义。词语级语义分析是句子、篇章语义分析的基础，句子级语义分析又是篇章语义分析的基础。

语义分析是自然语言处理的核心任务，其目标是实现对语言输入的语义理解，进而支撑后续的操作和处理。在理论上，语义分析涉及语言学、计算语言学、认知科学、神经科学等多个学科，语义分析的研究和进展可推动多个相关学科的发展。在应用上，语义分析对自然语言处理领域的其他任务都有一定的促进作用。如现代机器翻译，虽然目前的神经机器翻译系统已取得媲美人类甚至超过人类的翻译效果[41]，但要真正达到“信、达、雅”的标准，还需要有语义分析的参与。如现代的语义搜索引擎，从以前的匹配查询与文档转变为了理解用户提交的查询的意图，能够更精准的向用户返回最符合需求的搜索结果。另外，知识获取方面，它与语义分析是相互促进的，一方面，语义分析需要知识的支撑，更大、更全、更准确的知识库对语义分析有着至关重要的作用；另一方面，为了从自由文本中获取更多结构化的知识，语义分析又是必不可少的技术。

目前，语义分析的研究吸引了国内外大批学者，但大部分都集中于句子级语义分析方向上，词汇级和篇章级的研究工作甚少。主要因为词汇级语义分析，如词义消歧，已发展多年，技术已趋成熟，研究的重心转向句子级的语义分析；而篇章级语义分析由于完全体的篇章理解过于困难，因此衍生了多个与之相关的任务，如篇章的结构分析、话语分割、指代消解、共指消解等，任务分散且偏边缘，导致得到的研究关注很少，进展也缓慢。整体来说，语义分析虽然已取得了一定的进展，但技术还远未成熟完美。接下来本文对语义分析的研究进展与影响、技术展望和发展趋势作简要介绍。由于篇幅有限，接下来的内容只涉及备受关注的深度句子级语义分析，即语义解析。

3.3.3.2. 研究进展与影响

在深度神经网络模型崛起之前，语义分析领域基于文法和组合规则的模型占据主流。

近 5 年来，随着神经网络模型的兴起，特别是序列到序列模型（Seq2Seq）在自然语言处理多个任务上的成功，如机器翻译[42]，语义分析任务上也开始尝试将语义分析问题建模为序列到序列的问题。近 2 年，随着像 BERT[43]、GPT[44]这样的大规模预训练语言模型的提出，并在自然语言处理的多个任务上面取得 SOTA，整个 NLP 领域都转型采用预训练+精调的新研究范式[45]。为了更好的利用大模型里面的知识，NLP 领域还兴起了基于提示语（prompt）的方法浪潮[46]。深度语义分析领域也紧跟整个 NLP 领域的大潮，与之对应的先后出现了基于序列到序列的语义分析方法面向语义分析的预训练方法和基于大模型受限生成的方法。

其中基于序列到序列的语义分析方法[47–50]的核心在于将结构化的语义表示序列化，把语义表示看成一系列的语义单元。相比基于文法和组合规则的方法，Seq2Seq 方法非常简单，是端到端的，不需要人工设计特征，也不需要学习文法和组合规则。然而，Seq2Seq 的方法也忽略了一个问题，不同于机器翻译，语义分析的目标语言不是一种自然语言，而是一种形式化语言，它具有层次结构，Seq2Seq 方法只是简单地将语义表示偏平序列化，忽略了语义表示的层次结构信息，基于此，Dong et al.[47]提出了 Seq2Tree 的方法，其核心是一个层次化的解码器，解码时不再生成偏平化的语义表示序列，而是生成层次结构化的语义表示，简而言之，用一个层次树结构的形式来表征语义，序列化时，采用层次结构树的广度优先遍历的形式。考虑到 Seq2Seq 和 Seq2Tree 方法都忽略了语义表示 token 之间的紧密联系，Chen et al.[50]提出了一种 Seq2Action 的方法，该方法采用语义图作为语义表示，然后将语义图进行原子级分解，用设计好的动作序列来表示语义图的构建，进而用编码器-解码器模型框架来生成动作序列，并利用到语义表示 token 之间存在严格的句法和语义约束，提出了一种受限的解码方法。基于序列到序列的语义分析方法由于其简单而有效的特点，成为了目前语义分析领域最常用的基线模型。

与其它面向特定任务的预训练模型方法类似，面向语义分析的预训练模型[51–54]也包含两个关键：收集数据和设计自监督学习任务。针对 text-to-sql 的语义分析问题，典型的预训练模型是 GraPPa[53]，其采用了两种常用的用于 text-to-sql 问题的数据收集方法，一是从已有的跟表格有关的数据中抽取表格与自然语言对，二是利用同步文法在新采样的表格上自动生成（表格，自然语言，sql）数据对。预训练模型的输入不同于预训练语言模型的输入，这里的输入是将自然语言查询与表格的表头拼接起来的。自监督学习任务方面，为了在表示层面简历自然语言词语与表头的交互，设计掩码任务，即对输入进行随机的掩码，再进行复原，最后计算损失函数。为了进一步在表示层面学习表，通过预测表头的语义标签来实现。由于是预训练模型，使用方面可以像使用 BERT 一样方便，可适用于所有语义分析模型。

基于大模型的受限生成的方法[55–57]启发于像 T5 在 text-to-text 任务上的成功，以及 GPT 在文本生成任务上的成功。考虑到语义分析任务与 textto-text 问题的不同：语义分析生成的不是自然语言，而是形式化的语义表示，需要满足一定的文法约束，研究者们引入了一种中间语言：经典句式[58]，它是一种介乎于自然语言与语义表示之间的一种语言，又与自然语言类似，但又符合确定性的文法，它与语义表示之间可以通过同步文法进行确定性的转换。基于经典句式，语义分析可以转换成一种受限的复述生成。即给定输入句子，大模型利用复述生成其经典句式，在解码生成过程中可以利用约束来减小解码空间。这类模型的关键在于解码过程中约束的确定，目前一般采用启发式的基于文法的形式引入约束条件。由于大模型，如 T5[59]、BART[60]和 GPT 在 few-shot 和 zero-shot 问题上都表现出色，基于大模型的受限生成语义分析方法在 few-shot 和无监督的设定下也取得了很好的成绩。

3.3.3.3. 技术展望和发展趋势

语义分析技术发展迅速，整体上，紧跟自然语言处理领域的发展大潮，一方面部分方法启发于其他任务的先进技术，如基于序列到序列的语义分析方法，另一方面部分方法也启发了其他领域，如基于受限解码的事件抽取方法[61]。基于对现有技术的分析和总结，本文认为语义分析后续的研究发展趋势主要包括：

（1）通用的面向自然语言理解的预训练模型

目前的面向语义分析的预训练模型由于高质量的标注数据难以获取的问题，预训练模型还只在 text-to-sql 和 code generation 等数据相对容易获取的问题上得以实现。接下来，可以尝试同时面向更加通用的语义分析情境，如面向开放域的问答，语言到机器执行指令等，一个预训练模型，适用所有的语义分析任务。

（2）自学习的控制生成

目前，研究者都已意识到大模型加受限解码在语义分析问题上的威力。但整个过程还需要人工参与，如约束条件需要人来参与设计，用于经典句式与语义表示之间互相转换的同步文法需要人工定义。如何将这些人工参与的部分交给模型自主学习，实现自学习的 soft 的同步文法和自学习的 soft 的条件约束是下一步可研究的点。

（3）状态感知的预训练模型

目前的大模型与世界没有太多交互。而语义分析任务中有些情境需要与世界进行交互，如基于对话执行查询，基于对话执行指令操作等。如何训练一个面向自然语言理解的能与世界进行交互的大模型，即当世界的状态因为当前的动作发生改变时，大模型能否及时的感知到状态的变化，并在理解下个输入的过程中是基于已更新过的世界状态的，

也是一个可探究的点。

3.3.4. 信息抽取

3.3.4.1. 任务定义和目标

信息抽取（Information Extraction）的目标是从非结构化文本中抽取出结构化的信息，主要包括实体抽取、实体关系抽取（Relation Extraction，RE）、事件抽取（Event Extraction，EE）和事件关系抽取（Event Relation Extraction, ERE）等任务[62,63]。实体主要是指文本中名词性的短语，比如人名、地名、机构名、时间、日期、数字等。实体抽取也称为命名实体识别（Named Entity Recognition，NER），包括实体的识别和分类。实体识别就是从文本中找出哪个片段是一个实体。实体的分类就是判断找出的实体属于什么类别，比如：人名、地名等。实体关系抽取则是判断两个实体之间的语义关系，比如“姚明”和“上海市”这两个实体之间是“出生于”的关系，而“北京”与“中国”则是“首都”的关系。事件抽取任务是识别特定类型的事件，并把事件中担任既定角色的要素找出来，该任务可进一步分解为 4 个子任务：触发词识别、事件类型分类、论元识别和角色分类任务。

信息抽取技术是中文信息处理和人工智能的核心技术，具有重要的科学意义。通过将文本所表述的信息结构化和语义化，信息抽取技术提供了分析非结构化文本的有效手段，是实现大数据资源化、知识化和普适化的核心技术。被抽取出来的信息通常以结构化的形式描述，可以为计算机直接处理，从而实现对海量非结构化数据的分析、组织、管理、计算、查询和推理，并进一步为更高层面的应用和任务（如自然语言理解、知识库构建、智能问答系统、舆情分析系统）提供支撑。

3.3.4.2. 技术方法和研究现状

信息抽取的核心是将自然语言表达映射到目标知识结构上，并转换为可供计算机处理的知识。然而，自然语言表达具有多样性、歧义性和结构性，其中蕴含的知识具有复杂性、开放性以及规模巨大的特点，进而导致信息抽取任务极具挑战性。自上世纪 80 年代被提出以来，信息抽取一直是自然语言处理的研究热点。

在早期，大部分信息抽取系统（如 MUC 评测中的信息抽取系统）都采用基于规则的方法，该类方法依靠人工制定规则，其优点是可预判和解释，但面临着移植性差，很多场景很难甚至无法总结有效的规则。自 90 年代以来，统计模型成为信息抽取的主流方法[64]，通常将信息抽取任务形式化为从文本输入到特定目标结构的预测，使用统计模

型来建模输入与输出之间的关联，并使用机器学习方法来学习模型的参数，经典的方法包括使用条件随机场（CRF）将实体识别问题转化为序列标注问题。近年来，随着深度学习时代来临，研究者主要聚焦于如何使用深度神经网络自动学习有区分性的特征，进而避免使用传统自然语言处理工具抽取特征时存在的错误累积问题[65,66]。随着研究的深入，特别是大规模预训练语言模型的引入[43]，基于深度神经网络的信息抽取模型在公开数据集上达到了不错的成绩[67]，但是在实际应用场景效果还不尽人意。

理想设定与实际场景存在巨大鸿沟，近期越来越多的工作针对实际应用中的挑战展开。真实场景中实体、关系、事件具有长尾分布特点，许多关系和实体对的示例较少。对于金融、医疗等垂直领域，缺失标注数据现象更为明显，甚至数据的获取也很困难[68]，而神经网络作为典型的“数据饥渴”模型，在训练样例过少时性能会受到极大影响。针对小样本任务，Ding 等[69]发布了包含 8 种粗粒度和 66 种细粒度实体类的少样本命名实体识别；Han 等[68]发布了小样本关系抽取数据集 FewRel，Gao 等[70]在 FewRel 数据集的基础上提出了 FewRel 2.0，增加了领域迁移（domain adaptation）和“以上都不是”检测（none-of-the-above detection）。利用海量无监督数据得到的预训练模型得到有效的语义特征是少量样本快速学习知识的代表性方法，Baldini 等[71]使用 BERT 来对文本关系进行表示，并且提出了 Matching the blanks 的方法来预训练任务不可知（task agnostic）的关系抽取模型。

真实场景中的信息抽取还面临着复杂的语境，例如大量的实体间关系是通过多个句子表达的，同一个文档中的多个事件相互影响，文档级的信息抽取最近也收到广泛的关注，代表性的方法是使用图神经网络融合分布在文档中不同位置的实体的信息，并利用图算法进行信息的传递。Quirk 等[72]最早尝试构建文档级图，捕获相邻句子之间的关系。Christopoulou 等[73]构建以实体、实体提及（Mention）和句子为节点的文档图，并通过图上的迭代算法得到边的表示进行关系分类，之后有大量的研究者采用类似的方法对文档建模[74–76]。除了使用图网络外，研究者也开始尝试直接使用大规模语言模型建模文档，Xu 等[77]将 Mention 是否在同一个句子中、是否指向同一个实体编码作为实体结构信息送入到 BERT 编码层。Zhou 等[78]提出自适应阈值代替用于多标签分类的全局阈值，并直接利用预训练模型的自注意力得分找到有助于确定关系的相关上下文特征。在大规模预训练语言模型的研究上，研究者也尝试着加入知识增强语义表示，例如 ERNIE 中字、短语和实体三个级别的遮罩（MASK）训练[79]，Qin 等[80]通过对比学习的方式将实体判别、关系判别作为辅助任务帮助模型的训练。

3.3.4.3. 发展趋势

信息抽取技术研究蓬勃发展，已经成为了自然语言处理和人工智能等领域的重要分支。这一方面得益于系列国际权威评测和会议的推动，如消息理解系列会议（MUC，Message Understanding Conference），自动内容抽取评测（ACE，Automatic Content Extraction）和文本分析会议系列评测（TAC，Text Analysis Conference）。另一方面也是因为信息抽取技术的重要性和实用性，使其同时得到了研究界和工业界的广泛关注。信息抽取技术自身的发展也大幅度推进了中文信息处理研究的发展，迫使研究人员面向实际应用需求，开始重视之前未被发现的研究难点和重点。纵观信息抽取研究发展的态势和技术现状，本文认为信息抽取的发展方向如下：

(1) 高效的小样本学习能力

目前的小样本学习设定需要用一个巨大的训练集训练的，测试时只给出 N-way K-shot，在这 N*K 个样本上学习并预测。真实场景下的小样本学习不存在巨大的训练集，从 GPT3 开始，预训练-提示(Prompt)学习范式受到研究者的关注，该范式将下游任务也建模成语言模型任务，在只给出几条或几十条样本作为训练集，借助与大规模预训练语言模型中蕴含的大量知识，取得了不错的小样本学习效果取得了。此外，相对于传统的 Pretrain+Finetune 范式，Prompt 有得天独厚的，可以摆脱指数级的预训练参数量对巨大计算资源的需求，高效的利用预训练模型。基于上述分析，本文认为信息抽取的发展方向之一利用预训练—提示学习范式进行高效的小样本学习。具体包括：1）提示学习中信息抽取任务模板的设计；2）模板的自动学习与挖掘；3）预训练-提示学习范式进行信息抽取的理论分析。

(2) 多模态信息融合

目前信息抽取主要针对的是纯文本数据，而常见的文档具有多样的布局且包含丰富的信息，以富文本文档的形式呈现包含大量的多模态信息，从认知科学的角度来说，人脑的感知和认知过程是跨越多种感官信息的融合处理，如人可以同时利用视觉和听觉信息理解说话人的情感、可以通过视觉信息补全文本中的缺失信息等，信息抽取技术的进一步发展也应该是针对多模态的富文档。基于上述分析，本文认为信息抽取的发展方向之一是多模态信息的融合。具体包括：1）多模态预训练模型的设计；2）多模态信息抽取框架中跨模态对齐任务设计；3）多模态信息的提取和表示。

(3) 数据驱动和知识驱动融合

现有的神经网络信息抽取方法依靠深度学习以数据驱动的方式得到各种语义关系的统计模式，其优势在于能从大量的原始数据中学习相关特征，比较容易利用证据和事实，但是忽略了怎样融合专家知识。单纯依靠神经网络进行信息抽取，到一定准确率之

后，就很难再改进。从人类进行知识获取来看，很多决策的时候同时要使用先验知识以及证据。数据驱动和知识驱动结合是模拟人脑进行信息抽取的关键挑战。基于上述分析，本文认为信息抽取的发展方向之一是构建数据驱动和知识驱动融合抽取技术。具体包括：1）神经符号学习信息抽取框架的构建；2）学习神经网络到逻辑符号的对应关系；3）神经网络对于符号计算过程进行模拟。

3.3.5. 基于知识的自然语言处理

3.3.5.1. 任务定义和目标

基于知识的 NLP，是指利用人类各类型结构化知识（如语言知识图谱、世界知识图谱、常识知识图谱等）提升 NLP 模型语言处理能力的相关处理方法。通过融合符号表示的人类结构化知识及其带来的认知推理能力，赋予语言深度学习模型更好的可解释性与认知推理能力，突破当前 NLP 领域中广泛使用的深度学习技术所面临的可解释性差、可扩展性差和鲁棒性差等瓶颈问题。

3.3.5.2. 研究内容和技术现状

完成知识图谱到 NLP 深度学习模型的融合，涉及知识表示学习、融合知识的预训练语言模型等关键技术。

（1）面向 NLP 的知识表示学习（KRL）

离散符号表示的知识图谱，在计算上存在计算效率低下和数据稀疏等挑战问题。近年来，人们提出了基于深度学习的 KRL 的技术方案，并被广泛研究与应用。

语言知识图谱的 KRL：语言知识图谱，描述的是以形式化和结构化语言表达的语言学知识，可以轻松植入各种 NLP 系统，代表性有 HowNet、WordNet 等。词表示学习是许多 NLP 任务的基础步骤，代表性方法有 Word2Vec、GloVe 等，但这些方法都是将每个词映射成一个向量，不能够解决一词多义的问题。为解决该问题，许多学者提出利用语言知识图谱指导的词表示学习，通过其细粒度语言学知识增强词的语义表示。例如，1）基于 HowNet 义原编码的词表示学习方法(SE-WRL)[81]，将每个词看成一组义原的组合，将词义消歧和融合义原、义项、词的 Skip-gram 词表示学习进行联合建模。2）将词向量改造为语义词典的 Retrofitting 方法[82]，给出了通过鼓励链接词具有相似的向量表示来使用 WordNet 等语义词典中的关系信息来细化向量空间表示。近几年随着基于预训练

模型的背景表示学习的兴起，相关研究开始聚焦于如何利用语言知识图谱增强词的上下文表示。

世界知识图谱的 KRL：世界知识图谱，指以结构化符号表示的实体及其关系的知识库，代表性有 WikiData、DBpedia 等，其表示学习的核心问题是学习实体和关系的低维分布式表示。相关研究围绕的核心问题有：1）如何度量事实三元组的合理性；2）何种编码模型建模关系交互；3）如何融合异构信息。

度量函数，用于衡量事实的合理性。目前有两种典型的度量函数：1）基于距离的度量函数，通过计算实体之间的距离来衡量事实的合理性，其中$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$关系的上平移被广泛使用，代表方法有 TransE[83]、TransH[84]、TransR[85]等。2）基于语义相似性的度量函数，通过语义匹配来衡量事实的合理性。它通常采用乘法公式$\boldsymbol{h}^{\top}\boldsymbol{M}_r\approx\boldsymbol{t}$，代表方法有 RESCAL[86]、DistMult[87]、ComplEx[83]等。

编码模型，即对实体和关系的交互编码使用的具体模型架构，包括线性/双线性模型、分解模型和神经网络模型。线性模型通过将头部实体投影到靠近尾部实体的表示空间中，将关系表述为线性/双线性映射，代表方法有 DistMult[87],ComplEx[83]等。分解模型旨在将关系数据分解为低秩矩阵以进行表征学习，代表方法有 RESCAL[86]、TuckER[88]等。神经网络模型通过用更复杂的网络结构对关系数据进行编码，如 R-GCN[89]、KG-BERT[90]等，其中 KG-BERT 借鉴 PLM 思想，用 BERT 作为实体和关系的编码器。

异构信息，在知识图谱中除了实体和关系本身信息之外，还包含其他类型信息，如文本描述、实体属性、类别约束、关系路径、视觉信息等。利用这些额外信息增强实体和关系的知识语义表示，主要挑战在于异构信息编码和异构信息融合等问题。KEPLER[91]给出了预训练语言表示和知识表示联合学习的统一模型，如图 2 所示，其通过联合学习不仅能够将事实知识信息更好的嵌入到预训练语言模型中，同时通过预训练语言模型可以得到文本语义增强的知识表示。

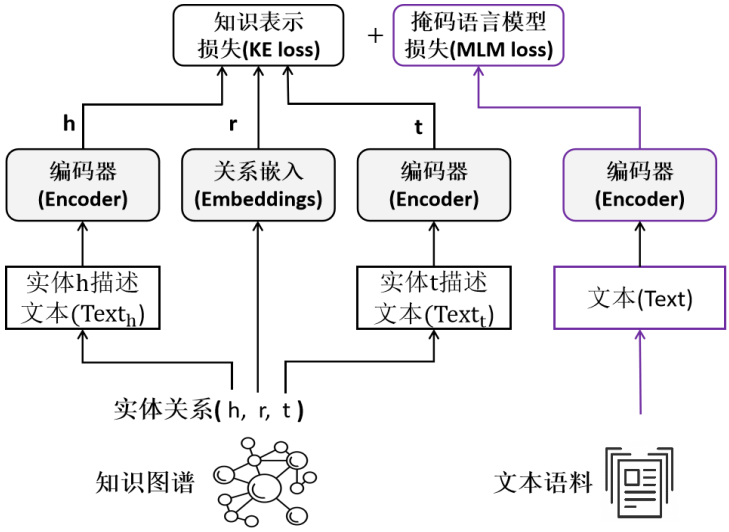


图 2 KEPLER 模型框架

（2）融合知识的预训练语言模型（PLM）

目前 PLM 主要采用互联网获取的海量通用文本语料训练得到，实现了对文本丰富语义模式的编码，但由于没有自觉运用结构化知识，依然严重缺乏知识运用和推理能力，缺乏可解释性和鲁棒性。为此，许多学者研究了融合结构化知识的 PLM 及其学习框架[92]，融合方法大致分为以下 4 种[93]：

知识增广：从输入端增强模型，有两种主流的方法：一种方式是直接把知识加到输入，另一方法是设计特定模块来融合原输入和相关的知识化的输入表示。目前，基于知识增广的方法已经在不同任务上取得良好效果，如信息检索[94]、问答系统[95]和阅读理解[96]。

知识支撑：关注于对带有知识的模型本身的处理流程进行优化。一种方式是在模型的底部引入知识指导层来处理特征，以便能得到更丰富的特征信息。例如，使用专门的知识记忆模块来从 PLM 底部注入丰富的记忆特征[97]。另一方面，知识也可以作为专家在模型顶层构建后处理模块，以计算得到更准确和有效的输出。例如，利用知识库来改进语言生成质量[98]。

知识约束：利用知识构建额外的预测目标和约束函数，来增强模型的原始目标函数。例如，远程监督学习利用知识图谱启发式标注语料作为新的目标，并广泛用于系列 NLP 任务，如实体识别[99]、关系抽取[100]和词义消歧[101]。或者利用知识构建额外的预测目标，比如 ERNIE[79]，CoLAKE[102]和 KEPLER[91]等工作，都是在原始的语言建模之外构建了相应额外的预训练目标。

知识迁移：则是从参数空间进行考量，获取一个知识指导的假设空间，从而让模型更有效。迁移学习和自监督学习分别关注从标注数据和无标注数据获取迁移学习和自监督学习分别关注从标注数据和无标注数据获取知识。作为一个迁移模型知识的典型范式，微调 PLM 在绝大多数 NLP 任务都可以取得良好的效果。在中文信息处理领域，一些中文 PLM 也相继被提出，如 CPM-1[103]、CPM-2[104]、PanGu-α[105]等，也都在各种中文任务中展现了良好性能。

3.3.5.3. 技术展望和发展趋势

结合国内外相关的研究工作，下面概括性地总结基于知识的 NLP 的技术趋势。一方面，面向 NLP 的深度学习技术能够自动学习语义的分布式表示，表达能力强，已在 NLP 多项重要任务中得到充分验证，为进一步融入知识指导信息的方法研究奠定了坚实基础。另一方面，知识表示与推理技术已经初步具备完整的方法体系，充分利用人类各

类型结构化知识赋予了人工智能不同的能力，为提升模型的可扩展性和鲁棒性提供了支撑。

尽管相关研究进展显著，但部分工作还非常初步，仍然有很多关键问题亟待解决，本文认为以下研究问题值得关注：

更大规模的知识表示：虽然已经出现了 GraphVite[106]、OpenKE[107]、DGL-KE[108]等系统工具，但这些工具还主要针对小规模知识图谱，这限制了大规模知识图谱的应用潜力。目前知识图谱的规模越来越大，如 Wikidata[109]已经含有了超过 9 千万实体、14.7 亿的关系，而且这种规模仍然呈现快速增长趋势。如何将现有知识表示学习方法适配到亿级实体规模的图谱上仍然是一个挑战。

PLM 的多元知识融合：目前在 PLM 中融合知识主要是围绕实体、实体关系等相关事实知识图谱，融合的知识类型和知识层次还比较单一，存在知识指导融合度低的问题。面向人类不同层次不同类型的丰富知识体系，探索融合这些多层次多类型知识的 PLM 框架和学习机制，是 PLM 技术未来研究的重要方向。

PLM 的持续知识增强：虽然 PLM 模型已经在多项任务上取得了超越人类的表现，但是现在 PLM 的模型通用智能水平增长仍遇到瓶颈。在可以预见的未来，PLM 模型的性能将持续增长。如何持续学习新知识、新数据提升模型语言处理能力，建立高效的知识持续植入的 PLM 学习机制，是 PLM 的关键研究方向。

PLM 的可靠知识编辑：PLM 在训练中需要事实知识并将其存储在模型参数中，以用于下游各种任务等，但大量事实知识存在时效性，随着时间推移可能会存在不准确或过时的问题。开发可靠的、无需重新训练的高效方法来修正模型中对应知识，是实现高质量可靠的 PLM 的关键问题[110]。

3.4.领域产业发展现状及趋势

近 5 年来自然语言处理相关任务在产业界有了更广泛的应用。特别是 2018 年以来，以 Elmo、BERT、GPT、ERNIE、悟道等为代表的大规模预训练模型在几乎全部自然语言任务上都取得了远超传统方法的卓越的性能，产业界对于自然语言处理算法研究的投入和应用也更加深入。通过对中文语言理解测评基准(CLUE)评测的排行榜以及近几年在 ACL、EMNLP、COLING 等会议上工作的统计，可以看到包括百度、腾讯、阿里巴巴、华为、科大讯飞、搜狗、字节跳动、快手、京东、美团、小米等小米等众多公司都在自然语言处理领域都投入了大量的研究和工程力量，取得了丰富的研究和应用成果，将自然语言处理算法大规模的应用于金融、教育、医疗、互联网、制造业等各类型行业。在抗击疫情方面，智能问诊、智能咨询、防疫问答、疫情预警、心理健康等自然语言处理

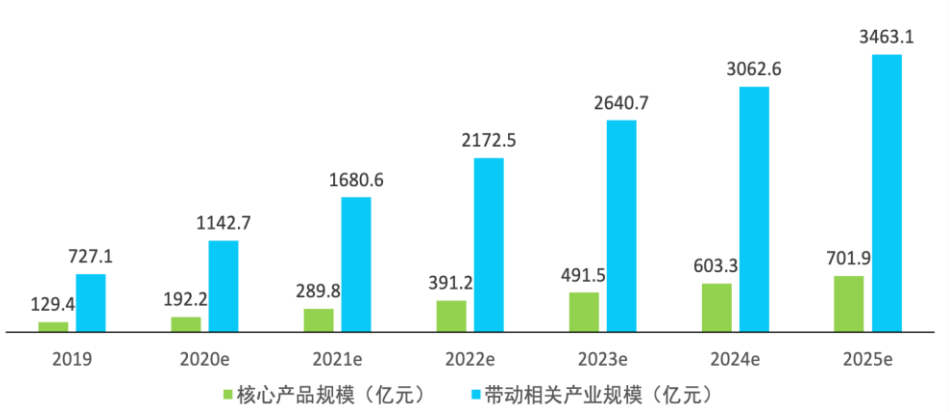
技术和系统也发挥了重要的作用。

自然语言处理技术真正进入到如何与产业更广泛和深入融合的前夜。现阶段自然语言处理产业发展具备以下几个主要特点：

1. 各行业智能化需求强劲，自然语言处理应用更加广泛

近几年各行业对智能化业务处理要求的上涨，加速了自然语言处理技术从互联网行业到传统行业的融合。在**法律领域**，自然语言处理技术在案例检索、法律条文推荐、判决预测、法律文本翻译、自动文书生成、合规审查等众多方面有效的辅助司法工作者提升案件处理效率和准确性。在**医疗领域**，医疗决策支持、医疗内容识别、病例检查、医学报告撰写、健康管理、自动导诊等方面也深度集成了自然语言处理技术。在**金融领域**，自然语言处理技术为量化投资提供了热点挖掘、舆情分析、事件驱动分析等多项重要因子，基于内容的用户画像、标签抽取等技术为为大数据风控提供了重要支持，此外自动客服、智能投顾等产品也广泛应用于整个行业。在**教育领域**的教、学、管、考等环节自然语言处理都发挥了重要作用，包括智能助教、智能批改、学情分析、自适应学习、机器组卷、教学分析等技术全方位提升教育教学的智能化程度，是实现因材施教的重要技术保障。在**制造业领域**，以自然语言处理技术为核心的 RPA（机器人流程自动化）等产品已经成为企业数字化转型的重要工具，在企业的法务、电商、财务、供应链、HR 等方面都大规模开展应用。

随着行业应用的不断深入，自然语言处理的产业规模也在快速增长。根据艾瑞咨询 2020 年发布的《中国人工智能产业研究报告（2020）》[111]数据，如图 3 所示，2020 年我国自然语言处理相关产品的市场规模达到 192.2 亿元，带动相关产业经济规模达 1142.7 亿元，预计 2025 年带动相关产业经济规模可以达到 3463.1 亿元，2019 到 2025 年核心产品的年均复合增长率可达到 30.8%。自然语言处理的整体产业规模，相较于 2015 年 22.4 亿元的中国自然语言处理市场规模有了大幅提升（数据来源《中国人工智能产业白皮书》，德勤研究，2018）。未来，随着自然语言处理通用能力的提升，相关应用范围和产业规模还会有更快速的增长。



数据来源:《中国人工智能产业研究报告(2020)》—艾瑞咨询，合并人机交互产品+知识图谱+NLP 核心产品

图 3 2019-2025 年中国 NLP 核心产品及带动相关产业规模

2. 自然语言处理算法范式快速更迭，企业优势明显发挥更大作用

2018 年在艾伦人工智能研究所(AI2)发布的 ELMo[112]以及 Google 发布的基于 Transformer 的 BERT[43]预训练模型之后，基于超大规模预训练的算法已经成为自然语言处理算法的新范式。2020 年 OpenAI 发布了包含 1750 亿参数的 GPT-3[44]。2021 年 1 月 Google 推出了包含 1.6 万亿参数的 Switch Transformer[113]模型，6 月北京智源研究院发布了“悟道 2.0”，参数量更是达到 1.75 万亿，创下当前全球最大预训练语言模型记录。2018-2021 年大规模预训练模型规模如图 4 所示，从中，可以看到国内的各大公司和研究机构也都先后投入大量资源开展相关研究。百度在 2021 年推出的文心 3.0（ERNIE 3.0）[114]模型参数量达到百亿，腾讯在 2021 年先后发布了具有百亿参数的“神舟“和十亿参数的“神农”两个模型，阿里达摩院也于 2021 年发布了具有万亿参数规模的 M6 模型[115]。这些研究推动着基于预训练模型的自然语言处理算法快速进步的同时，也注意到，由于超大规模模型训练所需的计算量和花费十分巨大，（按照 Nvidia 公司在其论文中的估算[116]，万亿规模模型训练通常需要花费数千万人民币的费用），企业成为了引领超大规模模型发展的主力军。同时，随着模型参数量的指数级增大，在超大规模预训练模型基础上进行微调的成本也大幅度上涨。因此，近两年在如何不改变预训练模型参数的同时完成各类任务的研究也开始越来越多的受到业界关注，包括：Prompt Tuning，Prefix-Tuning，P-Tuning 等。自然语言处理研究的范式在快速更迭中。

3. 自然语言处理平台商业化探索加速

现阶段，受制于自然语言处理算法能力的限制，相关技术需要与应用和产品相结合才能进行较好的商业化。自然语言处理涉及的环节多，相同的任务在处理不同领域的数据时，需要使用不同的标注数据甚至采取不同的模型框架进行完成，这严重制约了自然语言处理技术的直接商业化。虽然难度很大，但是近几年业界还是在自然语言处理技术直接商业上进行多种类型探索。

目前业界针对自然语言处理平台商业化的普遍做法是针对不同类型的用户和应用场景提供不同层面能力，从而尽可能的实现通用化和产品化。面向应用于新闻分析等通用领域，同时对自然语言处理算法了解较少的用户，提供 API 方式直接调用，使得用户快速获取包括分词、词性标注、命名实体、情感分析、中心词提取等自然语言处理处理能力。百度 AI 开放平台、阿里云达摩院 NLP 基础平台、科大讯飞开放平台、腾讯文智

开放平台等都提供该种方式的能力。在此基础上，针对具有一定使用经验并希望针对特定领域进行一定优化的用户，提供可以针对行业自适应标注、训练和服务等功能的平台也开始逐渐增多。百度 NLP 定制平台、阿里云达摩院 NLP 自学习平台、科大讯飞 AI+能力平台等都属于这个类型。对于具有自然语言处理算法模型开发能力的用户，很多公司也提供了开发服务平台，提供从数据管理、模型构建、模型管理、模型部署与服务等功能。百度飞桨 BML 全功能 AI 开发平台，阿里云机器学习平台 PAI、腾讯云智能钛机器学习平台等平台提供了这种能力。可以看到产业界在将自然语言处理技术与各行业深入融合的同时，也在自然语言处理平台产品化和商业化上也投入了大量资源。

经过这些年的发展，自然语言处理技术已经被越来越多企业应用，成为企业的核心竞争力，也有越来越多的企业在自然言处理技术上进行大量投入，自然语言处理技术在准确性上也取得了飞跃性发展。但是还要清楚的认识到，语言代表着人类的思想，人类的语言很复杂，真正的自然语言理解还有很长的路要走。目前的神经网络模型需要依赖大规模标注数据，模型鲁棒性亟待提高，同时还缺乏可解释性，也不具备显式的推理能力。如何融合机器学习与逻辑推理使其协同工作，充分利用数据和知识，使得机器真正理解人类语言，孕育着巨大的发展机遇。

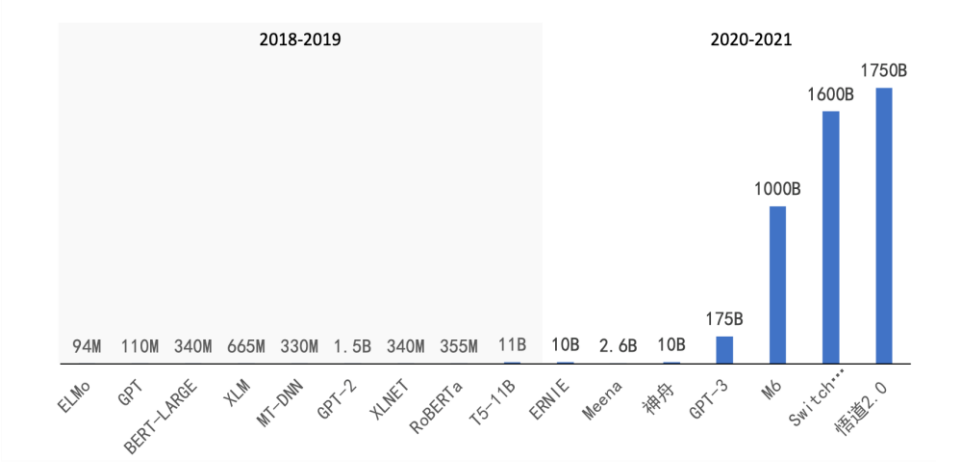


图 4　　2018-2021 年大规模预训练模型规模

3.5.总结及展望

本文首先介绍了自然语言处理（又称计算语言学）的定义、研究背景和意义，作为重要的认知智能任务，自然语言处理仍然面临众多问题，是目前制约人工智能取得更大突破和更广泛应用的瓶颈之一，被誉为“人工智能皇冠上的明珠”。接着介绍了自然语言处理的发展历史、现状与关键科学问题，其中“同质化”是自然语言处理历史发展的重要

趋势，并且随着模型规模越来越大，其“涌现”出了令人惊讶的“智能”。然后，重点从自然语言处理的范式迁移、词法句法分析、语义分析、信息抽取和基于知识的自然语言处理等五方面介绍了自然语言处理领域近五年（2017-2021）的发展情况。最后对自然语言处理产业发展现状及趋势进行了总结和展望。

未来，随着预训练模型等技术的进步，预计自然语言处理“同质化”的趋势将更加明显，并首先体现在跨模态数据上。基于 Transformer 的序列建模方法以及预训练模型在被成功应用于自然语言后，目前已在图像、视频、语音、表格数据、蛋白质序列、有机分子等数据上取得了优异的效果。这也使得未来构建一套**统一各种模态**的大规模预训练模型成为可能。

虽然预训练模型只是迁移学习的简单应用，但是其涌现出了令人惊讶的“智能”，其中“规模化”是必不可少的条件。正是由于规模化的重要性，越来越多的科研机构不断推出规模越来越大的预训练模型。不过，考虑到计算资源的限制以及大规模预训练模型训练时产生的大量碳排放对环境的影响，研制**更加高效**的预训练模型将是未来研究的重要方向。

最后，虽然基于深度学习的预训练模型显著提升了自然语言处理性能，但仍然存在鲁棒性差、可解释性弱、推理能力严重不足等瓶颈问题。其根本原因在于现有模型尚缺乏对人类知识的自觉运用机制。因此通过获取和利用**大规模多元知识**，有效提高自然语言处理的能力也是未来重要的发展方向。

计算语言学专委会在中国中文信息学会的领导下，将力争显著提升我国自然语言处理的研究水平，支撑我国信息技术向智能化发展，提高国民经济、信息处理等相关行业非结构化大数据处理及人工智能的创新能力和市场竞争力，将有力助推我国占领下一代信息技术和知识经济的科技制高点。

3.6.参考文献

[1] LIN T, WANG Y, LIU X, et al. A survey of transformers[J]. arXiv preprint arXiv:2106.04554, 2021.

[2] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J/OL]. SCIENCE CHINA Technological Sciences, 2020, 63(10):1872–1897. DOI: 10.1007/s11431-020-1647-3.

[3] SUN T, LIU X, QIU X, et al. Paradigm shift in natural language processing[J]. arXiv preprint arXiv:2109.12575, 2021.

[4] YANG P, SUN X, LI W, et al. SGM: Sequence generation model for multi-label classification[C/OL]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018:

3915-3926. https://aclanthology.org/C18-1330.
[5] SUN C, HUANG L, QIU X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 380-385. https: //aclanthology.org/N19-1035. DOI: 10.18653/v1/N19-1035.
[6] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 5849-5859. https://aclanthology .org/2020.acl-main.519. DOI: 10.18653/v1/2020.acl-main.519.
[7] YAN H, GUI T, DAI J, et al. A unified generative framework for various NER subtasks[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 5808-5822. https://aclanthology.org/2021.acl-long.451. DOI: 10.18653/v1/2021.acl-long.451.
[8] MAO Y, SHEN Y, YU C, et al. A joint training dual-mrc framework for aspect based sentiment analysis[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15):13543-13551. http s://ojs.aaai.org/index.php/AAAI/article/view/17597.
[9] YAN H, DAI J, JI T, et al. A unified generative framework for aspect-based sentiment analysis[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 2416-2429. https://aclanthology.org/2021.acl-long .188. DOI: 10.18653/v1/2021.acl-long.188.
[10] ZENG X, ZENG D, HE S, et al. Extracting relational facts by an endto-end neural model with copy mechanism[C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 506-514. https://aclanthology.org/P 18-1047. DOI: 10.18653/v1/P18-1047.
[11] LEVY O, SEO M, CHOI E, et al. Zero-shot relation extraction via reading comprehension[C/OL]//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, 2017: 333342. https://aclanthology.org/K17-1034. DOI: 10.18653/v1/K17-10 34.
[12] MCCANN B, KESKAR N S, XIONG C, et al. The natural language decathlon: Multitask learning as question answering[Z]. [S.l.: s.n.], 2018.
[13] ZHONG M, LIU P, CHEN Y, et al. Extractive summarization as text matching[C/OL]//Proceedings of the 58th Annual Meeting of the AssociationforComputationalLinguistics. Online: AssociationforComputational

Linguistics, 2020: 6197-6208. https://aclanthology.org/202 0.acl-main.552. DOI: 10.18653/v1/2020.acl-main.552.
[14] GAN L, MENG Y, KUANG K, et al. Dependency parsing as mrcbased span-span prediction[Z]. [S.l.: s.n.], 2021.
[15] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [Z]. [S.l.: s.n.], 2021.
[16] WANG X, PHAM H, YIN P, et al. A tree-based decoder for neural machine translation[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4772-4777. https://aclanthology.org/D18-1509. DOI: 10.18653/v1/D18-1509.
[17] XU J, DURRETT G. Neural extractive text summarization with syntactic compression[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 3292-3303. https://aclanthology.org/D19-1324. DOI: 10.18653/v1/D19-1324.
[18] ZHANG Y, CLARK S. Joint word segmentation and POS tagging using a single perceptron[C/OL]//Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics, 2008: 888-896. https://aclanthology.org/P08-1101.
[19] ZHANG Y, CLARK S. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model[C/OL]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA: Association for Computational Linguistics, 2010: 843-852. https://aclanthology.org/D10-1082.
[20] WANG H, YANG J, ZHANG Y. From genesis to creole language: Transfer learning for singlish universal dependencies parsing and pos tagging[J/OL]. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 2019, 19(1). https://doi.org/10.1145/3321128.
[21] TIAN Y, SONG Y, AO X, et al. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 8286-8296. https://aclanthology.org/2020. acl-main.735. DOI: 10.18653/v1/2020.acl-main.735.
[22] AKBIK A, BLYTHE D, VOLLGRAF R. Contextual string embeddings for sequence labeling[C/OL]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 16381649. https://aclanthology.org/C18-1139.
[23] CUI L, ZHANG Y. Hierarchically-refined label attention network for sequence labeling[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 4115-4128. https://aclanthology.org/D19-1422. DOI:
[24] 10.18653/v1/D19-1422.
[25] LIU J, ZHANG Y. In-order transition-based constituent parsing[J/OL].

[26] Transactions of the Association for Computational Linguistics, 2017, 5:413-424. https://aclanthology.org/Q17-1029. DOI: 10.1162/tacl_a _00070.
[27] CHEN W, ZHANG Y, ZHANG M. Feature embedding for dependency parsing[C/OL]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014: 816-826. https://aclanthology.org/C14-1078.
[28] DOZAT T, MANNING C D. Deep biaffine attention for neural dependency parsing[C/OL]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. https://openreview.n et/forum?id=Hk95PK9le.
[29] KITAEV N, KLEIN D. Constituency parsing with a self-attentive encoder[C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 2676-2686. https://www.aclweb.org/anthology/P18-1249. DOI: 10.18653/v1/P18-1249.
[30] ZHOU J, ZHAO H. Head-Driven Phrase Structure Grammar parsing on Penn Treebank[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 2396-2408. https: //aclanthology.org/P19-1230. DOI: 10.18653/v1/P19-1230.
[31] MRINI K, DERNONCOURT F, TRAN Q H, et al. Rethinking selfattention: Towards interpretability in neural parsing[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online: AssociationforComputationalLinguistics, 2020: 731-742. https: //aclanthology.org/2020.findings-emnlp.65. DOI: 10.18653/v1/2020 .findings-emnlp.65.
[32] ZHANG Y, ZHOU H, LI Z. Fast and accurate neural CRF constituency parsing[C/OL]//Proceedings of IJCAI. 2020: 4046-4053. https://doi. org/10.24963/ijcai.2020/560.
[33] YANG K, DENG J. Strongly incremental constituency parsing with graph neural networks[C]//Neural Information Processing Systems (NeurIPS). [S.l.: s.n.], 2020.
[34] YANG L, ZHANG M, LIU Y, et al. Joint pos tagging and dependence parsing with transition-based neural networks[J/OL]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 26. DOI: 10.1109/TASLP.2017.2788181.
[35] MÄRZL,TRAUTMANND,ROTHB. Domainadaptationforpart-ofspeech tagging of noisy user-generated text[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for
[36] Computational Linguistics: Human Language Technologies, Volume
[37] 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 3415-3420. https://aclanthology.o rg/N19-1345. DOI: 10.18653/v1/N19-1345.
[38] YASUNAGA M, KASAI J, RADEV D. Robust multilingual partof-speech tagging via adversarial training[C/OL]//Proceedings of the 2018 Conference of the North American

Chapter of the Association for
[39] Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for ComputationalLinguistics, 2018: 976-986. https://aclanthology.org/N18-1089. DOI: 10.18653/v1/N18-1089.
[40] NIVRE J, DE MARNEFFE M C, GINTER F, et al. Universal
[41] Dependencies v2: An evergrowing multilingual treebank collection [C/OL]//Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020: 4034-4043. https://aclanthology.org/2020.lrec-1.497.
[42] ZHANG M, ZHANG Y, FU G. Cross-lingual dependency parsing using code-mixed TreeBank[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 997-1006. https://aclanthology.org/D19-109 2. DOI: 10.18653/v1/D19-1092.
[43] ZHANG M, ZHANG Y. Cross-lingual dependency parsing via selftraining[C/OL]//Proceedings of the 19th Chinese National Conference on Computational Linguistics. Haikou, China: Chinese Information Processing Society of China, 2020: 807-819. https://aclanthology.org /2020.ccl-1.75.
[44] JURAFSKY D, MARTIN J H. Speech and language processing (draft) [J]. preparation [cited 2020 June 1] Available from: https://web. stanford. edu/~ jurafsky/slp3, 2018.
[45] NAVIGLI R. Word sense disambiguation: A survey[J]. ACM computing surveys (CSUR), 2009, 41(2):1-69.
[46] HE L, LEE K, LEWIS M, et al. Deep semantic role labeling: What works and what's next[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2017: 473-483.
[47] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. [S.l.:
[48] s.n.], 2017: 5998-6008.
[49] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J/OL]. CoRR, 2014, abs/1409.0473. http://arxiv.org/abs/1409.0473.
[50] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of NAACL-HLT. [S.l.: s.n.], 2019: 4171-4186.
[51] BROWN T B, MANN B, RYDER N, et al. Language models are fewshot learners[J]. arXiv preprint arXiv:2005.14165, 2020.
[52] HAN X, ZHANG Z, DING N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021.
[53] DING N, HU S, ZHAO W, et al. Openprompt: An open-source framework for prompt-learning[J]. arXiv preprint arXiv:2111.01998, 2021. [47] DONG L, LAPATA M. Language to logical form with neural attention[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany:

Association for Computational Linguistics, 2016: 33-43. https://aclanthology.org/P16-1004. DOI: 10.18653/v1/P16-1004.
[54] JIA R, LIANG P. Data recombination for neural semantic parsing [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 12-22. https://aclanthology.org/P16-1002. DOI: 10.18653/v1/P16-1002.
[55] XIAO C, DYMETMAN M, GARDENT C. Sequence-based structured prediction for semantic parsing[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1341-1350. https://aclanthology.org/P16-1 127. DOI: 10.18653/v1/P16-1127.
[56] CHEN B, SUN L, HAN X. Sequence-to-action: End-to-end semantic graph generation for semantic parsing[C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 766-777. https://aclanthology.org/P 18-1071. DOI: 10.18653/v1/P18-1071.
[57] YIN P, NEUBIG G, YIH W T, et al. TaBERT: Pretraining for joint understanding of textual and tabular data[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 84138426. https://aclanthology.org/2020.acl-main.745. DOI: 10.18653/v 1/2020.acl-main.745.
[58] HERZIG J, NOWAK P K, MÜLLER T, et al. TaPas: Weakly supervised table parsing via pre-training[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 4320-4333. https://aclanthology.org/2020.acl-main.398. DOI: 10.18653/v1/2020 .acl-main.398.
[59] YU T, WU C S, LIN X V, et al. Grappa: Grammaraugmented pre-training for table semantic parsing[J]. arXiv preprint arXiv:2009.13845, 2020.
[60] SHI P, NG P, WANG Z, et al. Learning contextual representations for semantic parsing with generation-augmented pre-training[J]. arXiv preprint arXiv:2012.10309, 2020.
[61] WU S, CHEN B, XIN C, et al. From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint ConferenceonNaturalLanguageProcessing(Volume1: LongPapers). Online: Association for Computational Linguistics, 2021: 5110-5121. https://aclanthology.org/2021.acl-long.397. DOI: 10.18653/v1/2021 .acl-long.397.
[62] SHIN R, LIN C, THOMSON S, et al. Constrained language models yield few-shot semantic parsers[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 7699-7715. https://aclanthology.org/2021.emnlpmain.608.
[63] SCHOLAK T, SCHUCHER N, BAHDANAU D. PICARD: Parsing incrementally for

constrained auto-regressive decoding from language models[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 9895-9901. https://aclanthology.org/2021.emnlp-main.7 79.

[64] WANG Y, BERANT J, LIANG P. Building a semantic parser overnight[C]//Association for Computational Linguistics (ACL). [S.l.:

[65] n.], 2015.

[66] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. arXiv preprint arXiv:1910.10683, 2019.

[67] LEWIS M, LIU Y, GOYAL N, et al. Bart: Denoising sequence-tosequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.

[68] LU Y, LIN H, XU J, et al. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction [C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 2795-2806. https://aclanthology.org/2021.acl-long.217. DOI: 10.18653/v1/2021 .acl-long.217.

[69] LIU K. A survey on neural relation extraction[J]. Science China Technological Sciences, 2020:1-19.

[70] LIU K, CHEN Y, LIU J, et al. Extracting event and their relations from texts: A survey on recent research progress and challenges[J]. AI Open, 2020, 1:22-39.

[71] SARAWAGI S. Information extraction[M]. [S.l.]: Now Publishers Inc, 2008.

[72] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C/OL]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, 2012: 1201-1211. https://aclanthology.org/D12-1110.

[73] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C/OL]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014: 2335-2344. https://www.aclweb .org/anthology/C14-1220.

[74] ZHENG H, WEN R, CHEN X, et al. PRGC: Potential relation and global correspondence based joint relational triple extraction[C/OL]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 6225-6235. https://aclant hology.org/2021.acl-long.486. DOI: 10.18653/v1/2021.acl-long.486.

[75] HAN X, ZHU H, YU P, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation [C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium:

Association for Computational Linguistics, 2018: 4803-4809. https://aclanthology.org/D18-1514. DOI: 10.18653/v1/D18-1514.

[76] DING N, XU G, CHEN Y, et al. Few-NERD: A few-shot named entity recognition dataset[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 3198-3213. https://aclanthology.org/2021.acl-long.248. DOI: 10.18653/v1/2021.acl-long.248.

[77] GAO T, HAN X, ZHU H, et al. FewRel 2.0: Towards more challenging few-shot relation classification[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 6250-6255. https://aclanthology.org/D19-1 649. DOI: 10.18653/v1/D19-1649.

[78] BALDINI SOARES L, FITZGERALD N, LING J, et al. Matching the blanks: Distributional similarity for relation learning[C/OL]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 2895-2905. https://aclanthology.org/P19-1279. DOI: 10.18653/v1/P19-1279.

[79] QUIRK C, POON H. Distant supervision for relation extraction beyond the sentence boundary[C/OL]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 1171-1182. https://aclanthology.org/E17-1110.

[80] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 4925-4936. https://aclanthology.org/D19-1498. DOI: 10.18653/v1/ D19-1498.

[81] WANG D, HU W, CAO E, et al. Global-to-local neural networks for document-level relation extraction[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 3711-3721. https://aclanthology.org/2020.emnlp-main.303. DOI: 10.18653/v1/2020.emnlp-main.303.

[82] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document-level relation extraction[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 1630-1640. https://aclanthology.org/2020.emnlp-main.127. DOI: 10.18653/v1/2020.emnlp-main.127.

[83] ZHOU H, XU Y, YAO W, et al. Global context-enhanced graph convolutional networks for document-level relation extraction[C/OL]// Proceedings of the 28th International

Conference on Computational
[84] Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 5259-5270. https://aclanthology.o rg/2020.coling-main.461. DOI: 10.18653/v1/2020.coling-main.461.
[85] XU B, WANG Q, LYU Y, et al. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction[C/OL]//Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021: 14149-14157. https://ojs.aaai.org/index.php/AAAI/article/view/17665.
[86] ZHOU W, HUANG K, MA T, et al. Document-level relation extraction with adaptive thresholding and localized context pooling[C/OL]// Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021: 14612-14620. https://ojs.aaai.org/ind ex.php/AAAI/article/view/17717.
[87] ZHANG Z, HAN X, LIU Z, et al. ERNIE: Enhanced language representation with informative entities[C/OL]//Proceedings of the 57th AnnualMeetingoftheAssociationforComputationalLinguistics. Florence, Italy: Association for Computational Linguistics, 2019: 14411451. https://aclanthology.org/P19-1139. DOI: 10.18653/v1/P19-11 39.
[88] QIN Y, LIN Y, TAKANOBU R, et al. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 3350-3363. https://aclanthology.org/2021.acl-long.260. DOI: 10.18653/v1/2021.acl-long.260.
[89] NIU Y, XIE R, LIU Z, et al. Improved word representation learning with sememes[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2017: 2049-2058.
[90] FARUQUI M, DODGE J, JAUHAR S K, et al. Retrofitting word vectors to semantic lexicons[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association forComputationalLinguistics: Human Language Technologies. [S.l.: s.n.], 2015: 16061615.
[91] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Advances in neural information processing systems. [S.l.: s.n.], 2013: 926-934.
[92] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 28. [S.l.: s.n.], 2014.
[93] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph

completion[C]//Twenty-ninth AAAI conference on artificial intelligence. [S.l.: s.n.], 2015.
[94] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]//Icml. [S.l.: s.n.], 2011.
[95] YANG B, YIH W T, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.
[96] BALAŽEVIĆ I, ALLEN C, HOSPEDALES T. Tucker: Tensor factorization for knowledge graph completion[C]//Proceedings of the 2019 ConferenceonEmpiricalMethodsinNaturalLanguageProcessingand the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 5185-5194.
[97] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. [S.l.]: Springer, 2018: 593-607.
[98] YAO L, MAO C, LUO Y. Kg-bert: Bert for knowledge graph completion[J]. arXiv preprint arXiv:1909.03193, 2019.
[99] WANG X, GAO T, ZHU Z, et al. Kepler: A unified model for knowledge embedding and pre-trained language representation[J]. Transactions of the Association for Computational Linguistics, 2021, 9:176194.
[100] YANG J, XIAO G, SHEN Y, et al. A survey of knowledge enhanced pre-trained models[J]. arXiv preprint arXiv:2110.00269, 2021.
[101] HAN X, ZHANG Z, LIU Z. Knowledgeable machine learning for natural language processing[J]. Communications of the ACM, 2021, 64 (11):50-51.
[102] GUU K, LEE K, TUNG Z, et al. Retrieval augmented language model pre-training[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2020: 3929-3938.
[103] XIONG W, DU J, WANG W Y, et al. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model[C]// International Conference on Learning Representations. [S.l.: s.n.], 2019.
[104] XU H, LIU B, SHU L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis[C/OL]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 2324-2335.
[105] https://aclanthology.org/N19-1242. DOI: 10.18653/v1/N19-1242.
[106] DING M, ZHOU C, YANG H, et al. Cogltx: Applying bert to long texts[J]. Advances in Neural Information Processing Systems, 2020, 33:12792-12804.
[107] GU Y, YAN J, ZHU H, et al. Language modeling with sparse product of sememe experts[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4642-4651.
[108] https://aclanthology.org/D18-1493. DOI: 10.18653/v1/D18-1493.
[109] XIN J, LIN Y, LIU Z, et al. Improving neural fine-grained entity typing with

knowledge attention[C]//Thirty-second AAAI conference on artificial intelligence. [S.l.: s.n.], 2018.
[110] HAN X, LIU Z, SUN M. Neural knowledge acquisition via mutual attention between knowledge graph and text[C]//Thirty-second AAAI conference on artificial intelligence. [S.l.: s.n.], 2018.
[111] HUANG L, SUN C, QIU X, et al. Glossbert: Bert for word sense disambiguation with gloss knowledge[J]. arXiv preprint arXiv:1908.07245, 2019.
[112] SUN T, SHAO Y, QIU X, et al. CoLAKE: Contextualized language and knowledge embedding[C/OL]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 3660-3670. https://aclanthology.org/2020.coling-main.32 7. DOI: 10.18653/v1/2020.coling-main.327.
[113] ZHANG Z, HAN X, ZHOU H, et al. Cpm: A large-scale generative chinese pre-trained language model[J]. AI Open, 2021, 2:93-99.
[114] ZHANG Z, GU Y, HAN X, et al. Cpm-2: Large-scale cost-effective pre-trained language models[J]. arXiv preprint arXiv:2106.10715, 2021.
[115] ZENG W, REN X, SU T, et al. Pangu-alpha: Large-scale autoregressive pretrained chinese language models with auto-parallel computation[J]. arXiv preprint arXiv:2104.12369, 2021.
[116] ZHU Z, XU S, QU M, et al. Graphvite: A high-performance cpu-gpu hybrid system for node embedding[C]//The World Wide Web Conference. [S.l.]: ACM, 2019: 2494-2504.
[117] HAN X, CAO S, LV X, et al. Openke: An open toolkit for knowledge embedding[C]//Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. [S.l.:
[118] n.], 2018: 139-144.
[119] ZHENG D, SONG X, MA C, et al. Dgl-ke: Training knowledge graph embeddings at scale[C]//SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery, 2020: 739–748.
[120] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10):78-85.
[121] DE CAO N, AZIZ W, TITOV I. Editing factual knowledge in language models[J]. arXiv preprint arXiv:2104.08164, 2021.
[122] 艾瑞咨询. 2020 年中国人工智能产业研究报告公开版[EB/OL].2020. https://report.iresearch.cn/report/202012/3707.shtml/.
[123] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT. [S.l.: s.n.], 2018: 2227-2237.
[124] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv preprint arXiv:2101.03961, 2021.
[125] SUN Y, WANG S, FENG S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2107.02137, 2021.

[126] LIN J, MEN R, YANG A, et al. M6: A chinese multimodal pretrainer [J]. arXiv preprint arXiv:2103.00823, 2021.
[127] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient largescale language model training on gpu clusters using megatron-lm[C]// Proceedings of the International Conference for High Performance
[128] Computing, Networking, Storage and Analysis. [S.l.: s.n.], 2021: 115.

第四章　少数民族语言文字信息处理研究进展、现状及趋势

国家历来重视少数民族语言文字信息化建设事业，为使各少数民族共享信息化时代的成果，制定了蒙古、藏、维吾尔、哈萨克、柯尔克孜、壮、朝鲜等少数民族文字编码字符集、键盘、字模等国家标准，研究开发了多种少数民族文字排版系统、智慧语音翻译系统，支持少数民族语言文字网站和新兴传播载体有序发展，不断提升少数民族语言文字的信息化能力和社会应用能力。近五年来，中国中文信息学会少数民族语言文字信息处理专业委员会以及相关高校、研究单位，在蒙古文、藏文、维吾尔文、哈萨克文、壮文、朝鲜文等少数民族语言文字信息处理方面取得了令人可喜的成绩。以下分文种阐述各语言文字信息处理的研究进展。

4.1.蒙古文语言文字信息处理研究进展

内蒙古大学、内蒙古社会科学院、呼和浩特民族学院等研究单位，以习近平总书记在内蒙古考察时的重要讲话、在内蒙古大学的重要指示精神和以习近平总书记关于教育的重要论述和对语言文字工作的重要指示，紧紧围绕“一带一路”边疆地区建设总目标，面向国家安全和区域发展战略需求，可持续性开展着基础研究和应用开发研究。

4.1.1.　内蒙古大学东北亚语言资源研究中心

1. 为了辐射多语言、多文种、多领域、多维度，内蒙古大学东北亚语言大数据中心围绕：（1）汉语、蒙古语、英语、俄语、日语、韩语等东北亚国家官方语言；（2）达斡尔、鄂温克、鄂伦春、布里亚特等小语种；（3）回鹘式蒙古文、八思巴文、托忒文、女贞文、突厥文等 10 多种古文字文献等建立了多文种资源库，它囊括语料库、词典库和文献库。语料库包括国内蒙古语单语语料库和多语语料库，蒙古国单语语料库和多语语料库以及东北亚其他语种语料库等；词典库包括了传统蒙古文词典、西里尔蒙古文词典和东北亚其他语种词典；文献库包括了回鹘式蒙古文、八思巴文、突厥文、女真文、契丹文、托忒文及其文献 3D 模型、文本化内容、研究论著目录及图形化内文等；

2. 基于上述资源，研发了多文种平行处理技术，跨文种智能搜索、多语种文本语音机器翻译、东北亚语言资源数字化博物馆、多模态语料库加工处理等，实现了多文种资源在机器翻译中的应用、跨文种搜索中的应用、语言教学中的应用、文化展示中的应用和知识表示中的应用。

3. 围绕语义、句法和语用处理需求和 OCR、文献处理研究工作，内蒙古大学近几

年陆续立项国家社科基金重大项目“回鹘式蒙古文文献数据库建设”和国家社科基金重点项目、一般项目等共计 8 项，在句法树库、语义知识库研发和熟语谜语研究方面开展着创新性研究工作。

4.1.2. 内蒙古社科科学院

1. 基础通用软件方面

（1）发布多文种办公软件 Onon Office，力求原生支持竖排编辑、竖排注释、竖排图表、竖排目标等特色版式，提供蒙古文按字形模糊查找与按音精确查找、自动校对、汉蒙翻译、新旧蒙古文互译等特色功能。

（2）发布 Onon 输入法新版本，特色在于输入法实现真正基于大规模语料的语言模型训练和基于语言模型的连续输入。

（3）发布 Onon 编码转换器，特色在于实现多数编码间转换理论值达到 100%，编码转换错误可预测水平，这将大力促进语料建设及数据兼容。

（4）所有基础通用软件适配国产操作系统，达到安可系统通用地步。

2. 行业应用软件

（1）多文种在线地图发布并持续更新服务，在基于二次开发的地名标注基础上，完成了自主地理信息系统的研发，并在特色数据标注，例如街面用字采集等方面进展突出。

（2）多文种文档分享平台持续更新服务，本年度主要在期刊碎片化加工及日报数据挖掘上进展突出。

3. 语料库建设

已完成 1 亿级单语语料加工，完成 7000 小时语音合成数据采集加工；完成 500 万句对对齐语料加工；完成回鹘式蒙古文文献平台建设工作。

4. 立项内蒙古社科院获得国家社科基金项目 1 项，内蒙古社科院获得内蒙古自治区关键技术攻关计划项目 2 项。

5．专利和著作权，斡仑科技获得外观专利 2 项；斡仑科技获得著作权 7 项等。

4.1.3. 呼和浩特民族学院

近年来极其重视中文信息处理相关研究工作，计算机科学与信息工程学院和蒙古学学院（翻译学院）等单位在学术研究、人才培养、合作交流等方面积极开展了工作。

学校立项重点项目“人工智能与知识图谱重点实验室”和“蒙古文智能计算与机器翻译”创新团队，“重点实验室”依靠内蒙古自治区蒙古文信息处理技术重点实验室，于

原基础上扩充组建了一支博士成员为核心骨干的科研队伍，学校提供了 90 平米的专用实验室、配备两台服务器、24 台工作站等设备，具备了较好的科学研究条件。“重点实验室”依托呼和浩特民族学院的研究基础和优势资源，围绕蒙汉英多语种语义网、蒙汉英文机器翻译、蒙古文舆情分析、蒙古语语音识别研究、蒙古语资源建设等方面开展一系列的探索与研究，并与学校其他相关资源平台交叉融合发展，短短的一年多时间取得了显著的成绩。

1.“重点实验室”研究团队已成为内蒙古自治区蒙古语言文字信息处理技术研究和人才培养的一支重要力量。近几年，“重点实验室”研究团队已发表学术论文 50 余篇，其中被 SCI（或 EI）收录 10 余篇，出版专著（教材）等 10 余部，承担国家自然科学基金项目 2 项、国家社会科学基金项目 2 项，自治区科技攻关项目、自治区自然基金项目、教育部人文社会科学规划项目、国家语委科研项目等省部级科研项目 12 项、教育部产学合作协同育人项目 5 项，发明专利（实用新型专利）和软件著作权共 6 项，近几年的科研经费累计超 500 万元。

2.“重点实验室”拥有 10 余项自主研发的软件系统。“重点实验室”研究团队在内蒙古自治区蒙古语言文字信息化专项扶持项目资金的资助下，成功研制了《蒙古文网站内容管理及信息发布系统》、《蒙古文网站实时追踪与信息自动采集系统》、《蒙古文 UNICODE 编码自动批处理转换软件》、《蒙古文命名实体与实体关系自动识别系统》、《蒙古文词法分析与词性自动标注系统》等 12 套软件系统，其中的部分研究成果被广泛应用于相关领域，产生了良好的社会效益。

3.“重点实验室”拥有大体量的蒙古语数字化的教育教学资源。“重点实验室”研究团队在内蒙古自治区蒙古语言文字信息化专项扶持项目资金的资助下，研制完成了“蒙古语有声资源库”（以下简称“资源库”）。截止到 2018 年底，共收录制作完成 2000 余小时的数字化的蒙古语有声资源，并于 2019 年上线（http://www.nm-cy.cn）运营，最高日访问量超 10000 人次。“资源库”还荣获全国民族院校优秀教学成果 A 级奖、内蒙古自治区民族教育优秀科研成果二等奖。以“资源库”为支撑的大学生创新创业项目“蒙 IT 驿站”和“乌力格尔”（故事会）分别荣获 2017 年全国“互联网+”大学生创新创业大赛铜奖，2017 年内蒙古自治区“互联网+”大学生创新创业大赛金奖、2019 年内蒙古自治区“互联网+”大学生创新创业大赛“青年红色筑梦之旅”赛道银奖。

4.“重点实验室”在蒙汉文机器翻译研究方面取得了新的重大进展。“重点实验室”研究团队从 2013 年开始收集整理蒙汉平行语料，2015 年在 10 万句平行语料的基础上搭建了基于短语的蒙汉统计机器翻译系统，2017 年 9 月在 40 万句平行语料的基础上搭建了基于深度学习技术的蒙汉神经机器翻译系统。该系统采用循环神经网络（RNN）的

编码器—解码器模型，与以往的机器翻译系统相比，翻译质量大大提高，基于神经网络的技术是该系统的一大突破，该系统作为 2017 年度内蒙古自治区蒙古语言文字信息化专项扶持重点项目“融合大数据与多语言开放域的蒙汉文知识图谱构建及其应用技术共享服务平台”的阶段性研究成果，受到同行专家的一致好评，并被多家主流媒体转载报道。此后，继续收集整理蒙汉平行语料，到 2020 年 1 月时语料规模达到 130 万句以上，并使用 Transformer 模型重新构建了翻译模型，以此模型为基础搭建了基于大规模语料的“蒙汉在线机器翻译系统”，在新闻、日常用语等领域翻译准确率达到了 90%以上。

5.“重点实验室”已成为国家语言资源监测与研究少数民族语言中心蒙古语研究基地。2014 年，学院获批了隶属于国家民委、国家语委和中央民族大学的“国家语言资源监测与研究少数民族语言中心”的蒙古语研究基地。这是全国第四个研究基地、首个蒙古语研究国家基地。“重点实验室”主要成员在中心+基地的模式下，与国内知名高校和科研院所合作完成了国家重点研究项目。如，与中央民族大学、清华大学、西藏大学联合承担国家自然科学基金重点项目“跨语言社会舆情分析基础理论与关键技术研究”，负责完成了“蒙古文舆情监测语料库构建”子项目的研发任务；与中国科学院和中央民族大学联合承担“蒙古语语音语言数据采集及处理方法研究与实现”，负责完成“蒙古语语音语言数据采集”子项目的研发任务。

6.近期取得的标志性成果可概述为：研制开发《蒙古秘史》全文检索平台（www.mnuuts.com）。该平台由呼和浩特民族学院计算机科学与信息工程学院金罡博士率领的团队研发完成。该平台将《蒙古秘史》中名词、形容、动词以同位关系、上下位关系、整体与部分关系组织为树形结构数据库，并将此数据库融入到《蒙古秘史》全文检索平台，为《蒙古秘史》语料库检索增加了语义检索功能。蒙古语语义网在语料库语言学或文本挖掘方面的另一个典型的应用是蒙古语五畜知识平台（www.wuchu.com）。该平台由呼和浩特民族学院布音其其格博士的团队研发完成。蒙古语的 （“五畜”）是指蒙古高原上主要经营的 （马）、 （骆驼）、 （牛）、 （绵羊） 和 （山羊）。蒙古语中五畜相关词汇非常丰富，比如关于五畜的岁数、性别名称、用具的名称、身体部位名称以及畜产品名称等相关五畜文化的词汇规模很大，并且区分程度也是非常细致。本系统搜集了五畜相关词汇，然后对词汇进行语义分类并完成词汇语义解释，常用词汇使用下位语义场方式提供了说明，最终以数据库形式存储了语义信息，旨在使五畜相关词汇较全面、系统、正确的保存，为使用者提供快捷方便的词汇语义查询学习服务。系统中除了五畜相关词汇以外还增加了五畜相关民间文学作品，例如谜语、谚语、神话故事、赞颂词、民歌歌词等内容，使系统中的知识更加丰富，后续将不断扩充完善数据，通过动态更新满足各类用户的知识学习使用需求。此外，“蒙古族儿童民间经典故事有

声资源库的构建及关键技术研究”、“蒙古语同形词语义词典”、“汉蒙机器翻译系统”、“蒙汉英语义网”等相关成果也陆续集成到实验室的研究中，努力打造资源、技术融合的开放平台。

4.1.4. 内蒙古师范大学蒙古文信息处理

4.1.4.1. 蒙汉机器翻译研究进展

在蒙汉机器翻译方面，结合国家自然基金项目和自治区自然基金项目，针对蒙古语言自身特征，结合机器翻译的最新方法,开展了基于神经网络的蒙汉机器翻译研究。

首先在基于 transformer 的蒙汉神经机器翻译模型上，将蒙古文词干，构形附加成分等形态信息融入到神经机器翻译模型；针对蒙汉神经机器翻译的未登录词问题，采取基于语义相似度的未登录词替换、基于语言模型的未登录词替换和基于蒙汉对齐词典的未登录词替换等三种方法进行了研究；结合蒙古文的特点研究了基于神经网络的蒙古文次切分方法和基于神经网络的蒙古文命名实体识别方法。

为了有效利用单语数据来提升蒙汉神经机器翻译的性能，提出了基于 BERT 数据增强的蒙汉神经机器翻译方法。同时研究了反向翻译方法产生的蒙汉伪平行语料对蒙汉神经机器翻译的影响。为了提高模型提取语言内部特征的能力，研究了基于降噪自编码器的蒙汉神经机器翻译方法。有效缓解在蒙汉神经机器翻译任务中平行语料库稀缺问题。

4.1.4.2. 蒙古文在线教育研究进展

结合内蒙古自治区科技计划重大项目、自治区民委的蒙古文语言文字专项基金项目、内蒙古自治区自然基金项目，针对蒙古文教育资源匮乏的现实，结合当前最先进的 MOOC 大规模在线教育技术，开展了蒙古文在线教育研究工作。

我们自主研发了国内首个蒙古文 MOOC 在线互动学习平台，包括平台学生端学习管理系统（LMS）和教师端课程内容管理系统（CMS）；建设完成 4 门蒙古文 MOOC 课程资源，资源包括授课课件、授课视频、习题库和试题库等；完成平台系统的部署、测试、维护、升级等日常运营，提供课程资源的服务；另外，研发了蒙古文单词学习自适应学习手机 APP 软件，包括蒙古文单词书写笔顺的动态演示子模块；依托我们的在线互动学习平台和 APP 软件，收集学习相关数据，进行了基于民族教育大数据的学习者学习风格模型构建研究。蒙古文 MOOC 平台及课程资源建设，在 2021 年获得内蒙古师范大学第十五届教学成果二等奖。

4.2.藏语语言文字信息处理研究进展

西藏大学、西北民族大学、青海师范大学等高校的研究机构以习近平总书记在西藏、青海考察时的重要讲话精神和以习近平总书记关于教育的重要论述和对语言文字工作的重要指示，紧紧围绕“一带一路”边疆地区建设总目标，面向国家安全和区域发展战略需求，可持续性开展着基础研究和应用开发研究。下面从所承担的科研项目、研发的平台建设及成果转化、发表的学术论文和国家标准、重点实验室建设和人才培养、获奖情况、举办的学术会议及最新研究领域等六个方面做工作进展总结，具体为：

4.2.1. 科研项目和资源建设

西藏大学目前承担了 2 项国家重点研发计划，另参与了 1 项国家重点研发计划。青海师范大学目前承担 1 项国家重点研发计划项目“公共文化服务装备研发及应用示范”（项目编号：2020YFC1523300）。 西藏大学完成的国家重点研发计划项目“藏文文献资源数字化技术集成与应用示范”（项目编号：2017YFB1402200）研发了全球首款藏文古籍版面分析与多字体文字扫描识别系统。应用该技术成体系地深度数字化了藏医药古籍 1000 余册、建立了藏医药古籍全文数据库、初步构建了藏医药知识图谱，为藏医药知识挖掘和产业发展提供了数据支撑，并应用于国家级布达拉宫藏文古籍数字化保护和利用等各类藏文文献数字化项目，深度数字化 4000 余册各类藏文古籍。

近年来西北民族大学依托所承担的国家科技支撑计划项目，构建了九年义务教育藏语网络教学资源，大学本科藏语言文学专业基础课和专业课网络教学资源，藏语农林牧科普教育网络教学资源等远程教育资源库。

4.2.2. 平台建设及成果转化

4.2.2.1. 阳光藏汉双向机器翻译系统

西藏大学集成自主研发的藏文自动分词和词性标注、藏汉双向机器翻译、跨语言搜索引擎等藏语自然语言处理技术成果，首次研发跨语言互联网藏文舆情分析技术及其软件系统，并工程化应用于西藏自治区互联网信息办公室、拉萨市互联网信息办公室等涉及国家安全的部门。“阳光藏汉双向机器翻译系统”面向社会服务，用户遍布 40 余个国家和地区，产生了广泛的社会影响。

4.2.2.2. 汉藏双语银河麒麟桌面操作系统

为加快推进国产自主可控替代计划，西藏大学与国防科技大学和麒麟软件有限公司合作研发了汉藏双语银河麒麟桌面操作系统。与金山公司合作，2021 年 10 月发布了国内首个纯国产化藏文版办公软件——藏文版 WPS Office 正式发布。银河麒麟桌面操作系统（藏文版）V10 发布，该产品集成了基于音节的藏语文本编辑距离计算方法与排序方法、藏文拼写检查，且支持汉藏、英藏文本混合编排，界面语言支持汉语、藏语、英语的灵活切换。藏文版操作系统的设计充分考虑藏族用户习惯，集实用性、高性能、创新性、可维护等优势于一体，满足藏文用户的多元化需求。

4.2.2.3. 国家通用语/藏语远程教育平台

今年来西北民族大学研发的国家通用语/藏语远程教育平台，是国家科技支撑计划项目“少数民族语言文字信息处理共性关键技术研究与示范应用”的科研成果。该平台是基于云平台构建，支持多终端访问，并在手持移动设备上开发了藏汉双语辅助教学系统，同时，完成了面向藏族的汉语普通话学习系统，平台包含多种宝贵的教学资源库。

4.2.2.4. “一带一路”特色农产品多语言电子商务平台

西北民族大学的“一带一路特色农产品多语言电子商务平台”来源于国家科技支撑计划课题“民族特色农产品多语言网络交易展示平台关键技术集成与应用示范”，课题主要目的是以互联网为依托，搭建基于云计算的民族特色产品多语言交易平台，以信息化技术将民族特色产品推向市场，提升民族地区特色产品的品牌竞争力，为民族地区培养信息化人才，以电子商务提升民族地区经济和社会发展，为民族地区经济社会进步作出积极的贡献。

“一带一路特色农产品多语言电子商务平台”被时任甘肃省省长唐仁健同志誉为“大宝贝”，被甘肃省政府总结为具有“丝路的民族的世界的，甘肃的中国的全球的”特征，成果的产业化对甘肃省乡村振兴、大数据产业发展和抢抓“一带一路”机遇等有重大价值，该成果在 2019 年、2020 年多次被甘肃省省长通过批示和专题会议的形式推进项目转化落地。并将此平台建设作为重点工作写入甘肃省政府 2019 年工作报告，2019 年 11 月 27 日被省政府列入《新时代甘肃融入“一带一路”建设打造信息制高点实施方案》，为响应省委省政府的指示精神，加快推进“一带一路特色农产品多语言电商平台”的落地转化，西北民族大学作为科研成果产出单位，省属国有企业丝绸之路信息港股份有限公司

作为成果产业落地实施单位，合作成立了科研成果转化公司，甘肃省省长为公司揭牌。

该平台是国内外第一个以汉语为核心的多语言平行对照电商平台，国内外第一个以“一带一路”沿线特色农产品为特定交易对象的电商平台，平台发挥西北民族大学民族语言学科优势，瞄准民族地区商贸交流存在语言障碍、信息服务薄弱、农产品销售困难的现实困境，尤其是聚焦目前国内外缺乏民族语言电子商务平台的瓶颈制约，助推特色农产品进入国内消费市场，打入国际市场，带动农村贫困地区经济发展，在此基础上，由特色农产品延伸扩展到不同领域的产品，为甘肃省乃至全国民族地区和“一带一路”沿线国家及世界范围内各民族地区加快发展提供了一条新路径。

平台自 2019 年起，有 500 多家企业受益，3000 余种特色农产品通过汉语/英语/藏语/蒙古语/维吾尔语五种语言在网销售。受到了国内媒体的持续关注，中国民族报，在第 2 届丝绸之路高峰论坛当天，在报纸头版以“为国际减贫提供中国智慧”为标题对平台做了大篇幅报道。2020 年、2021 年，央视网、中国新闻网、新华社等媒体，对于洪志教授及团队也做了专题报道：“女科学家研发多语言电商平台牵‘一带一路’做‘世界生意’”，“Multilingual e-commerce platform to help China's ethnic farmers in foreign trade”等。

4.2.2.5. 互联网+藏语网络信息处理平台

青海师范大学依据所承担的科技成果转化专项“互联网+藏语信息处理平台建设项目”（编号：2017-GX-146），研发了藏语网络信息处理平台，实现了藏语网络信息自动采集与分析子系统，建立了藏文搜索引擎，其网站种子数量占国内外藏文网站域名总数的91%以上，2 小时内完成对所有藏文文本数据和重新索引和排序。实现了藏汉(汉藏)语言自动翻译子系统，构建了 200 万句对的藏汉双语平行语料。通过构建藏汉双语情感词词典和藏语情感分类语料库，完成了基于神经网络的藏语文本情感分类方法研究，实现了藏语舆情分析及应用子系统。

4.2.3. 学术论文和著作

1. 据不完全统计，近五年来，在知网发表或 CCL、少民语言文字信息处理学术研讨会和少数民族青年自然语言处理学术研讨会等学术会议录用的藏文信息处理研究相关学术论文达 130—150 篇。

2. 制定国家标准 5 项：2018 年 6 月，西藏大学牵头，西北民大和青海师范大学为主要参与单位制定了《信息处理用藏文分词规范》（GB/T 36452-2018）和《信息处理用藏语词类标记集》（GB/T 36337-2018）两个国家标准；

同年，青海师范大学牵头，西藏大学和西北民族大学为主要参与单位制定了《信息处理用藏语短语分类与标记规范，GB_T36472-2018》和《信息技术 藏文字符排序规范》（GB/T 36335-2018）两个国家标准；西北民族大学牵头，西藏大学和青海师范大学为主要参与单位制定了《信息处理用藏文文献文本信息标记规范》（GB/T 36338-2018）国家标准。目前有 5 名教师任全国信息技术标准化委员会第三届字符集与编码分技术委员会委员，8 名教师任藏文信息技术国家标准工作组委员。

自 2017 年开始，与“网智天元科技集团股份有限公司”联合成立“藏大网智大数据研究中心”，与有关单位共同编制《西藏环保现状和环境变化趋势》决策咨询报告、发布《西藏大数据》内刊。

4.2.4. 重点实验室建设和人才培养

中华人民共和国科学技术部、青海省人民政府和青海省科学技术厅的大力支持下，经多次论证和会商，青海师范大学申报的“省部共建藏语智能信息处理及应用国家重点实验室”，于 2021 年 2 月获准立项建设，这是少数民族信息处理领域第一个获批的国家级重点实验室。实验室的研究领域为信息科学领域，实验室下设四个研究中心，主要围绕青藏高原地区科学技术的发展、多元文化融合及社会安全稳定发展等方面的需求，开展图数据与算法、藏语自然语言处理、藏语模式识别和民族文化智能处理等四个方向的研究工作。实验室根据需要设立首席科学家、学术带头人、特聘教授、客座教授、客座研究员等学术岗位，按照国内一流学科标准，制定学术岗位聘任条件和聘任流程。努力将国家重点实验室建设成为在西部地区有重要影响的科技创新基地，打造成为为藏语信息处理、人才培养、开放合作的重要平台和高地。

4.2.5. 获奖情况

2020 年青海师范大学 “藏汉(汉藏)机器翻译关机键技术及应用示范”获青海省科学技术奖一等奖(编号：2020-KJJB-1-5)；2017 年西藏大学“藏语自然语言处理关键技术研究和应用”获西藏自治区科学技术奖一等奖。2021 年青海师范大学“藏汉机器翻译关键技术及应用创新团队”获得了由青海省人民政府颁发的青海省科学成果奖创新驱动奖；西北民族大学国家通用语/藏语远程教育平台成果获中国中文信息学会钱伟长中文信息处理科学技术二等奖、甘肃省科技进步三等奖、甘肃省高校科技进步一等奖等。

4.3.维吾尔文语言文字信息处研究进展

新疆大学多语种信息技术重点实验室、新疆多语种信息技术研究中心、新疆民族语音语言信息处理实验室、新疆师范大学等研究单位，以习近平新时代中国特色社会主义思想为指引，深入贯彻落实党的十九大和十九届二中、三中、四中、五中、六中全会精神、第三次中央新疆工作座谈会精神、中央民族工作会议精神、习近平总书记关于教育的重要论述和对语言文字工作的重要指示批示，紧紧围绕“一带一路”新疆核心区建设和自治区“社会稳定长治久安”总目标，面向国家安全和新疆区域发展战略需求，立足新疆经济和社会发展，开展前瞻性、基础性、战略性、系统性研究。

4.3.1. 基础研究方面

面向机器翻译、信息抽取、文本分类、语义分析、语音识别、语音评测、国家通用语言文字推广等领域，在复杂形态语言的形态特征分析、语言模型与翻译模型建模、命名实体识别技术、实体关系抽取、情感分析、跨语言跨模态舆情监测研究等方面开展基础和应用研究。

主要针对多源、多通道、多形态、多格式、多错误（非标准）等复杂场景下海量文本分析处理开展研究，建立了资源稀缺语言高质量语料资源库和知识库，构建了机器翻译全流程应用平台。建立了大规模高质量的多语言语料知识库，研发了一系列文本智能处理工具库，为开展多语言自然语言处理奠定了坚实基础和技术支撑；构建基于复杂形态语言非标准文本特征分析的机器翻译模型并研发多语言机器翻译系统；创建了复杂场景下海量文本处理的工程技术体系。项目成果拓展了信息文化沟通交流渠道，提升了信息获取与掌控能力，显著推动了自然语言信息处理领域的技术进步。

承担了民族民间文化资源传承与开发利用技术集成与应用示范、大数据驱动的汉语与英语及中国少数民族语言之间的机器翻译等国家重点研发计划项目，维吾尔语汉语语音翻译系统关键技术研究、多少数民族语言连续语音识别方法及应用、基于深度语义的汉维机器翻译研究、基于无监督学习方法的口语理解与人机对话行为研究、维吾尔文网络社会集群行为感知的关键技术研究、中亚诸语言形态分析理论与方法研究等国家自然基金项目。

获中国科学院杰出科技成就奖 1 项、中国科学院科技促进发展奖二等奖 1 项、获得新疆维吾尔自治区科技进步一等奖 1 项、新疆维吾尔自治区科技进步二等奖 2 项，学科带头人获得新疆维吾尔自治区科技进步特等奖 2 项。

4.3.2. 应用研究方面

2017 年，麒麟软件/中标软件作为国产操作系统厂商参加了由新疆大学牵头的“国产多语种操作系统技术规范”项目研制工作。该项目研究支持多语种少数民族文字信息处理的通用接口规范和应用编程接口规范，形成《信息技术多语种桌面操作系统通用规范》，指导操作系统和应用程序开发。

2021 年 6 月由吾守尔院士、倪光南院士牵头在北京举办了“少数民族语言信息化推进战略研讨会”，国产操作系统联盟、新疆维吾尔自治区工信厅、麒麟软件有限公司等单位领导及少数民族语言信息处理相关专家学者参会，为少数民族语言信息化建设凝心聚力、定锚扬帆。

研发了维吾尔语语音识别、语音合成、机器翻译、舆情分析平台等系统，承担了《智能语音 AI 项目》、《英、汉、维多语言自然语言处理与机器翻译平台》、《面向公共安全的多语种网络舆情监测系统》、《文书图的自动生成分系统开发》、《国家通用语言文字模拟学习与评测网络服务开发》等横向项目。

4.3.3. 产业落地方面等

相关研究单位与鹏城实验室、中国科学院深圳先进技术研究院、中电科社会安全风险感知与防控大数据应用国家工程实验室、等研究机构及阿里巴巴、百度、腾讯、科大讯飞、中科软、麒麟软件有限公司、中电科新疆联海创智信息科技有限公司、中国电信新疆分公司等知名企业合作开展机器翻译、语音识别与翻译、语言分析等方面的合作开发，联合研制的汉-维互译双向语音机器翻译系统在自治区反恐维稳工作中发挥重大作用，在 24 万驻村干部承担的“访惠聚”和脱贫攻坚工作中应用，已推广应用 30 多万套。正在开发面向国家通用语言文字学习相关技术平台，努力为铸牢中华民族共同意识、社会稳定和长治久安提供技术支撑。

国家语言资源监测与研究少数民族语言中心已研发实现的应用系统包括：跨语言网络舆情动态监测与分析平台（国家自然科学基金重点项目）；“十四五”少数民族语言文字规范标准体系建设规划；中国少数民族语言使用国情地图；民汉机器翻译系统；民汉信息检索系统；跨语言领域知识图谱构建；民族语文政策知识库；蒙藏维常识性语义知识库；藏维文词法分析相关标准和规范；蒙、藏文编码转换系统；蒙、藏、维语料辅助标注工具等 20 余项。

4.4.壮语语言文字信息处理研究进展

4.4.1. 研究背景与意义

壮文是壮语的文字载体，目前存世的壮族文字有古壮文（亦称方块壮字、古壮字）和现代壮文。古壮文起源于唐代（一说起源于秦汉），大量使用于神话、故事、歌谣、剧本、对联、碑刻、家谱及契约等文本中，如今民间仍有不少壮族群众在使用。现代壮文是国家在 1950 年代为壮族人民创制并于 1980 年代修订的一种以拉丁字母为基础的文字，主要应用于重要文献/书籍翻译、法定单位名称牌匾/公章、部分主要公共场所/街道/地名的名称标识等。古壮文信息处理的研究目的是民族文化遗产的抢救与传承,而现代壮文信息处理的研究目的则在于现实应用,两者都具有重要的研究价值和现实意义。同时，壮文信息处理技术对开展东南亚语言信息处理研究具有积极的促进作用。

4.4.2. 发展现状与问题

受语言技术综合人才相对缺乏、政策关注度较低等因素影响，近年来，壮文信息处理进展相对较慢，但其延伸研究（东南亚语言信息处理研究）则得到了较大的发展。

（1）古壮文信息处理

将古壮字纳入国际标准是古壮文信息处理近几年的工作重点，在壮文信息技术国家标准工作组的努力下，已有 2200 多个古壮字被 ISO/IEC 10646 收录。近期从壮族古籍中统计出的古壮字已达 4 万多字，无论是国际标准的收录量，还是 2010 年研发的古壮文处理系统的容纳量，都远远无法满足当前壮族古籍整理的需求。为此，2021 年全国少数民族古籍整理研究室、中国社会科学院民族学与人类学研究所、广西少数民族古籍保护研究中心及南宁市平方软件新技术有限责任公司启动了少数民族古文字数字化处理系统古壮字子项目（项目期：2021~2023 年），项目系统旨在兼容国际标准，同时容纳所有整理出来的其它古壮字，并支持壮族古籍的录入、编辑、管理、检索和发布等。

（2）现代壮文信息处理

文本机器翻译和语音翻译在壮语区的应急语言服务、非日常基层工作以及语言学习等方面都有着现实的迫切需求。近年来，现代壮文信息处理研究主要集中在运用神经网络技术实现机器翻译和语音翻译上。2017 年南宁市平方软件新技术有限责任公司研发完成汉壮神经网络机器翻译系统；2018 年中国民族语文翻译局上线了壮汉机器翻译系统和语音翻译系统。由于存在以下几方面的问题，当前汉壮机器翻译和语音翻译效果不太理想：（1）文本方面，由于新壮文 1982 年才开始使用，语料主要来源为政府文件翻

译稿，数量少，且数据稀疏问题严重。（2）在语音方面，语音语料主要来源为播音文件，同样存在数据稀缺和稀疏问题；同时，壮语语音划分为 13 个土语区（标准壮语以武鸣区的壮语为标准音），不同地区语音差别大，如果仅实现标准壮语语音翻译，则无法切实现实需求问题。

（3）延伸研究

广西与东南亚国家陆海相连，壮族与不少周边国家民族语言相似、文化相通，壮文与东南亚语言的信息处理技术在一定程度上可以互相迁移、互相促进。为服务“一带一路”建设，南宁市平方软件新技术有限责任公司及广西达译科技有限公司在壮文信息处理研究基础上，开展东南亚语言信息处理的研发工作，完成了汉-越南语/泰语/印尼语/马来语/缅甸语/老挝语/柬埔寨语神经机器翻译系统、计算机辅助翻译系统、跨语言信息检索系统，以及东南亚互联网信息采集与挖掘系统，同时在越南语、马来语的语音识别与合成上也取得了突破。

4.4.3. 总结及展望

继续推进古壮字标准化工作；

针对壮族古籍的整理、分析及出版需要，研发含盖所有古壮字的新处理系统，以及古壮-现代壮、汉壮辅助翻译系统等古籍整理辅助工具；

进一步深入调研，着力解决汉壮机器翻译和语音翻译方面的语料问题；

进一步推进东南亚语言信息处理研究，使壮文信息处理和东南亚语言信息处理互相促进，服务“一带一路”建设。

4.5.朝鲜语语言文字信息处理研究进展

朝鲜文信息技术工作组成立于 2013 年，成立后在吉林省民委、延边州人民政府的支持和帮助下，在朝鲜文信息技术标准化建设和朝鲜文信息处理应用技术方面取得了较好的成绩，对中文信息处理领域做出了应有的贡献。

完成了标准化项目研发。共制定了 2 项国家标准，GB/T 34957-2017《信息技术 基于数字键盘的朝鲜文字母布局》、GB/T 34958-2017《信息技术 朝鲜文通用键盘字母数字区的布局》；完成了 3 项吉林省地方标准项目，DB22-T 2798-2017《朝鲜文信息技术术语和定义》、DB22-T 2799.1-2017《信息技术 朝鲜文通用字符 24 点阵字型 第 1 部分：白体》、DB22-T2798.2-2021《信息技术 朝鲜文术语和定义 第 2 部分：算术与逻辑运算》，改变了自 1989 年后 20 多年来朝鲜文信息技术标准化建设停滞不前的现状，其中《信息

技术 朝鲜文术语和定义 第 2 部分：算术与逻辑运算》标准化项目被吉林省科技厅认定为科技成果（证书编号 2021452）。

开发了中国朝鲜文输入法。集中技术力量研发了基于视窗系统、安桌系统、苹果系统和 LINUX 系统的 4 种朝鲜文输入法，填补了空白，从理论上改变了国内朝鲜文输入法只能依靠外来技术的尴尬局面。

开发了朝鲜文字型。历时三年时间开发了 15 种朝鲜文字型，（清明体、黎明体、书写体、白鹤体（平）、刚正体、光明体（长）、行书体、未来体、弓书体、昭南体、青峰体、岩石体、龙云体）实现了“零”的突破，打破了长期以来只能依靠外来字型的局面。

2021 年 11 月为止完成了朝鲜文编码字符集标准映射表。完成了 4300 个朝鲜文字符的国际标准、中、朝、韩三国字符集标准映射表，详细分析了各个标准之间存在的差异化问题，提出了修订和完善国际标准的相关方案，并完成了国家有关部门资助的《古朝鲜文编码字符集研究报告》，《国际标准 ISO/IEC 10646 与国家标准 GB/T 12052-1989 朝鲜文字符差异分析报告》，《现代朝鲜文编码字符集研究报告》，《关于在 ISO/IEC 10646 中增添 10 个古朝鲜文音节字与 1 个古朝鲜文字母的提案》等重要研究论文。

成功举办了两届国际学术研讨会，促进了国际交流。2018 年 8 月在延吉举办了中朝韩国际学术研讨会，通过研讨会重启了朝、韩之间关闭 11 年之久的学术合作交流之门。2019 年 7 月会同中国知网在长春举办了东北亚科学技术研讨会，再一次促进了中朝韩在科学技术方面的交流与合作，同时使中国朝鲜文杂志成功登陆了中国知网学刊大数据平台，进一步拓展了中国朝鲜文的影响力。

中国朝鲜文信息技术标准化建设在多方支持和工作组的努力下实现了零的突破、填补了空白、开创了局面，在为国家争取话语权主导权方面做出了应有的贡献，可以说取得了骄人的成绩。

第五章　机器翻译研究进展、现状及趋势

5.1.研究背景与意义

机器翻译（machine translation, MT）是指利用计算机实现从一种自然语言到另外一种自然语言的自动翻译。被翻译的语言称为源语言（source language），翻译到的语言称作目标语言（target language）。

简单地讲，机器翻译是打破语言壁垒，实现无障碍自由交流的关键技术，是自然语言处理领域的核心研究方向，它几乎涉及自然语言处理中的所有问题，被认为是自然语言处理乃至人工智能领域最具挑战的技术。

由于人们通常习惯于感知（听、看和读）自己母语的声音和文字，甚至很多人只能感知自己的母语，因此，机器翻译在现实生活和工作中具有重要的社会需求。从理论上讲，机器翻译涉及语言学、计算语言学、人工智能、机器学习，甚至认知语言学等多个学科，是一个典型的多学科交叉研究课题，因此开展这项研究具有非常重要的理论意义，既有利于推动相关学科的发展，揭示人脑实现跨语言理解的奥秘，又有助于促进其他自然语言处理领域、甚至语音技术和图像视觉处理等领域的快速发展。从应用上讲，无论是社会大众、政府企业还是国家机构，都迫切需要高质高效的机器翻译技术。特别是在“互联网+”时代，以多语言多领域多模态呈现的大数据已成为我们面临的常态问题，机器翻译成为众多应用领域革新的关键技术之一。例如，在商贸、体育、文化、旅游和教育等各个领域，人们接触到越来越多的外文资料，越来越频繁地与持各种语言的人通信和交流，越来越多的商品实现全球购全球卖的目标，文学和影像作品也越来越频繁地在各个国家传播，从而对机器翻译的需求越来越强烈；在国家信息安全和军事情报领域，机器翻译技术也扮演着非常重要的角色。可以说离开机器翻译，基于大数据的多语言信息获取、挖掘、分析和决策等其他应用都将成为空中楼阁。

尤其值得提出的是，在现在和未来很长一段时间里，建立于丝绸之路这一历史资源之上的“一带一路”将是我国与周边国家发展政治经济，进行文化交流的主要战略。据统计，“一带一路”涉及 60 多个国家、44 亿人口、110 余种语言，可见，机器翻译是“一带一路”战略实施中不可或缺的重要赋能技术。

机器翻译概念自 1949 年被正式提出以来，其技术经历了规则方法、统计方法和深度学习方法，特别是自 2017 年基于自注意力机制的 Transformer 模型提出以来，机器翻译技术在大规模的平行语料、高效的端到端大模型和充足的计算资源的共同推动下取得

了突破性进展。然而，正如上述所说，机器翻译是一个非常复杂的技术，一方面，该技术涉及到自然语言理解、语义转换和自然语言生成，每个模块都面临诸多难题；另一方面，机器翻译面临语音翻译、图像翻译、视频翻译、多语言翻译和低资源小语种翻译等更多更复杂的应用场景，每个场景面临数据匮乏和语义鸿沟等难题，因此，机器翻译技术还不完美，仍需要研究人员不断努力，真正突破语言壁垒、实现任意时间、任意地点和任意语言的自动翻译，完成人们无障碍自由交流的梦想。

5.2.领域发展现状与关键科学问题

随着深度学习的兴起，研究人员开始使用端到端的深度学习模型进行机器翻译的建模 (Cho et al., 2014; Ilya et al., 2014; Bahdanau et al., 2015)。通过深度神经网络，神经机器翻译系统可以直接对翻译过程进行序列到序列的建模，并通过反向传播算法从大量平行语料中直接对神经网络的参数进行学习。这样的翻译系统不再依赖从数据中挖掘的带有噪音的翻译对应关系，也不再进行基本单元的组合和评分，而是通过对一系列向量表示的数值运算完成整个翻译过程。上述简单的建模方式充分发挥了 GPU 等设备带来的计算能力上的飞跃，能够有效发掘数据中隐含的翻译相关的规律，从而获得了出色的翻译效果。

2017 年以来，Google 的 Vaswani 等人提出的 Transformer 模型(Vaswani et al., 2017)综合利用注意力、自注意力机制和层叠网络更加有效地进行了文本的建模，成为当前机器翻译系统研制的主流范式，其结构对其他自然语言处理任务也带来了深远的影响。目前，在大规模数据和 GPU 算力的支持下，机器翻译系统的性能稳步提升，机器翻译技术在行业中得到更加广泛的应用。同时，得益于 TensorFlow、PyTorch 等深度学习框架的兴起和流行，以及 dl4mt、OpenNMT、Tensor2Tensor、Hugging Face、Fairseq 等开源工具的贡献，机器翻译的基础系统的研发可以快速进行，这使得越来越多的研究者更方便地进入了机器翻译领域，推动了机器翻译研究的新高潮。

机器翻译的核心挑战包括两个部分，一是学习源语言和目标语言中语义相同的表达方式，二是生成一个完整流畅的目标语言句子。在当前研究中上述两个挑战仍然是机器翻译研究的关键科学问题。

5.2.1. 双语语义等价关系学习

近年来的机器翻译实践表明，在拥有大规模双语数据的条件下，基于深度学习的翻译系统往往能够达到比较好的翻译水平，在某些数据集上甚至可以达到与人工译员相当的翻译质量。然而，大规模双语数据的条件并不总能很好的满足。这主要体现在领域和

语言对两个角度。

从领域的角度来说，虽然中英、英德、英法等语言具有大规模的双语数据，但这些数据往往来自于一个或多个特定的来源，如新闻、专利、学术文献等，一般统称为领域。不同领域的双语数据，其语言、语法特点往往会呈现出某些特定规律，这使得从不同领域的数据中学习的翻译知识也存在一定的差别，给多领域混合学习造成困难。更重要的是，某个特定领域的双语数据规模往往要比该语言对整体双语数据规模小得多。这使得在特定领域获取高质量的翻译系统面临重要挑战。由此引发了机器翻译领域自适应研究，主要关注在特定领域提升机器翻译效果，着重于发掘领域之间的知识迁移、特定领域的词典等翻译资源利用等；多领域机器翻译研究，主要关注同时提升翻译模型在多个不同领域的翻译效果，着重于建模领域之间的共性知识和独有知识等。

从语言对的角度来说，仍然存在许多平行数据缺乏的语言对（称为资源稀缺或低资源场景），为机器翻译系统的学习带来巨大困难。在这些资源稀缺的场景中是否能够学习双语语义等价关系、如何学习双语等价关系，成为机器翻译研究的重要挑战。由此引发了无监督机器翻译研究，主要关注在没有平行语料的情况下进行翻译学习的方法和技术，着重于利用词典、单语数据等知识来源；多语言机器翻译研究，主要关注在多个不同语言对之间共享的机器翻译模型，着重于发掘在不同语言对之间的翻译知识共享；多模态机器翻译研究，主要关注融合图像、语音等不同模态数据进行机器翻译，着重于获得在不同模态的数据之间可以共享的数据表示，建立在不同模态数据之间的关联等。

5.2.2. 目标语言生成

生成流畅通顺的自然语言句子一直是自然语言处理中一个重要研究问题。在机器翻译研究中，生成的同时还需要保证句子的连续性，这实际上要求生成过程能够考虑到源端和目标端的上下文信息，给生成过程带来了更大挑战。在存在大规模平行数据的条件下，利用深层神经网络进行学习已经可以达到一定的生成质量，但是在实用场景中可能面临更复杂的上下文环境，比如在篇章翻译、同传翻译、并行翻译等场景下，存在许多新的研究挑战。

从篇章翻译的角度来说，篇章中的不同句子之间往往存在一定的逻辑关联，相邻句子中的单词也往往具有指代、单复数、一致性等方面的关联性。篇章机器翻译研究的目标即是保持整个篇章翻译的一致性和流畅性，主要关注在翻译过程中识别和利用篇章级的上下文的方法。

从同传翻译的角度来说，需要实时将语音信号转换为对应的文本信号，不仅面临着语音到文本转换过程中语音识别相关的问题，同时还面临着翻译过程中可能仅能看到部

分源语言上下文，源语言信息不完全等问题。同传翻译的研究一方面关注不同模态之间关联信息的利用，一方面还关注部分源语言上下文带来的歧义问题。

从翻译并行性角度来说，传统翻译按照语言的顺序（自左向右）地生成目标语言，每个符号生成时已经完成了其前所有符号的生成，这样的生成方式效率相对较低。如果能在翻译过程中能够有效利用周围目标端上下文信息，将有可能并行生成多个片段或者符号，极大提升翻译效率。并行翻译的研究主要关注在并行生成之前对前后目标端上下文的关联建模，或者在多次编辑修改过程中利用上下文提升翻译准确性，在一定程度上达到并行生成的目的。

5.3.领域关键技术进展及趋势

5.3.1. 机器翻译模型

5.3.1.1. 任务定义和目标

机器翻译模型是一种数学模型，该模型能够利用已有的平行句对数据，并从这些数据中学习到某些规律，从而实现从源语言句子到目标语言句子的转换。机器翻译模型经历了基于实例的模型、基于统计的模型和最近的基于神经网络的模型。特别地，近年来端到端的神经机器翻译获得了迅速发展，其翻译质量较统计机器翻译取得了显著提升，已成为当前机器翻译领域的研究热点和主流 (Bahdanau et al., 2015; Bapna et al., 2018; Chen et al., 2018; Dai et al., 2019; Dehghani et al., 2019; Gehring et al., 2017ab)。端到端的神经机器翻译模型采用编码器(encoder)-解码器(decoder)框架实现序列到序列的转换。在给定包含N句对的源语言-目标语言平行语料$D = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$的情况下，机器翻译模型的优化目标是使得目标函数$L(\theta)$在数据集$D$的翻译概率最大化，即：

$$L(\theta) = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)};\theta),$$

其中$P(y|x;\theta)$为源语言句子x到目标语言句子y的翻译概率，θ为模型参数。

5.3.1.2. 研究进展与影响

根据编码端和解码端网络结构的不同，端到端的神经机器翻译模型可以大体分为基于循环神经网络的模型、基于卷积神经网络的模型和基于自注意力的模型。下面，本文

首先分别讨论以上三种类型翻译模型的研究进展；然后介绍最近越来越受关注的非自回归翻译模型的研究进展。

基于循环神经网络的神经机器翻译模型：基于循环神经网络的神经机器翻译模型（Sutskever et al., 2014; Bahdanau et al., 2015）一经提出，便成为当时机器翻译的主流模型。随后，特别是 2015~2018 年期间，出现了大量的相关工作用于缓解该翻译模型的潜在问题。例如，针对模型的注意力机制忽略了历史的对齐信息从而造成过译和漏译现象，Tu et al. (2016)和 Li et al. (2018) 提出了在模型中融合源端单词的覆盖度的不同方法。研究者探讨了在模型中引入各种先验知识，包括句法和语义等语言学知识（Li et al., 2017; Wu et al., 2017; Wang et al., 2018; Song et al., 2019）、词对齐信息（Liu et al., 2016; Wang et al., 2017）、词典（Zhang et al., 2017）等。Zhang et al. (2018; 2019) 提出了改进的循环神经网络。Wu et al. (2016), Zhou et al. (2016), Wang et al. (2017), Chen et al. (2018)构建了深层循环神经网络模型。

基于卷积神经网络的神经机器翻译模型：类似于递归神经网络，卷积神经网络也可以对序列进行建模，并较前者具有并行性高的特点。Gehring et al. (2017a)提出了基于卷积神经网络的编码器和基于循环神经网络的解码器的翻译模型。随后，Gehring et al. (2017b)提出了一种全新的基于卷积神经网络的翻译模型 ConvS2S。有别于循环神经模型需要对句子从左至右进行编码，ConvS2S 使用多层卷积网络对源端和目标端序列进行编码和解码，底层的卷积网络用于捕捉相距较近的词之间的依赖关系，而高层卷积网络用于捕捉较远词之间的依赖关系。为了减少模型参数规模，Kaiser et al. (2018) 将深度可分离卷积网络应用到机器翻译中。

基于自注意力机制的 Transformer 模型：基于编码器-解码器框架，Vaswani et al. (2017)提出了一种基于自注意力的新型翻译模型 Transformer。该模型并不使用任何循环或卷积神经网络，而是使用一种自注意力网络对序列进行建模。自注意力机制通过直接对单词之间进行建模，可以非常高效地捕获序列内任意两个词之间的依赖关系。同样，Transformer 模型具有并行性高和训练速度快等特点。Transformer 模型一经提出就受到了广泛的关注，代表了目前最优的机器翻译模型之一。基于标准 Transformer，近年来研究者们提出了许多新的改进模型，在实验中这些新的模型往往取得了较标准 Transformer 更佳的翻译性能或者更快的训练/推理速度。Transformer 模型上的改进大体可以分为两类：1) 针对 Transformer 模型内各模块，特别是多头注意力网络(Guo et al., 2019; Wu et al., 2019; Wang et al., 2020; Tay et al., 2021)和位置编码(Shaw et al., 2018)的改进等；2) 针对 Transformer 系统结构的改进(Bapna et al., 2018; Hassan et al., 2018; Dai et al., 2019; Dehghani et al., 2019; Wang et

al., 2019)。更详细的关于 Transformer 模型的改进和变体可参见 Lin et al. (2021)。

非自回归神经机器翻译模型：以上翻译模型均为自回归神经机器翻译模型，即解码器在预测目标端的第t个单词时，依赖于前面位置的输出$y_{<t}$。因此，在翻译时只能从左至右逐词生成译文，翻译速度较慢。而非自回归神经机器翻译对目标语言的每个词独立地进行预测，不依赖于前面位置的输出，因此能一次性地预测出整句译文，显著提升了模型的翻译速度。Gu et al. (2018)首先提出了非自回归神经机器翻译模型，该模型沿用了 Transformer 的模型结构。与传统的自回归模型相比，存在着下面几方面不同：1) 编码端除了生成每个单词的向量之外，还为每个词预测其对应的目标端单词个数；2) 解码端的输入不再是目标端单词，而是源端单词的重新组合；3) 解码端各层中还包含了位置注意力网络。此外，考虑到源语言句子可以对应于目标端多种不同的译文，这会对非自回归模型的训练造成很大的干扰。为此，利用知识蒸馏技术，用自回归模型的输出译文替换平行语料中的目标端译文，可以有效地提升非自回归模型的翻译质量。由于目标端序列信息的缺失，非自回归模型的性能与自回归模型有较大的差距。后续的研究提出了多种提升非自回归模型翻译性能的方法，包括如何更有效地利用自回归模型（Wei et al., 2019; Zhou et al., 2020; Gu and Kong, 2021）、增强解码端的输入（Guo et al., 2019）和迭代式解码（Lee et al., 2018; Ghazviniejad et al., 2019; Gu et al., 2019）等。其中，通过在解码过程不断对译文进行修复，采用迭代式解码技术的非自回归模型不仅在性能上已接近自回归模型，同时在解码速度上仍大幅优于自回归模型。

5.3.1.3. 技术进展和发展趋势

近些年来，神经机器翻译模型得到了迅速发展，以 Transformer 为基础框架的翻译模型得到了广泛的应用。根据翻译模型本身存在的问题和实际的应用需要，本文认为机器翻译模型在未来几年仍需关注如下几个问题。

多头自注意力模型作为 Transformer 模型的重要组成部分，虽然能够有效的捕获不同层次的语言学信息，但其结构复杂，特别是当网络层数增加时，存在着大量的计算冗余，因此轻量级注意力模型的研究具有重要的实践意义。

目前的研究成果表明，加大模型的维度和网络深度能够提高翻译的性能。但大模型的训练和推理不可避免地会增加时间和空间上的开销。如何对模型进行压缩，利用预训练的大模型训练小模型，也是未来的技术发展趋势之一。

人类对语言的理解具有很强的鲁棒性，即使输入的句子存在拼写错误，很多时候并不会影响对其含义的理解。但对翻译模型来讲，其译文可能会很糟糕。甚至只需对输入仅做微小的改变，模型就会产生完全不同的输出。因此，如何增强翻译模型的鲁棒性对

于提升其实用性具有重要的意义。

5.3.2. 低资源机器翻译

5.3.2.1. 任务定义和目标

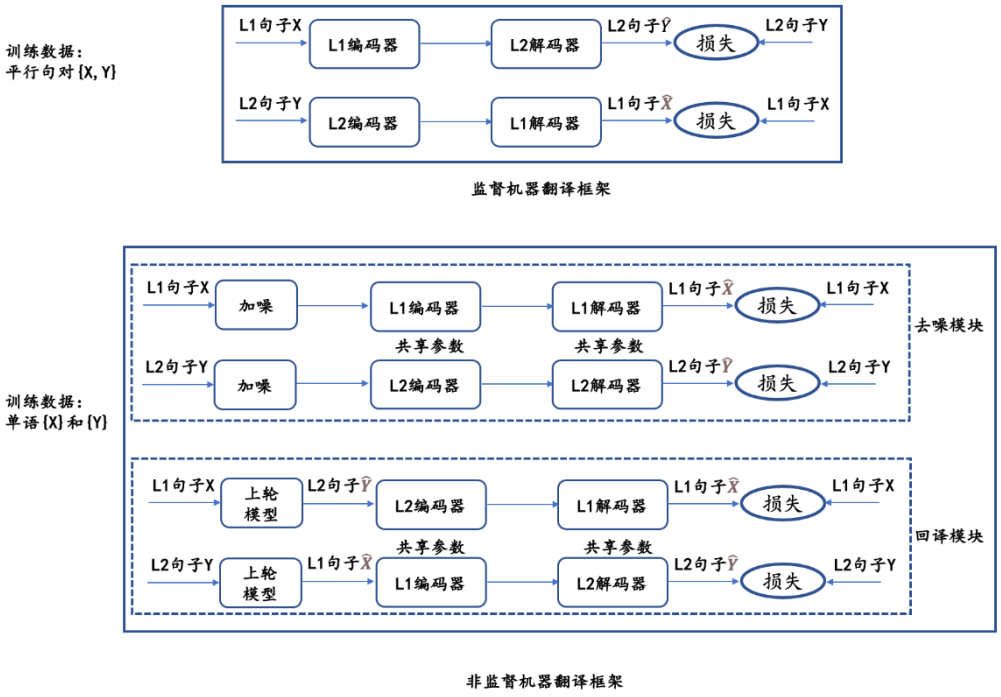
随着深度学习技术的发展，基于神经网络的机器翻译系统在性能上得到了大幅的提升 (Bahdanau et al. 2014, Vaswani et al. 2017)。这种全新的翻译框架使用基于编码器-解码器的神经网络代替传统的统计方法从大规模人工标注的双语平行语料中学习翻译知识模拟自动化翻译，在一些具有丰富语料资源的语言对上(比如中英，英法等) 所生成的译文已经达到了人类可以接受的程度。大规模双语平行语料则成为构建一个好的机器翻译系统的必要条件。事实上，由于国家和文化之间的交流不平衡，世界上绝大数语言之间缺乏平行语料。同时，构建大规模的平行语料库需要耗费的人力、物力以及财力也是非常昂贵的。所以目前的机器翻译主要应用在欧洲语言之间，以及世界上政治、经济贸易和文化较为发达的国家的语言之间，比如中英，英俄，中日，英日等语言对。相反地，东南亚语言、非洲语言和国内的少数民族语言等由于缺乏充足的平行语料，通常被称为稀缺语言对或低资源语言对。在缺乏充足平行语料的条件下构建机器翻译系统，则称为低资源机器翻译。其目标是降低机器翻译对平行语料的依赖，利用尽可能少的平行语料构建性能更好的机器翻译系统。特别地，在没有任何平行语料的条件下构建的机器翻译，则称为非监督机器翻译。

5.3.2.2. 研究进展与影响

中间语言和数据增强：目前比较普遍的低资源翻译是基于中间语言（pivot）的间接翻译。例如，中文-罗马尼亚语的平行语料较少，而中文-英语以及英语-罗马尼亚语的平行语料较多，因此可以将英语作为中间语言连接“中-英”和“英-罗”两个翻译系统，从而完成“中-罗”的机器翻译。这种方法的优势是简单且有效，但本质上还是依赖了中间语言与源语言和目标语言的平行语料，并没有很好的降低机器翻译对平行语料的依赖。与昂贵的人工标注平行语料相比，各种语言的单语语料的获取则显得非常容易(比如各个国家的新闻、书籍、网页等媒介中的语料)。因此，研究人员提出从单语语料产生伪平行语料的数据增强方法，主要有回译（back-translation）(Sennrich et al. 2016)和自训练（self-training）(Zhang and Zong, 2016; Sun et al., 2021)两种方法，以提升低资源机器翻译系统的性能。回译方法利用反向翻译模型，将目标语言的大量单语语料翻译成源语言，以构建{伪源语言，目标语言}的伪平行语料。而自训练方法则利用前向翻译模型，将源语言的大量单语语料翻译成目标语言，构建{源语言，伪目标语言}的伪平行语料。这两种方法都能够从单语语料出发，构建大量的伪平行语料，从而弥

补真实平行语料的不足。然而，由于机器翻译本身的缺陷，产生的伪平行语料的质量与人工标注还有很大差距，极大的限制了低资源的翻译系统的性能。

非监督机器翻译：此外，基于回译技术,研究人员还提出了仅需要单语语料的非监督机器翻译方法(Artetxe et al. 2019; Artetxe et al. 2018; Lample et al., 2018a; Lample et al., 2018b)。非监督机器翻译期初是从相似语言对之间开始，比如英语-法语和英语-德语，这类语言对具有类似的词汇和词表，在语言学上有天然的映射关系。在2018 年，研究者利用预训练双语词表示（bilingual word embedding），去噪自动编码器(denoising auto-encoder),回译和共享语言表示(sharing latent representation)实现了基本的非监督机器翻译框架，其基本的框架下图所示。



图：监督的机器翻译（上图）与非监督的机器翻译（下图）框架

多语言语言模型：随着预训练语言模型的蓬勃发展，多语言的预训练模型（multilingual pre-trained model）也在低资源机器翻译得到了广泛应用，比如 XLM，MASS，mBART,mRASP,mRASP2 等（Lample and Conneau 2019; Lin et al. 2020; Liu et al. 2020; Pan et al. 2021; Song et al. 2019)。这类模型通常利用多种语言的单语或者平行语料，通过设计一定的预训练任务对模型进行训练。预训练后的模型可作为机器翻译模型的初始化参数，然后通过机器翻译的微调得到最终的翻译模型。多语言预训练可以帮助模型形成语言无关的（language-agnostic）隐层表示，并提升模型在多种语言之间的对齐和映射能力。因此，对于低资源翻译任务来说，多语言预训练可以利用高资源语言的语料，帮助模型提升在低资源语言上的性能。除了作为模型的初始化参数以

外，预训练模型还可应用于平行语料的挖掘和生成，进一步减少了人工标注平行语料的需求。其中，平行语料挖掘的代表性工作是 CRISS (Tran et al., 2020)，它利用多语言预训练模型语言无关的隐层表示，从两种语言的单语语料中挖掘平行语料。而 Han et al. (2021)则提出了利用 GPT-3 模型生成平行语料的方法。

5.3.2.3. 技术进展和发展趋势

语言是人类智能的最高形式，机器翻译是自然语言处理和人工智能领域的核心问题和经典任务之一。在国家的经济和文化发展上，机器翻译也是关乎国计民生的技术和产业。目前我国经济和文化加速融入世界经济，与全世界的交流更加广泛。与此同时，我国也在开始引领区域和国际的发展，创立并领导国际性合作组织，比如“一带一路”、区域全面经济伙伴关系协定（RCEP）等。在这个过程中，机器翻译可以极大地缩小沟通的障碍，发挥着更加重要的作用。另一个方面，机器翻译的科学研究与应用几乎没有鸿沟，最新的科研技术可以直接快速地应用在相关领域，进而加快我国经济与世界经济接轨的速度。目前的低资源机器翻译主要还处于算法初步设计阶段，这项技术有着广泛的前景，如果能在通用的场景下进行低资源的机器翻译，不光在机器翻译和自然语言处理的学术界推动学科的发展，还对国计民生产生重大的积极影响。

5.3.3. 多语言机器翻译

5.3.3.1. 任务定义和目标

多语言机器翻译（Multilingual Machine Translation）旨在设计一个能够处理多个语言对的翻译模型和系统，例如一个多语言翻译系统不仅可以完成汉语到英语的翻译，还可以实现英语到汉语、汉语到法语和西班牙语到汉语等语言对的翻译。机器翻译技术发展至今，通常需要针对每个语言对（例如汉语到英语）构建一个机器翻译模型，从而若想实现n个语言之间的互译则需要构建$n(n-1)$个翻译系统。由于模型参数规模非常庞大，这种标准的机器翻译模型设计方式不仅将导致训练和部署将耗费巨大的存储和计算资源，而且也无法共享和利用相似语言之间的翻译知识。基于编码器和解码器框架的神经机器翻译使得多个语言共享编码器或解码器成为可能，多语言机器翻译方法应运而生。

多语言机器翻译的终极目标是利用可获得的所有语言的平行语料，在单个翻译模型中实现尽可能多的语言之间的自动翻译。可见，多语言机器翻译极大降低了模型训练和部署的成本，也便利了不同语言之间翻译知识的互相迁移和利用，对推动机器翻译的规模化产业应用提供了理论和技术支撑，因此多语言机器翻译研究具有重要的学术意义和

应用价值。

5.3.3.2. 研究进展与影响

简单地讲，多语言机器翻译就是基于多个语言对的平行语料构建单一共享的翻译模型实现多个语言对的自动翻译。可以看到，模型参数共享机制是多语言机器翻译研究的核心；此外，不同语言之间平行语料规模差异显著，如何平衡训练数据实现高效训练也是一个研究重点。下面，本文从上述两个方面分别介绍多语言机器翻译的研究进展与影响。

多语言机器翻译的参数共享：多语言机器翻译的关键在于如何设计不同翻译任务之间的参数共享程度，以同时建模语言之间的通用知识和每种语言的特有知识。在基于神经网络的多语言机器翻译方法的早期阶段，研究者通常使用语言独立的建模方法，为不同语言设计不同的编码器和解码器，并通过共享的注意力模块为不同语言建立联系(Dong et al., 2015; Firat et al., 2016)。例如，在一种源语言到多种目标语言的翻译模型中，不同翻译任务共享源语言端的表示，而目标语言端则使用独立的解码器(Dong et al., 2015)。这种方法学习到同一种语言在不同翻译任务上的翻译知识，最大程度地保留了语言独有知识。随后，Ha et al. (2016)和 Johnson et al. (2016)提出了全参数共享的模型，所有源语言和所有目标语言共享翻译知识。在这种方法中，所有不同的源语言共享编码器，所有不同的目标语言共享解码器，使得模型的大小不再受到语言数目的约束，进而推动了大规模多语言机器翻译的发展。通过融入更多语言对的训练数据，大规模多语言机器翻译可以在同一个模型中同时处理上百个语言对的翻译。

随着语言种类和训练数据的增加，单个多语言翻译模型遇到了表示瓶颈(Bapna and Firat, 2019)，不同语言差异带来的影响超过了语言间知识迁移的影响，使得多语言翻译模型的翻译质量出现下降。为了在保留通用的翻译知识的同时解决语言之间的冲突，研究者在全参数共享模型的基础上提出了不同的考虑语言特性的建模方法，降低参数共享程度 (Blackwood et al., 2018; Sachan and Neubig, 2018; Lin et al., 2021; Wang et al., 2018, 2019; Xie et al., 2021; Zhang et al., 2020)。Blackwood et al. (2018)使用统一的编码器和解码器，通过语言相关的注意力机制将不同语言映射到不同的语义空间来进行特定语言建模。Sachan and Neubig (2018)细致地研究了在基于自注意力的全参数共享翻译模型中，将不同模型组件根据语言特性划分为语言共享和语言独享类型。近期，研究者提出动态参数共享的方法，由模型根据不同翻译任务的训练数据和任务特征进行语言共享参数和语言独享参数的自动分配(Zhang et al., 2020; Lin et al., 2021; Xie et al., 2021)。通过融合共享参数和语言独享参数，多语言

机器翻译的翻译质量得到了显著提升，在多数语言对上都达到了与双语翻译模型（每个语言对构建一个翻译模型）可比或者更优的水平。

多语言机器翻译的高效训练：多语言机器翻译的训练数据来源于不同语言规模差异巨大的平行语料，例如汉语和英语拥有数千万甚至数亿的平行句对，而汉语和波斯语可能仅有几十万或者几万的平行句对。通常，多语言机器翻译的训练目标是最小化训练数据上的平均损失，从而导致资源丰富语言（例如汉语和英语）训练充分，而比资源稀缺语言（例如汉语和波斯语）无法得到有效训练。因此，训练数据间的不平衡是训练多语言机器翻译面临的严重挑战。

解决上述不平衡问题的一个常用策略就是通过启发式采样方法保持不同语言训练数据的平衡。例如，Johnson et al. (2017)和 Arivazhagan et al. (2019)通过对资源稀缺语言的训练数据进行上采样，使其从规模上接近资源丰富语言的训练数据，从而有效缓解资源稀缺语言训练不充分的问题。这种启发式方法虽然能够提升资源稀缺语言的翻译质量，但是需要针对不同数据集设置不同的采样参数，而且忽略了不同语言的学习难度以及语言之间的相似性。于是，Wang et al. (2020)设计了一种数据打分器自动地学习如何赋予每个语言的训练数据不同的学习权重。具体地，他们使用一种基于梯度的元学习方法通过最大化多语言开发集上的梯度相似性来学习不同语言上的采样分布。更进一步，为了避免计算和存储额外的梯度信息，Zhou et al. (2021)从目标函数的角度提出一种新的基于分布式鲁棒优化的学习目标，该目标旨在最小化多语言训练数据上最坏情况下的期望损失，从而实现多语言采样分布的自动学习，显著提升了多语言机器翻译的性能。

5.3.3.3. 技术进展和发展趋势

近些年多语言机器翻译发展迅速，在实际场景中得到了广泛应用，本文认为多语言机器翻译的未来发展趋势有如下几点：

参数共享机制自动学习：参数共享机制是多语言机器翻译的核心，虽然近年来的研究工作提出了若干多语言机器翻译的参数共享机制，但是大部分工作仍然需要先验知识设计语言共享和语言独享策略。如何针对不同的语言集合、不同的数据规模和不同的语言特性高效自动地学习参数共享机制，是多语言机器翻译的发展趋势。

融合预训练语言模型：多语言预训练语言模型通过大规模单语数据，提升了多种语言的表示能力，在跨语言自然语言理解任务、低资源和无监督机器翻译上都取得了很好的性能。针对多语言翻译设计合理的预训练目标，在提升表示能力的同时建模不同语言之间的语义映射关系，是多语言机器翻译的发展趋势之一。

多语言机器翻译的连续学习：多语言机器翻译通常是在给定的多个语言集合上学习单一共享的翻译模型，往往忽略了数据和语言的动态增长的特性。例如，在汉语、英语、日语、法语和德语五种语言规模给定的训练数据上训练的多语言翻译模型将面临五种语言训练数据不断增多和语言种类不断增加的挑战。因此，如何设计多语言机器翻译模型，使其具有连续学习的能力也将是未来的一个发展趋势。

5.3.4. 语音翻译

5.3.4.1. 任务定义和目标

语音翻译是指利用计算机将一种语言的语音信号（输入）翻译为另外一种语言的文字或者语音（输出），涉及语音处理（语音合成、语音识别）、机器翻译等人工智能技术。根据翻译方式的不同，语音翻译有交替翻译和同声传译两种形式。交替翻译指源语言句子语音信号全部输入完毕以后，再开始翻译。同声传译指在不打断讲话者的条件下，将讲话内容持续实时地翻译给听众，要求翻译系统既要保持较高的翻译质量，又要保证较低的时间延迟。

5.3.4.2. 研究进展与影响

与传统的文本翻译不同，语音翻译面临环境噪声、说话人口音、表达口语化等挑战。此外，机器同传还面临如何平衡翻译质量和时间延迟的关键难题。近年来，随着深度学习、语音处理、机器翻译等技术的迅速发展，语音翻译取得了长足进步，并且广泛应用于旅游、学习、会议等场景。

目前，语音翻译主要有两种模型。一种是级联模型（cascaded model），另外一种是端到端模型（end-to-end model）。级联模型将语音识别、机器翻译、语音合成三个模块顺序连接，语音识别模块将源语言语音信号转写为源语言文本，机器翻译模块将源语言文本翻译为目标语言文本，最后语音合成模块将目标语言文本合成为目标语言声音。级联模型将 3 个子模块松耦合的连接在一起，各个模块可以分别优化，是目前实际系统中使用的主流模型。对于机器同传，级联模型中需要解决翻译质量和时间延迟的平衡问题。通常用的策略有固定策略（fixed policy）和自适应策略（adaptive policy）。固定策略事先确定一个固定长度对源语言流文本进行切分，而不依赖于具体上下文，如设定固定切分窗口 (Sridhar et al., 2013)、等待词策略(Dalvi et al., 2018; Ma et al., 2019)等。自适应策略动态调整源语言的读入长度，如基于规则的策略 (Cho et al., 2016)、基于强化学习的策略 (Gu et al., 2017)、基于模仿学习的方法 (Zheng et al., 2019)、基于语义单元的策略 (Zhang et al., 2020)、基于单调无限回溯注意力机制的

策略 (Arivazhagan et al., 2019)等。

端到端模型直接从源语言语音信号到目标语言文本或者语音进行建模。相比于级联模型包含多个模块和处理步骤，端到端模型在结构上更简洁，近年来的研究也表现出该方法较大的潜力。如基于课程表学习的端到端模型 (Kano et al., 2017)、两阶段模型(Sperber et al., 2019)、基于多任务学习的模型 (Anastasopoulos et al., 2018)、基于预训练模型的方法 (Bansal et al., 2018)、基于知识蒸馏的方法 (Liu et al., 2019)、语音识别与翻译交互解码 (Liu et al., 2020)、融合语音和文本的多模态预训练模型 (Zheng et al., 2021)等。不过，受语音翻译数据规模的限制，端到端模型目前面临严重的数据稀缺问题，整体上还未取得对比级联模型的明显优势。

5.3.4.3. 技术进展和发展趋势

随着技术的进步以及翻译质量的持续提升，语音翻译日趋实用。从展现形式上，目前的机器同传产品可以分为“字幕式”和“语音式”。 所谓“字幕式”，就是将语音识别和翻译结果用字幕的方式投影到大屏幕上，现场观众通过看字幕了解演讲内容，形象地来说，就是“看同传”。 所谓“语音式”，是将同传的结果以语音的形式播放出来，也可以称为“听同传”。 从部署方式上来看，机器同传系统可以分为离线部署和云端部署。离线部署，将同传系统封装在一个离线机器里，其优点是不受网络影响，同时可以缩短系统时间延迟。缺点是如果多场会议并行，会增加机器的部署成本。云端部署，将同传系统部署在云端，优点是部署和维护都比较方便，可以迅速接入会议。缺点是受网络抖动、带宽等影响，可能会增加时间延迟甚至出现同传中断的情况。从硬件形式上来看，根据不同的场合和需求，产品也灵活多样。有面向会议场景的一体机，面向旅游场景的翻译机、翻译耳机等。

针对语音翻译的挑战，未来的研究方向包括语音识别纠错、鲁棒性翻译模型、数据集建设、面向场景的评价体系和评价指标、以及融合语音、语言、视觉信息的多模态翻译等。

5.3.5. 多模态机器翻译

5.3.5.1. 任务定义和目标

多模态机器翻译是神经机器翻译的重要研究方向之一。其任务定义是以源语言文本和图像、视频等多模态信息作为输入，自动产生目标语言译文。相比于传统的纯文本机器翻译模型，多模态机器翻译主要致力于有效利用视觉等多模态信息消除源语言文本的歧义，进而提升机器翻译模型的性能。2016 年，Elliott 等人第一次提出基于神经网络

的多模态机器翻译模型 (Elliott et al., 2016)，同年 Specia 等人组织了第一届多模态机器翻译评测 (Specia et al., 2016)，可以说，学术界和产业界同时注意到了该任务的研究意义和应用价值。

纵观多模态机器翻译近年来的研究进展，相关工作主要致力于融入图片信息的多模态神经机器翻译模型建模和分析研究，并在此基础上根据不同应用拓展出多个不同的多模态神经机器翻译模型。

5.3.5.2. 研究进展与影响

融入图片信息的多模态机器翻译：在这方面，相关工作主要聚焦于以下四点：图像特征表示、模型建模、训练和分析。

图像特征表示：由于多模态机器翻译源自于图像文本描述研究（Image Captioning），现有研究多借鉴计算机视觉研究所使用的图像特征，具体包括：1）**全局特征**：将图像表示为一个向量；2）**局部特征**：将图像等分为多个区域，每个区域表示为一个向量；3）**物体特征**：提取图像物体，每个物体表示为一个向量。现有研究中广泛应用了这三种特征，它们对多模态机器翻译系统性能提升均有帮助。

模型建模：模型建模是多模态机器翻译的研究重点，旨在研究如何将图像特征表示融入翻译模型中，相关方法主要包含： **1）双重注意力机制。**Calixto 等人分别对文本和图像进行注意力机制操作并融合两个上下文向量 (Calixto et al., 2017)。在此基础上，Libovicky 等人提出了扁平注意力机制和层次注意力机制 (Libovicky et al., 2017)。**2）图像信息作为文本信息补充。**该类方法使用图像特征初始化编码器，或使用图像特征拼接文本序列作为编码器输入。Elliott 等人的研究表明仅使用图像特征初始化编码器能取得更好的效果 (Elliott et al., 2016)。Delbrouck 等人提出使用文本隐状态关注图像信息以补充文本隐状态 (Delbrouck et al., 2017)。Yin 等人基于融合跨模态语义单元关系的多模态图来建模神经机器翻译模型 (Yin et al., 2020)。Yao 和 Wan 提出了多模态自注意机制，过滤无关图像信息并强化与图像相关的文本信息的抽取 (Yao and Wan, 2020)。3）**跨模态联合表示。**这类方法更为直接，主要关注如何学习跨模态联合表示来帮助后续翻译。在这方面，Calixto 等人引入变分神经网络来学习代表文本和图像语义的隐变量 (Calixto et al., 2019)。Lin 等人提出利用解码器隐状态来指导胶囊网络提取文本和图像的上下文联合表示 (Lin et al., 2020)。

模型训练：如何在模型训练中有效融入图片信息也是研究者关注的焦点之一。Zhou 等人提出联合翻译和图像-文本的语义匹配任务进行多任务训练 (Zhou et al., 2018)。Wang 等人则引入对比目标函数，强化模型捕捉与句子相关的图像物体信息 (Wang et

al., 2021)。Caglayan 等人则首次引入预训练来增强多模态机器翻译模型 (Caglayan et al., 2021)。Huang 等人提出利用图像描述模型与回翻产生伪平行数据 (Huang et al., 2020)。

模型分析：针对图像信息对于机器翻译的作用，不少研究者进行了深入分析。Elliott 等人根据模型使用相关/不相关图片的性能差异来量化图片信息作用 (Elliott, 2018)。Wu 等人使用门控机制将图片信息融入文本隐状态，根据门控权重来量化图片信息作用 (Wu et al., 2021)。Yang 等人以无性别差异的土耳其语作为中间语言，在回翻中加入图像信息以研究图片对译文选择的影响 (Yang et al., 2021)。上述工作得出了一致结论，即在文本信息有限时，图片信息能发挥较大的作用。

其它多模态机器翻译模型：除了上述模型，研究者们从应用角度出发，进一步提出了多个多模态机器翻译模型，主要包括：1）**电商产品多模态机器翻译。**与图片模态机器翻译不同，电商多模态翻译具有专有名词多，关系更复杂，数据稀疏的特点。在这方面，Song 等人提出了第一个大规模面向电商产品翻译的数据集，并设计了一个统一的跨模态、跨语言模型 (Song et al., 2021)。2）**拍照多模态机器翻译。**在这方面，传统方法主要采用先 OCR 识别再进行机器翻译的级联式方法，但是这样存在 OCR 识别错误传播的缺陷。对此，Mansimov 等人探索了端到端的拍照多模态机器翻译模型 (Mansimov et al., 2020)。Jain 等人提出了两阶段的拍照多模态机器翻译训练方法 (Jain et al., 2021)。3）**视频多模态机器翻译。**视频多模态翻译面临着视频信息冗余，训练数据稀缺的难题。针对数据稀缺的难题，Sanabria 等人和 Wang 等人分别提出了 How2 和 VATEX 的多模态数据集 (Sanabria et al., 2018; Wang et al., 2019)。Gu 等人提出了一种基于空间层次注意力机制的视频多模态机器翻译模型，同时解决了动词歧义和名词歧义的问题 (Gu et al., 2021)。**4）同传多模态机器翻译。**与上述任务不同，同传多模态翻译任务面临的挑战是，如何在保证时延的前提下利用图像信息。Caglayan 等人将双重注意力机制工作应用到该任务上并对其进行大量的分析 (Caglayan et al., 2020)，Ive 等人则利用强化学习的方法探索了不同类型的视觉信息和集成策略对同传翻译模型的质量和延迟的影响 (Ive et al., 2021)。

5.3.5.3. 技术进展和发展趋势

未来多模态机器翻译的发展趋势包含以下几方面。第一，现有多模态机器翻译数据集十分稀缺。相应地，有两方面研究值得关注：如何高效构建多模态平行数据集，如何充分利用模态缺失的数据集。第二，如何建模多语言多模态间的共性和特性，这对发挥翻译模型的潜力具有重要影响。第三，现在多模态翻译模型学到的知识还较为粗浅，融入外部知识，例如预训练模型知识、人类先验知识，将有助于提升模型性能。

5.3.6. 译文质量评价

5.3.6.1. 任务定义和目标

译文质量是整个翻译工作的核心，译文质量的高低决定了翻译任务的成败。人们评估机器翻译系统输出译文质量的过程被称为译文质量评估（Translation Quality Evaluation）。可将其看作是一个对译文进行打分或排序来反映翻译质量优劣的过程。

译文质量评估的核心问题是评估标准，也即什么样的译文是高质量译文。一般而言，高质量译文不仅需要准确无误反映原文内容、流畅通顺运用目标语言，有时还需要满足目标语言的文化习惯、能准确传达文字深层的信息，也即人工翻译中所提出的“信、达、雅”的标准。按这个思路，人们常用的评估标准有流畅度（Fluency）和忠诚度（Fidelity）两项 (Church and Hovy, 1993)。其中流畅度反映了译文在词义、结构、风格等方面表达目标语言的合理程度，越通顺的译文流畅度越高；忠诚度反映了译文能够准确传达源文意思的如实程度，越全面、准确的表达源文的意思，译文的忠诚度越高。

一方面，对于机器翻译研究和开发者而言，译文质量评估的主要目的，是通过给出量化结果来指导、支撑技术对比和优化。在机器翻译的发展进程中，译文质量评价有着非常重要的作用。正是由于 BLEU (Papineni et al., 2002) 等代表性自动评价方法的提出，机器翻译研究人员可以不需要人工干预并在短时间内得到译文质量的评价结果，从而加速了机器系统研发部署的进程。从某种意义上说，译文质量评价技术的发展及系列评测活动，直接引领和带动机器翻译的蓬勃、快速发展。

具体的，近年来在机器翻译研究中被广泛认可的自动评价指标包括 BLEU、NIST、METEOR 等。其中，BLEU 通过比较机器输出译文与参考译文之间的匹配 n 元语法单元来评价机器输出译文的质量。NIST 方法则在 BLEU 的基础上，依据匹配的 n 元语法单元的罕见程度当做权重，影响最终译文质量的评价。METEOR 的提出是为了补足 BLEU 方法的短板，同时它还具有词干提取、同义词匹配以及精准词匹配等特有功能。到 2021 年为止，译文质量评测任务在各类评测比赛中增添了丰富的评测类型与数据集来源，参赛队伍规模与水平远超以往。

另一方面，对机器翻译使用者而言，译文质量评估的主要目的，是帮助他们有效地在不用产品或不同引擎间做出选择。对于更为激进、更大程度上依赖智能技术的用户来说，译文质量评估技术还可用于承担译审的角色，即对于人工或机器输出的译文、自动评估译后编辑工作量，帮助筛选出需要人工干预、进行译后编辑的译文，甚至进一步帮助译员快速定位译文中的错误位置、内容及类型。在这个意义上，译文质量评估技术、自动译后编辑技术（Automatic Post-editing）和人机交互翻译技术具有密不可分的关

系 (Zhang et al., 2020)，他们的配合使用可以实现智能化、低成本的译文错误定位和修正。

具体的，国际影响力较大的 WMT 系列评测与国内知名的 CWMT/CCMT 系列评测分别于 2012 年、2018 年引入了语句级别的自动译文质量评价任务。该任务目标是准确的预测译文的 HTER 分数 (Snover et al., 2006)，代表着译文需要人工译后编辑的程度。

5.3.6.2. 研究进展与影响

根据人机协作方式和使用场景的不同，人们常将译文质量评估技术分为人工评估方法、自动评估方法两大类。而后者又可细分为有参考译文的自动评估方法、无参考译文的自动评估方法。

1）人工评估。是指评估者根据翻译结果好坏对译文进行评估。2013 年，全球语言和翻译行业的资源中心，翻译自动化用户协会 TAUS（Translation Automaton User Society）正式发布实施了动态翻译质量框架，其中囊括了丰富的译文质量评估知识库和多重标准的质量评估工具。我国对译文质量也形成了国家标准。2003 年以来，由中国翻译工作者协会等专业翻译机构起草，中国国家标准化管理委员会陆续发布了针对翻译服务的标准《翻译服务译文质量要求（GB /T19682—2005），规定了翻译服务译文质量的基本要求、特殊要求、其他要求、译文质量评定和检测方法等。

在人工评估时，一般由多个评估者匿名对译文打分，之后综合所有评估者的评估结果给出最终的得分。人工评估可以准确反映句子的翻译质量，是最权威、可信度最高的评估方法，但是其缺点也十分明显，不仅需要耗费人力物力，而且评估的周期长，不能及时得到有效的反馈，难以支撑快速迭代发展的机器翻译研发。

在面向机器翻译的人工评估方法中，直接评估（Direct Assessment）法被业界广泛认可使用 (White et al., 1994)，这种评估方法需要评估者审查翻译内容是否正确传达源文句意，使用词语和表达是否符合目标受众，以及是否考虑地域和文化因素来对机器译文的绝对评分，最终得到的百分制的分数用来表征机器译文的质量。另一种经典的人工评估方法是相对排序（Relative Ranking）(Callison-Burch et al., 2007)。这种方法通过对不同机器翻译的译文质量进行相对排序得到最终的评估结果。

2）自动评估。狭义的自动评估（Automatic Evaluation）是指在人类专家翻译的结果作为参考答案的基础上，将译文与答案进行自动比对，用其近似程度作为评估结果。即译文与答案越接近，评估结果越好。自动评估具有高效、稳定的优点。

在具有代表性的自动评估方法中，基于词串比对的方法聚焦于译文词语及n-gram的

翻译准确性 (Snover et al., 2006)。其思想是将译文看作符号序列，通过计算参考译文和机器译文间的序列相似性来评价机器翻译的译文质量。另一种基于词对齐的方法在机器译文和参考译文的单词之间建立一对一的对应关系 (Banerjee and Lavie, 2005)，这种评价方法在引入准确率的同时还能显性引入召回率作为评价所考虑的因素，从而反映机器翻译的忠实度。此外，基于检测点的方法和多策略融合的评价方法分别从具体问题的语言学检测点和多角度对译文进行综合评估。

3）质量估计。无参考译文自动评价在机器翻译领域又被称作质量估计（Quality Estimation），与上段中的介绍的自动评价技术不同，质量估计旨在不参照标准译文的情况下，对机器翻译系统的输出进行评价。

质量估计的本质是通过训练教会机器在没有参考译文的情况下判断译文的总体质量好坏，因而实用价值很强。在人助机译的工作模式下，可用于间接反映译后编辑的工作量。根据对机器翻译系统的不同输出层次进行划分，质量估计可以具体划分为四个级别：单词、短语、句子、文档。

常见的技术路线是将质量估计模型的基本框架设计为两部分：1）特征抽取模块：用于在数据中提取能够反映翻译结果质量的“黑盒”特征。 2）质量评估模块：利用前者从源语言和目标语言中抽取的特征表示，通过回归或分类算法去设计质量评估模型

传统机器学习方式中，句子都是由某些特征表示的。因此需要人工设计能够对译文质量评估有指导性作用的特征 (Esplà-Gomis et al., 2015)，常用的特征有：复杂度特征：衡量源语言端句子的复杂程度和翻译难度。一般认为源语言结构越复杂越难翻译，其对应的翻译结果质量可能也会比较差。 2）流利度特征：反应翻译结果的流利程度，包括目标语句中的字符数及目标语句的语言模型概率。3）充分度特征：表征翻译结果是否准确传递出源文想要表达的意义。通过提取源文与翻译结果相对应的特征，来判断两者的结构和含义是否一致。

神经网络技术为质量评估任务带来了能够自动学习评估译文质量特征的质量向量（Quality Vector）。因此，无论是从译文与答案的相似度匹配精度，还是对源文本身词义、结构、语义等抽象特征的深度掌握 (Chen et al., 2017)，基于神经网络的质量评估架构与传统架构相比都产生了质的飞跃。

近年来，随着各种预训练模型在自然语言处理的一系列任务中取得瞩目效果，在低资源语种和特定领域中借助预训练模型来提取句子代表性特征。这种方法大大减少了质量评估模型所需的训练数据规模和网络结构的复杂度，并取得了显著效果 (Kepler et al., 2019)。

5.3.6.3. 技术展望与发展趋势

译文质量评估仍面临诸多挑战。第一，译文多样性和评估标准多样性呼唤新的译文评估方法。由于自然语言高度的歧义性、灵活性和多样性，机器翻译的参考答案本身并不唯一。此外，对译文准确、全面的评价准则很难制定，自动评估与人工评估的结果之间也往往存在较大偏差。这些深层问题的解决，有赖于译文评估方法的创新与突破。第二，现有译文评价技术路线中，也有很多问题暴露出来，亟待分析解决。包括机器译文中的错误分析和错误分类问题，增强译文质量评估模型的健壮性问题，有效降低参考译文标注成本的问题，等等。

5.4.领域产业发展现状及趋势

5.4.1. 机器翻译产业发展现状

得益于机器翻译技术的不断突破，机器翻译产业也得到了蓬勃发展，在许多行业中有所应用。例如,《京津冀协同发展语言服务调查报告》调查数据显示，北京地区有 87.7%的需求方愿意在语言服务中使用机器翻译。随着机器翻译产品和应用大量涌现，创新创业日益活跃，许多知名企业及创业公司纷纷加入此赛道。据企查查数据显示，目前中国机器翻译服务的在营企业有 5680 家，各大互联网巨头也将机器翻译作为重要的赛道。

目前，涉及机器翻译技术的企业主要分为两类，除了有自身业务需求的高技术型企业外，还有许多新加入机器翻译赛道的创新型初创企业。这些企业既为 B、G 端市场（面向企业和政府）提供定制化翻译系统和云平台，同时也通过翻译耳机、翻译笔等智能硬件产品布局 C 端市场（面向个人用户），打造出服务亿万用户的翻译产品。

5.4.2. 机器翻译产业发展特征

机器翻译技术的变迁带来了产业应用的变革，因此其产业发展也具有鲜明的特色。

5.4.2.1. 从简单使用渗透到翻译生成环节

机器翻译技术尚未发展成熟之前，受翻译品质、数据规模、技术成熟度等因素限制，只能辅助一些人力所不能及的多语言阅读场景，比如，海量互联网文本翻译等。但是，随着神经机器翻译带来的翻译品质的提升，如今机器翻译已经从传统的辅助阅读等应用逐渐演化成可独立使用的翻译产品。例如，机器翻译已经可以作为独立的翻译软件，直

接服务于知识产权、医药等领域。从某种程度上来说，机器翻译也在逐渐改变传统翻译行业的现状，虽然机器翻译的品质和应用还没有达到完美，也无法完全替代高端人工翻译，但其对于海量翻译生产任务有着极大的促进作用。

5.4.2.2. 形态、场景、应用的多样化

除了传统的文字翻译，机器翻译的应用场景逐渐扩展到对文档、语音、图片、图像等多模态内容的翻译。例如，在一些体育赛事报道中，在线视频浏览的场景中，机器翻译已经成功与语音和图像处理技术相结合，进行多语言内容的传播方式。包括现在许多国际会议的同声传译也都由机器翻译自动完成。此外，图片翻译等功能也大量使用在手机应用中，为出国旅游、跨国交流提供便利。而且，使用机器翻译进行外语学习也受到广泛关注，包括词典笔等智能翻译小设备进一步丰富了机器翻译的应用形态。

5.4.2.3. 技术和产品的研发重心发生转变

长期以来，机器翻译的技术研发一直依托于高校和专门的科研机构。这种局面随着神经机器翻译对大规模算力和语料的要求不断上升，也发生了转变。最明显的特征是，机器翻译已经成为了企业技术研发的重要基础技术之一，大量有实力的企业都建立了机器翻译团队专门对机器翻译的技术和产品进行攻关。导致整个产业也在逐渐从简单的应用产品研发，转向对机器翻译基础方法和高技术产品的研发，这大大加速了机器翻译研究的进展。例如，这些年被广泛使用的 Transformer 模型就是来自于企业的研究团队，而这类模型也是最先被应用在机器翻译的在线服务产品中。从某种意义上说，现在机器翻译的技术研发已经逐渐演变成由高校、科研院所、技术型企业等多方联合主导，产业化对机器翻译技术的发展方向的影响更加明显。

5.4.3. 机器翻译产业发展面临的挑战

机器翻译的产业发展进入了快速上升期，但也面临着诸多挑战。主要有以下几个方面：

5.4.3.1. 挑战一：如何走好“最后一公里”

随着近年来机器翻译技术步入发展快车道，市场需求大量涌现，机器翻译的产学研融合迎来了新契机。但是，目前机器翻译技术尚未形成单独的产业链，也就是说，单纯的机器翻译技术还无法形成可直接交付给用户的产品或者服务。想要直接使用机器翻译，

仍然需要对其进行改造、定制，并进行产品化。而这部分内容的研发在传统机器翻译研究者的视角中并不太受关注。换言之，机器翻译如果不解决“最后一公里”的问题，其应用无法很好地进行推广。这也迫使机器翻译研究的方向选择进行调整。例如，机器翻译应用中更加关心如何让机器翻译系统通过少量的数据进行适应，对错误样例进行快速修正，也就是低成本的少样本学习。再比如，对于用户来说，如何进行机器翻译系统与用户的交互，如何使机器翻译在使用过程中能够接受用户的反馈。以上都是机器翻译应用中需要解决的问题。

5.4.3.2. 挑战二：更小更快的系统

神经机器翻译带来翻译品质巨大提升的同时，也对算力提出了更高的要求。比如，在实验室中使用机器翻译系统，动辄需要几个 GB 甚至十几个 GB 的显存要求，对浮点运算的能力也有极高的要求。这些，从一定程度上增加了机器翻译大规模应用的成本，特别是在一些运算资源受限的场景中，如何使用更小的存储，如何让机器翻译系统运行得更快也成为了机器翻译产业应用需要解决的问题。比如，在端侧部署机器翻译需要考虑模型压缩、低精度计算等手段，在不降低（或者尽可能少的降低）机器翻译品质的同时保证翻译的延时和存储空间在合理范围。再比如，在超大并发的情况下，如何合理地利用设备，如何有效的进行并行化。这些都是机器翻译产业化对技术提出的更高要求。

5.4.3.3. 挑战三：市场需求碎片化

虽然机器翻译的应用场景很多，市场需求却非常碎片化。简单来说，机器翻译看似“哪里都可以用”，实际上并不是“哪里都可以用”。造成这种现象的原因，一方面是由于机器翻译技术还没有成熟到完全达到人类翻译的水平，更重要的是由于机器翻译的应用需求非常琐碎，单一的一套产品很难解决所有问题，至今还没有出现可以对每一个行业、场景“通吃”的系统。甚至，由于机器翻译中仍然存在错误，应用时还需要大量的定制化开发和调优，来进行补救。这也在一定程度上增加了机器翻译应用的成本。当然，机器翻译需求端的变化也是机器翻译系统研发人员所需要面对并适应的。未来，随着技术的日益成熟和应用场景的不断挖掘，一定会出现标准化的、可规模化的机器翻译产品，这需要机器翻译研究者和使用者的共同努力。

5.5.总结及展望

本章首先介绍了机器翻译的定义、研究背景和意义，作为自然语言处理和人工智能领域的核心任务，机器翻译仍面临很多问题和挑战，需要更多努力和创新，真正打破语

言壁垒，实现各种语言各种场景的自动翻译。本章接着介绍了机器翻译领域的发展历史、现状与关键科学问题，其中双语语义等价关系学习和目标语言生成时两个核心科学问题。然后，重点从机器翻译模型、低资源机器翻译、多语言翻译、语音翻译、多模态翻译和译文评价等六方面介绍了自然语言处理近五年（2017-2021）的发展情况。最后对机器翻译产业发展现状及趋势进行了分析、总结和展望。

机器翻译是自然语言处理领域落地应用最多的技术的之一，在日常社会生活中已经无处不在，在汉语-英语、英语-法语等大语种以及口语、新闻和专利等资源丰富领域，译文质量已经接近或达到普通人类水平。未来，机器翻译技术的研究和应用前景仍然非常广阔，有望在以下几个方面实现突破：

机器翻译模型架构： Transformer 模型自 2017 年提出以来，几乎成为机器翻译唯一的模型框架，五年来虽然很多研究揭示出 Transformer 模型的问题，但是一直没有出现可以替代的建模架构。尽管核心模型的更新换代难以预测，但是本文认为更优的模型架构探索一定是未来机器翻译的一个研究趋势。

极低资源机器翻译方法：近五年来，回译等数据增强技术、大规模预训练技术、无监督机器翻译技术以及多语言翻译技术极大提升了低资源场景下的机器翻译质量，但是离实用水平仍有很大差距。世界上目前正在使用的语言有 4000 余种，而几乎 99%的语言都是低资源、甚至是极低资源语言。因此，如何结合现有技术或结合现有技术以及各种形态的数据资源，实现极低资源场景的高质量机器翻译将是未来的一个研究热点。

大范围场景感知的机器翻译方法：当前，机器翻译技术主要还是以文本句子为单位进行逐句翻译。一方面，句子包含的语义往往不完整，需要更大范围的上下文信息进行补充和消歧；另一方面，文本模态包含的语义也经常不够完整，需要语音和图像视觉等模态的信息进行补充和完善。近年来虽然语音翻译和多模态翻译得到了快速发展，但是结合上下文信息和多模态信息的大范围场景感知的机器翻译方法还有待探索，也将会成为未来的一个研究趋势。

轻量机器翻译模型：机器翻译的应用场景非常丰富，对服务形式的要求也呈现多样化趋势，但是目前最为主要的还是以云端的形式提供机器翻译服务。未来，端侧应用需求越来越多，而机器翻译模型动则上千万的参数规模，难以实现端侧部署。因此，如何在保持译文质量不受显著影响的基础上精简模型参数实现机器翻译模型的轻量化是规模化产业应用的前提，必然是未来的研究趋势。

5.6.参考文献

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, pp. 82–91.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019b. Monotonic infinite lookback attention for simultaneous machine translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 1313–1323.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 194–203.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. In 5th International Conference on Learning Representations, ICLR 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015a. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015b. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 58–68.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In Proceedings of the 2018 Conference on Empirical Methods in Natural

Language Processing, pp. 3028–3033.
Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1538– 1548.
Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 3112–3122.
Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Lo¨ıc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, pp. 2350–2361.
Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021, pp. pages 1317–1324.
Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 1913–1924.
Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 6392–6405.
Chris Callison-Burch, Cameron S. Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta) Evaluation of Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, ACL 2007, pp. 136–158.
Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 76–86.
Zhiming Chen, Yiming Tan, Chenlin Zhang, Qingyu Xiang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017. Improving machine translation quality estimation with neural network features. In Proceedings of the Second Conference on Machine Translation, WMT 2017, pp. 551–555.
Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? arXiv preprintarXiv: 1606.02012,2016.
Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi

Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder‐decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724– 1734.

Kenneth Ward Church and Eduard H. Hovy. 1993. Good applications for crummy machine translation. In: volume 8. 4. Springer, 1993, pp. 239‐258.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pp. 7057‐7067.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978‐2988.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, pp. 493‐499.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers. In 7th International Conference on Learning Representations, ICLR 2019.

Jean-Benoit Delbrouck and Stephane Dupont. 2017. Modulating and attending the source image during encoding improves multimodal translation. In 31st Conference on Neural Information Processing Systems (NIPS 2017).

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 1723‐1732.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2974‐2978.

Desmond Elliott, Stella Frank, and Eva Hasler. 2016. Multilingual Image Description with Neural Sequence Models. In 4th International Conference on Learning Representations, ICLR 2016.

Miquel Esplà-Gomis, Felipe Sánchez Martı´nez, and Mikel L. Forcada. 2015. Ualacant word-level machine translation quality estimation system at WMT 2015. In Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, pp. 309– 315.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 866–875.

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2017a. A convolutional encoder model for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 123–135.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, pp. 1243–1252.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, pp. 6111– 6120.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-Autoregressive neural machine translation. In 6th International Conference on Learning Representations, ICLR 2018.

Jiatao Gu and Xiang Kong. 2021. Fully Non-autoregressive neural machine translation: Tricks of the trade. In Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, pp. 120–133.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O. K. Li. 2017. Learning to translate in real time with neural machine translation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, pp. 1053–1062.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pp. 11179–11189.

Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, pp. 87–92.

Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019a. Non-autoregressive neural machine translation with enhanced decoder input. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pp. 3723–3730.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019b. Star-Transformer. In Proceedings of the 2019 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, pp. 1315–1325.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In Proceedings of the 13th International Workshop on Spoken Language Translation. pp. 1–7.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever. 2021. Unsupervised neural machine translation with generative language models only. arXiv preprint arXiv:2110.05448.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. axXiv preprint arXiv:1803.05567.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander G. Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 8226–8237.

Julia Ive, Andy Mingren Li, Yishu Miao, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. Exploiting multimodal reinforcement learning for simultaneous machine translation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, pp. 3222–3233.

Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. Image translation network. In Image Translation Model.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, pp. 339–351.

Lukasz Kaiser, Aidan N. Gomez, and Franc¸ois Chollet. 2018. Depthwise separable convolutions for neural machine translation. In 6th International Conference on Learning Representations, ICLR 2018.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, pp. 1700–1709.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end English-Japanese speech translation. In Interspeech 2017, 18th Annual Conference of the International Speech

Communication Association, pp. 2630– 2634.
Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019. Unbabel's participation in the WMT19 translation quality estimation shared task. In Proceedings of the Fourth Conference on Machine Translation, WMT 2019, pp. 78–84.
Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In 6th International Conference on Learning Representations, ICLR 2018.
Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5039–5049.
Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1173– 1182.
Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, pp. 8556–8562.
Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. 2017. Modeling source syntax for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 688–697.
Yanyang Li, Tong Xiao, Yinqiao Li, Qiang Wang, Changming Xu, and Jingbo Zhu. 2018. A simple and effective approach to coverage-aware neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, pp. 292–297.
Jindrich Libovicky and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 196–202.
Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin,Zhengyuan Yang, Yubin Ge, Jie Zhou, and JieboLuo. 2020a. Dynamic context-guided capsule network for multimodal machine translation. InMM'20: The 28th ACM International Conference on Multimedia, Virtual Event, pp. 1320–1329.
Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021a. A survey of transformers. axXiv preprint arXiv:2106.04554.
Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In Proceedings of the 2020 Conference

on Empirical Methods in Natural Language Processing, EMNLP 2020, pp. 2649–2663.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021b. Learning language specific sub-network for multilingual machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, pp.293–305,

Lemao Liu, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, pp. 3093–3102.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, pp. 726–742.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, pp. 1128–1132.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020b. Synchronous speech recognition and speech-to-text translation with interactive decoding. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, pp. 8417–8424.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 3025–3036.

Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. Proceedings of the First International Workshop on Natural Language Processing Beyond Text, EMNLP Workshop 2020, pp. 70–74.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, pp. 244– 258.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a

method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311－318.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 261－271.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Lo¨ıc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In 32nd Conference on Neural Information Processing Systems (NIPS 2018).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, pp. 464－468.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, pp. 223－231.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: masked sequence to sequence pre-training for language generation. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, pp. 5926－5936.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019b. Semantic neural machine translation using AMR. Transactions of the Association for Computational Linguistics, pp. 19－31.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pre-training. In MM ’21: ACM Multimedia Conference, pp. 2843－2852.

Lucia Specia, Stella Frank, Khalil Sima’an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, pp. 543－553.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. Transactions of the Association for Computational Linguistics, pp. 313－325.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje,

and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, pp. 230–238.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2021. Self-training for unsupervised neural machine translation in unbalanced training data scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACLHLT 2021, pp. 3975– 3981.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014a. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014b. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014 (NIPS 2014), pp. 3104–3112.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020. Synthesizer: Rethinking self-attention in transformer models. In Proceedings of the 37th International Conference on Machine Learning.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS), pp. 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, pp. 2720–2728.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. 2017a. Deep neural machine

translation with linear associative unit. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 136－145.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 1810－1822.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020a. Linformer: Self-attention with linear complexity. axXiv preprint arXiv:2006.04768.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, YuanFang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 4580－4590.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2017b. Neural machine translation advised by statistical machine translation. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 3330－ 3336.

Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig. 2018a. A tree-based decoder for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4772－4777.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020b. Balancing training for multilingual neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8526－8537.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018b. Three strategies to improve one-to-many multilingual translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2955－ 2960.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019c. A compact and language-sensitive multilingual translation method. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1213－1223.

Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. 2019. Imitation learning for non-autoregressive neural machine translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 1304－1312.

John S. White, Theresa A. O’Connell, and Francis E. O’Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, AMTA 1994.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In 7th International Conference on Learning Representations, ICLR 2019.

Shuangzhi Wu, Dongdong Zhang, Nan Yang, Mu Li, and Ming Zhou. 2017. Sequence-to-dependency neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 698–707.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. axXiv preptint arXiv:1609.08144.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, pp. 6153– 6166.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, pp. 5725–5737.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 4346–4350.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 3025–3035.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or Not? Learning to Schedule Language-Specific Capacity for Multilingual Translation. In 9th International Conference on Learning Representations, ICLR 2021.

Biao Zhang and Rico Sennrich. 2019. A lightweight recurrent network for sequence modeling. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 1538–1548.

Biao Zhang, Deyi Xiong, Jinsong Su, Qian Lin, and Huiji Zhang. 2018.

Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4273–4283.

Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior knowledge integration for neural machine translation using posterior regularization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 1514–1523.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, pp. 1535–1545.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020a. Learning adaptive segmentation policy for simultaneous translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, pp. 2280–2289.

Tianfu Zhang, Heyan Huang, Chong Feng, and Xiaochi Wei. 2020b. Similarity-aware neural machine translation: reducing human translator efforts by leveraging high-potential sentences with translation memory. Neural Computing and Applications (2020), pp. 17623–17635.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simultaneous translation with flexible policy via restricted imitation learning. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pp. 5816–5822.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pre-training and speech translation. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, pp. 12736–12746.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in nonautoregressive machine translation. In 8th International Conference on Learning Representations, ICLR 2020.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. 2021. Distributionally robust multilingual machine translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5664–5674.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. Transactions of the Association for Computational Linguistics, pp. 371–383.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3643– 3653.

第六章　信息检索技术研究进展、现状及趋势

6.1.研究背景与意义

信息检索技术的发展与人类获取信息的需求密切相关。尽管信息检索概念是由 CalvinMooers 在 1951 年最早提出，但包括倒排索引、关键词查询、超链接等在内的各类组织、整理与查找信息的技术则有远为悠久的历史。随着互联网技术和社会信息化的快速发展，人类所能够访问到的信息规模急剧增加，对于更加高效的信息检索技术的迫切需求也成为了这一趋势的必然结果，搜索引擎、推荐系统等以信息检索技术为核心的产品应运而生，改变了我们的生活、学习与工作环境，成为了信息化社会必不可少的基础设施。

作为信息技术特别是智能信息处理技术的重要前沿领域之一，信息检索技术与产品形态近年来一直呈现高速发展演进的趋势。我们认为，这种趋势的驱动力主要来自以下三方面：

首先是**互联网产业推动力的影响**。信息检索技术逐渐从个别产品的功能形态转变成为各类流行产品普适存在的必备技术手段。一方面，互联网用户能够访问的信息资源规模继续以空前速度增长；另一方面，信息资源平台之间相对隔离的现状造成了通用搜索引擎能够满足的信息需求差强人意。各类垂直信息资源平台纷纷推出独立的搜索或推荐系统（如微信搜一搜、头条搜索以及各类电子商务搜索与推荐平台等）以便利用户的信息访问，而这些垂直搜索/推荐系统特异性的功能和信息组织需求推进了信息检索技术的进步，也使得更加便利的统一化的信息获取工具（如对话式搜索系统、智能信息助手等）成为当前研究的热点方向。

其次是**技术内生发展规律的影响**。信息检索技术逐渐从情报科学研究的对象转变为计算机科学和人工智能研究的核心问题之一，促进了对于信息检索研究核心价值和本质性问题的再思考。一方面，研究者开始对于信息检索中的核心概念如相关性（relevance）、匹配（matching）等进行重新思考，并尝试构建相应的计算模型；另一方面，认知科学研究工具的进步（如使用脑机接口设备、眼动设备、皮电设备等对人类的生理信号进行收集分析等）使得对信息检索中的关键性问题如信息需求的产生与满足、相关性判断的机制等进行实证研究成为可能。对这些本质性问题的思考催生了新的研究方向与范式（如探索式搜索、搜索即学习），也酝酿着未来的检索技术产品形态。

第三是**计算方法工具进步的影响**。机器学习技术特别是深度神经网络技

术的进步改变了包括多媒体信息处理、自然语言处理、信息检索在内的智能信息处理技术的面貌。一方面，图神经网络、表示学习等技术的进步使得对于语义和知识的计算一定程度上成为可能，从本质上提升了信息检索系统的性能；另一方面，稠密向量检索、预训练等技术在改进检索性能的同时提出了远高于传统技术的算力需要，也使得信息检索系统的构建方式产生了革命性的改变。由于信息检索直接服务于人类的信息获取过程，深度学习技术与信息检索技术的相互融合和促进也有助于对于更深层次认知规律的研究与理解。

接下来，我们将首先对信息检索领域近年来的发展现状与科学问题进行分析；随后将围绕预训练技术、推荐系统、问答系统、表示与匹配、个性化技术、量子信息检索、用户模型、用户交互、性能评价等主题对信息检索领域的关键技术进行综述，并对信息检索技术的产业应用情况与趋势进行讨论；最后，我们对信息检索技术的未来发展，特别是中国学者可能在其中发挥的作用进行展望。

6.2.领域发展现状与关键科学问题

本章我们将从信息检索的基础理论、核心技术和应用方向三个角度阐述本领域的发展现状。基础理论层面将重点回顾对检索核心概念与评价机制的前沿理论探索，核心技术层面主要介绍近年来新兴技术对检索带来的变革，而应用方向则将介绍检索技术应用的新模式、新问题以及由此带来的新的研究方向。结合上述的研究现状，我们将进一步探讨本领域当前所面临的关键科学挑战。

6.2.1. 理论：超越传统相关性的研究

信息检索的核心是度量用户查询与文档之间的相关性（relevance）。早期的研究从系统和算法的角度出发，认为相关性是查询意图与文档主题/话题的匹配程度[1]，建模方法以相对客观的主题相关性计算为主[2-3]。随着对用户在检索过程中重要性认识的不断深入，相关性越来越多地被认为是用户认知层面判断产出的结果[4-5],它依赖于用户的信息需求状态、知识状态以及对相关信息的感知/认知能力，由此相关性的建模也逐渐转变为以用户为中心的更为主观的计算为主[6-7]。与此同时，相关性认识的转变也带动了评价机制的革新，区别于传统基于TREC/Cranfield的系统评价方法，现有评价方法研究更多地考虑用户的因素，将相关性定义为一个多维度概念，包括检索结果的实用性（utility）[8]、满意度（satisfactory）[5]等等。

近年来，随着检索系统越来越智能化，对信息检索基础理论的研究也涌现出更

多的视角。一方面，由于统计性偏差或者检索算法本身设计的缺陷可能引入歧视性行为，关于检索公平性的研究初现端倪[9]。检索公平性包括结果公平性和过程公平性[9]，研究者们从不同公平主体[10-12]（用户、信息、用户信息组合）、不同公平粒度[13-14]（单次公平、均摊公平）、不同公平目标[15]（基于对待的公平、基于影响的公平）等角度展开研究，提出了衡量整体分布不一致程度的一致公平[16-17]、衡量效益分布和价值分布差异的校准公平[12-13]、无嫉妒公平[18]以及反事实公平[19]等指标，也涌现出了通过调整训练数据集[14]、改进排序和重排序算法（如基于正则化[12]、对抗学习[19]、强化学习[20]、数学规划[21]等）等途径提升检索公平性的新方法。

另一方面，随着检索技术越来越多地被用在医疗、金融、军事、政治等关键领域，可解释性也越来越受到研究人员的重视，如何让检索模型、检索过程、检索结果具有更好的可解释性显得尤为重要。已有工作主要研究检索可解释所需满足的基本性质[22]，结合这些约束要求设计新的模型结构[23]，对检索结果生成解释文本[24]，并利用可视化等技术手段实现检索过程的可解释分析[7]。可解释检索如何评价一直是个难题，线上测试和人工标注[25]条件要求较为苛刻，线下的评价标准还不够成熟，尚未形成检索可解释性的理论基础。

6.2.2. 技术：数据、模型、知识驱动的新技术体系

信息处理技术的不断突破带来了信息检索核心技术的变革与发展，近年来，由大数据、大模型、大算力带来的人工智能领域的一系列新技术范式，也对信息检索的关键技术产生了重大的影响。

随着人工智能模型的发展，深度学习、强化学习、对抗训练、图神经网络等前沿算法驱动了一系列新兴的信息检索技术。以连续向量表征为基础的深度学习模型，凭借其强大的表征学习和建模能力，为信息检索中语义信息表征[26-27]、相关性推理决策[28-29]以及复杂交互过程的建模[30-31]等提供良好的支撑。由于信息检索本质上是关于用户与信息或用户与检索系统之间的交互，强化学习方法被引入到信息检索领域构建交互式检索模型，根据用户的实时反馈不断更新检索策略，并优化用户的预期累积满意度[32]。此外，对抗训练可以通过引入对抗样本或信号获得更加鲁棒的检索模型。现有工作主要基于博弈理论的极小化极大算法，迭代优化生成检索模型和判别检索模型[33]。近两年来，图神经网络的飞速发展，在信息检索领域引起了许多研究者的关注，人们利用图神经网络良好的结构捕捉能力，从查询文档交互信息中提取隐式匹配信号，从而实现检索质量的提升[34]。

在智能模型之外，知识图谱、用户行为等多种来源的知识成为驱动检索技术发展的另一股力量。一方面，研究者将知识图谱中蕴含的丰富知识作为有用的辅助信息引入检索过程中，例如，建模词项和实体之间的交互信息[27,35]，或者将知识图谱嵌入到基于交互的神经排序模型中[36]，不仅能有效应对数据稀疏、语义失配等问题，也能帮助产生多样化、可解释的检索结果；另一方面，研究者利用搜索过程中用户的多种上下文行为所蕴含的隐式知识，包括搜索关键词、浏览结果、点击结果、网站访问情况、书签情况等，对文档进行个性化排序建模，返回更有针对性的搜索结果，从而提高用户体验[37-38]。近年来，基于大规模无监督数据预训练得到的语言模型在自然语言处理领域获得了极大的成功，这种预训练-调优框架也对信息检索领域产生了重大影响。目前，大量工作探索了直接迁移预训练模型的方法[30-31,39]，在信息检索的任务中微调已训练好的预训练模型参数，常用的方法包括将查询-文档对拼接输入到 BERT 中计算相关性得分[30-31]或利用生成式预训练模型 BART 建模文档和查询之间的生成过程[40]，大幅提升了信息检索任务的性能。最近，研究者们开始思考为自然语言处理而设计的预训练模型是否真的可以满足信息检索中语言理解的内在需求问题，面向信息检索定制的预训练方法也应运而生，包括改造预训练模型结构[41]和重构预训练自监督任务等手段[42-43]，这些为信息检索定制的预训练模型在小样本、低资源场景下显示出了更强大的检索性能。

6.2.3. 应用：从搜索、推荐到智能助理

随着社会信息化的不断进步和检索系统能力的提升，更加泛在化、智能化的应用模式不断涌现，信息检索的研究也从传统聚焦于 Web 检索、移动搜索逐渐向问答系统、推荐系统、智能助理等多元化方向发展。

问答系统旨在针对用户提出的自然语言式问题，从海量候选文本中快速定位相关数据、精准找出问题答案，它可被广泛应用于智能客服、智能音箱、智能车载等众多场景。为了保障答案的精准，对信息可信性、正确性的研究变得十分重要，一些工作通过平衡答案覆盖度与系统风险[44]、引入显式的先验知识（例如，知识图谱[45]、共识信息[46]）等方法，以提升问答结果的可靠程度；其次，面向开放域的问答系统，需要在非静态的环境中学习。一个典型的挑战是领域漂移[47]，即不同领域的数据会陆续到达，这就需要模型快速适配到新的领域，同时不能遗忘在以前领域上学习到的知识，持续学习等新方法成为领域适配任务的可能途径[48]；最后，随着近些年深度学习和预训练模型的发展，研究人员也开始尝试突破传统的级联式问答系统，向端到端的问答技术迈进[49]。

推荐系统是与信息检索高度相关却又截然不同的信息获取模式，其本质上是在用户需求不明确的情况下，通过对用户偏好与需求的刻画，自动地从海量内容中寻找符合其感兴趣的信息进行个性化推送。近年来，推荐系统的研究呈现井喷式发展，不仅在工业界得到众多厂商的关注，在主流的信息检索会议中也常常占据主导地位。近年来，研究人员从知识驱动的机器学习入手，尤其是基于因果推理的认知技术，探索不同类型交互行为的内在原因，实现更精准的、可解释的推荐建模。例如，利用知识图谱进行知识增强[50]、基于因果去偏差[51]或因果反事实[52]的技术对推荐系统进行纠偏等，以提升推荐系统的精准度。此外，网络信息资源逐渐从单一内容模态演化到多模态融合信息，相关工作引入跨模态信息[53]缓解交互行为的稀疏性提升推荐效果、或实现跨模态的信息推荐[54]。为了推动推荐系统的发展和落地，TensorFlow Recommenders（TFRS）、RecBole 等开源项目为开发和研究人员提供了方便高效的工具和服务来搭建完整系统。

此外，随着智能设备的不断普及，智能助理慢慢成为搜索系统演进的新形态。智能助理旨在利用信息检索、自然语言处理、语音识别等技术通过统一的对话交互界面来一站式给用户提供需要的信息和服务。目前苹果、谷歌、微软、亚马逊、百度已投入大量资源，积极研发并推出了 Siri、Google Assistant、Alexa、Cortana、小度等具有代表性的智能助理。围绕着智能助理的研究，一些新的研究方向也涌现出来，例如交互式、对话式搜索，用户以自然语言对话的形式与搜索系统进行多轮交互以获取信息、进行决策或者完成任务，在这个过程中，系统不再是被动等待查询、单轮优化的简单模式，而是需要更加主动地辅助用户展开多轮信息交互。目前已有一些相关研究工作，例如通过理解用户复杂模糊的信息需求[55-56]，生成和用户信息需求更契合的问题[57]、平衡主动提问和返回结果之间的风险[58]等技术建模与学习对话式搜索过程，但该研究领域仍然处于起步阶段，对该方向技术的构建以及评估的准则还需要深入研究。

6.2.4. 检索前沿挑战

随着搜索引擎逐渐成为互联网用户认识世界的主要手段，检索系统也将从传统的访问网络信息资源的主要门户逐渐演化成为辅助人们认知与决策的重要渠道之一。这种演化对检索系统所依赖的理论、技术以及评价都提出了新的需求，主要体现在如下三个方面:

•计算理论方面:一方面，相关性作为人类认知行为的产物，具有复杂、动态的特点，目前对相关性的基础理论认识仍缺乏深入的探索。近年来，脑科学与认知科学

的进一步发展可能为相关性理论的认识提供新的动力和方向；另一方面，随着用户对信息需求的精准程度要求越来越高，结果的正确性也成为信息检索系统的重要指标之一，突破信息的相关性走向正确性，构建信息正确性的基础理论也是未来亟需重点探索的方向。

•技术范式方面:在传统“索引-检索-排序”的范式中，检索系统是由多个模块构建的一个复杂系统，难以支持用户复杂的决策过程。我们可以尝试突破这种模式，将传统检索系统的索引、检索和排序组件重构为一个单一的统一模型，来替代长期存在的“pipeline”模式。通过将给定语料库的所有知识编码到一个可以用于广泛任务的模型中，我们有望消除传统方法对索引的依赖，形成 IR 的新范式，为用户信息获取与认知决策的需求提供类似人类专家的高质量答案。

•评价指标方面:除了对相关性进行度量，未来也需要评估信息检索的可信性，确保用户认知的正确。信息检索的可信性可以从三个维度进行评价：检索数据的可靠性和真实性、检索模型和检索过程的可解释性和鲁棒性、以及检索结果的公平性和一致性。如何科学、客观地将检索中多维度、多评价指标问题综合成为一个单指标形式，利用产生的综合评价指数对可信性进行评价，是该领域当前面临的重要瓶颈。

6.3.领域关键技术进展及趋势

6.3.1. 预训练技术

6.3.1.1. 任务定义

近年来，预训练方法受到研究人员的广泛关注，特别是在 2018 年下半年以来，预训练的研究开始呈现爆发式发展，是自然语言处理领域最大的突破之一。预训练方法的思想是，在利用标注数据之前，先利用大规模无标注的语料进行长时间的无监督或是自监督的预先训练（pre-training），获得任务无关的通用语言建模和表示的能力。然后，在具体的下游任务上不需要对模型结构进行较大的改动，只需要在原有模型基础上根据下游任务的需要构造相应的输出层，并使用任务语料对模型进行少量训练，这一步骤被称为微调（fine-tuning）。预训练作为一种新的迁移学习范式几乎提升了所有下游任务的性能，极大的推动了自然语言处理[59-60]、语音识别[61]、计算机视觉[62]等方向的发展，它也因此成为学术界和工业界公认的前瞻性研究领域，相关研究成果在实际场景中得到广泛应用。

6.3.1.1.1. 任务目标

预训练的研究目标是通过在大规模语料上进行自监督学习，让模型能够学习通用的语言学知识提升下游任务的性能。相比于从零开始训练任务模型的方法，预训练模型能够利用大量无标注数据训练得到的语言学知识，使得微调阶段对于标注数据的需求大大降低，大幅提升下游任务的性能。在面向信息检索的预训练研究上，一方面研究关注如何开发利用现有预训练模型中的知识来提升检索模型的性能[63]，另一方面也在探索更加符合信息检索特性的预训练方法[42,64]，进一步提高信息检索系统的能力。

6.3.1.2. 任务进展

当前，预训练技术已经广泛的应用到检索系统的各个模块，根据预训练模型在检索任务中应用深度的不同，基于预训练检索方法的发展大致经历了三个阶段：第一阶段是直接迁移预训练模型，将现有训练好的预训练语言模型直接应用在信息检索的任务中，并根据下游任务微调模型参数；第二阶段是改造预训练模型结构，使其更加符合信息检索任务的特点；第三阶段是重构预训练任务目标，使得模型在预训练阶段就直接学习下游任务需要的知识。

（1）直接迁移预训练模型。预训练模型在信息检索中最早被应用在检索系统的重排序任务中，直接利用 Transformer 结构中的自注意力机制来建模查询和文档的交互，构造基于交互的相关性匹配方法。例如，Nogueira 等人[30]将查询和文档拼接后输入到 Bert 模型中，将输出的[CLS]表达作为二者交互的特征来计算相关性得分。另一种基于表示的相关性匹配方法则是将预训练模型作为查询和文档的语义编码器进行独立编码，然后通过相似度函数计算二者的相关性。例如，Qiao 等人[65]将查询和文档分别输入到 Bert 模型中分别得到两个语义向量，并基于余弦相似度函数来计算最终得分。此外，预训练模型也被应用到查询理解[56]、查询扩展[66]、词重要度估计[67]等检索任务中。

（2）改造预训练模型结构。尽管预训练模型能够有效提升检索任务的性能，但其网络结构多是针对自然语言处理任务的需求而设计，在检索任务中仍存在适配不足的问题。首先，预训练模型的输入长度最大为 512 个字符，难以满足检索任务中文档篇章长度变化极大的挑战，为此研究人员提出了各种段落级的相关性建模方法，典型的模型包括 Bert-FirstP[31]，IDCM[68]，PARADE[69]等。其次，预训练模型包含大规模的网络参数使得计算复杂度极高，难以满足检索任务对于模型低延时的效率需求，一方面，

研究人员通过解耦模型底层的交互将在线计算转化为离线计算提升计算效率，典型的模型有 PreTTR[41]，ColBert[70]等；另一方面，研究人员也提出利用模型蒸馏技术将大模型转变成小模型提升效率的方法，典型的模型有 Distilled Ranker[71], TCT-Colbert[72], Simplified TinyBERT[73]等。

（3）重构预训练任务目标。除了网络结构的适配外，研究人员也探索了设计面向信息检索需求的预训练目标，从大规模语料中学习排序任务的特征，进一步提升检索的性能。例如，Chang 等人[64]根据 Wikipedia 页面的链接关系，提出了三种句子采样的方法来构造查询-文档样本对进行自监督学习；Ma 等人提出了两种代表词预测的任务，即 PROP[42]和 BPROP[43]，在多个检索数据集上都取得了当前最佳的效果；Ma 等人[74]提出了利用互联网语料中大规模的超链接与锚文本来生成伪查询-文档对，构造自监督任务来预训练语言模型的方法。目前，构建面向信息检索的预训练方法仍然处于初步探索阶段，需要更进一步的研究。

6.3.1.3. 任务影响

预训练技术带来了一种全新的学习范式，即预训练-微调的范式，它为自然语言处理、语音识别、计算机视觉等 AI 领域带来了巨大的影响，也深刻的影响了信息检索领域的发展。目前基于预训练的检索模型取得了比以往神经检索模型更好的效果，它已经成为了学术界研究的热点方向。例如在向量检索模型中，预训练模型使得基于向量的检索方法首次超过基于倒排的检索方法[26]，掀起了语义召回模型研究的浪潮[70,75]。此外，在信息检索评测榜单上，基于预训练模型的排序算法已经完全占领了榜单前列[76]，相比传统的方法取得了较大的性能优势，尤其是在更大的数据集上，预训练模型体现了巨大的威力。与此同时，预训练模型在工业界也已经落地应用，包括谷歌 [1]，微软 [2]和百度[77]，都公开发文表示已经将预训练模型应用到了在线搜索服务中，取得了显著的效果。

6.3.1.4. 发展趋势

近年来大多数研究都集中在预训练方法在信息检索中的应用与适配方面，随着预训练研究的不断推进，研究人员开始探索预训练与检索深度耦合的方法，一个重要的趋势是结合检索任务的特性来构造满足信息检索需求的预训练模型。当前前沿的方向有：（1）设计面向信息检索任务的预训练模型，包括构建适合排序任务的预训练模型架构、提出符合相关性建模需求的预训练目标、以及针对检索的预训练模型持续

学习范式。（2）开发利用多源异构数据的预训练检索模型，例如多语言、多模态、或知识增强的预训练模型，利用更丰富的数据资源增强文本表示，从而提高在信息检索任务上的表现和性能。（3）端到端的信息检索系统学习：现有的信息检索系统大多采用“检索-排序-重排序”的多级流水线架构，每个模块的学习过程通常是独立的，目前预训练模型在各个模块都取得了一致的较好效果，从而使得端到端优化成为可能。（4）以模型为中心的检索范式：传统的检索架构依赖额外的外部文档索引，近年来大规模预训练模型在知识编码方面展现了超强的能力，能够直接生成对信息需求的响应，这种以模型为中心的检索范式颠覆了传统的以索引为中心的检索范式，为新一代信息检索应用模式和形态带来了新的机遇和挑战。

6.3.2. 推荐系统

6.3.2.1. 任务定义

推荐系统是一种信息过滤系统，旨在获知用户画像、物品信息以及上下文场景的基础上，通过构建一个函数以预测用户对候选物品的喜好程度，再基于预测分数对候选物品进行排序，生成推荐列表。

6.3.2.2. 任务目标

用户和物品是推荐系统两个基本元素，因此，推荐系统的任务目标可以从用户和物品两个维度来阐述：从用户角度，推荐系统的目标是为用户推荐其感兴趣的物品；从物品角度，推荐系统的目标是将物品推荐给其受众以获得最大的收益。具体来讲，常见的推荐系统的评价指标有以下几类：

•准确性：如何精准地为用户推荐其感兴趣的物品以及如何将物品精准推荐给目标受众

•转化率：如何将物品点击转化为购买率以追求利益最大化

•多样性：如何在保证准确性的同时，为用户提供更加多样化的物品，避免信息茧房问题

•公平性：如何消除推荐系统中存在的性别歧视、年龄偏见以及种族歧视等不公平现象

•可解释性：如何打破推荐模型的黑盒属性，为推荐结果提供解释以提高用户对平台的信任

6.3.2.3. 任务进展

当前推荐系统主要有三个研究方向：（1）基于协同过滤的推荐；（2）基于特征交互的推荐，以及（3）交互式推荐。

（1）基于协同过滤的推荐。此类方法的核心是利用多个用户的协作行为来预测目标用户的行为。早期的工作直接计算用户或物品行为的相似性[78]。之后，矩阵分解模型由于其简单且有效的性质而得到广泛应用[79-80]。近些年来，神经网络与推荐系统结合是研究的主流。其大致可以分为基于历史行为的神经协同过滤，基于图神经网络的协同过滤以及基于其他新兴技术的方法。基于历史行为的算法借助用户的历史行为来学习更好的用户兴趣表征，它们通常将用户交互过的物品的表征进行池化来建模用户的兴趣，也会引入注意力机制来区分不同历史行为的影响[81-83]。基于图神经网络的推荐模型则是将协同信号建模成用户-物品交互二分图的高阶链路，并通过邻居聚合来获得节点的表征[84-86]。近期，有工作表明，利用自监督学习或者在双曲空间进行表征学习能获得更高质量的表征[87-88]。

（2）基于特征交互的推荐。此类方法的核心是引入不同的特征并建模特征之间的高阶交互来捕捉用户和物品之间的关联。分解机模型是这类方法的雏形，它建模了特征间的两两二阶交互[89]。基于此，不同改进算法被提出，例如 FFM 计算域级别的特征交互[90]，NFM 设计了一种二阶交互层作为后续深度网络的输入[91]，AFM 引入注意力机制来区分不同特征交互的重要性[92]，xDeepFM 提出了一种压缩交互网络来显式建模高阶交互[93]，AutoInt 则借助自注意力网络来捕捉更细粒度的交互[94]，FiGNN 结合图神经网络来建模特征的高阶交互[95]。近期，不少工作结合 AutoML 技术来自适应选择有效的特征交互或者对特征进行分组交互，并取得了不错的推荐精度[96-97]。

（3）交互式推荐。交互式推荐是一种动态推荐方法，其通过“提问–反馈”的显式交互模式，一方面可以增加用户在产品使用中的参与感，另一方面也有助于系统更实时地捕获到用户即时偏好，从而进一步提升推荐效果。这种方式打破了传统推荐技术中信息不对称的固有限制，可以在与用户的动态交互中显式地获取用户偏好，从而执行更符合用户当前兴趣的可解释性推荐[98-99]。在交互式推荐这种动态场景中，一种常见的问题建模方式是将用户–系统交互序列建模成一个马尔可夫决策过程，并用基于强化学习的方法作为交互策略，使系统能够自动地根据当前交互状态做出下一步最优的动作[100-101]。这种自动决策的优势在于能够取代繁杂的人工规则。例如，在电商场景中，用户已经购买过的商品以及同类型商品不应

该继续出现在后续推荐列表里。静态模型无法自动学习到这一规则，但基于强化学习的模型能够自动地从用户的反馈中学习到这一规则模式。强化学习的最终目的是实现整体性能最优，即最大化整个交互轨迹的累计收益。在真实场景中，一个精心设计的强化学习模型能够让用户感到持续的满足感、新鲜感，也能让商家获得最大的长期收益[102]。

6.3.2.4. 任务影响

推荐系统在现实世界中有着极其广泛的应用，涉及电子商务、新闻、娱乐以及教育等众多产业，对个人、公司乃至国家都具有重大影响，是当下互联网中一项举足轻重的技术。从宏观角度，推荐系统关乎国计民生。它可以挖掘用户潜在兴趣，提高用户的消费欲望，进而拉动内需，促进消费和产业升级；它也可以引导主流价值观，弘扬正能量，提高国民素质。从微观角度，推荐系统能对用户进行精准的个性化服务，不仅可以提升用户对产品的满意度，也能受益产品的提供方，提高公司利润。

6.3.2.5. 发展趋势

虽然推荐系统在近些年取得了突破性的进展，然而它也存在诸多问题亟待解决。首先，推荐系统中偏差问题非常严重，如果缓解各种类型的偏差，例如流行度偏差、位置偏差、曝光偏差等，是下一代推荐系统的重中之重。因果推断为这类问题提供了一种可行的解决方案，已逐渐成为当下的研究热点。其次，信息茧房问题是当代推荐系统的一大弊病，如何打破信息茧房、消除用户极化现象也是下一代推荐系统需要攻克的难点。此外，推荐系统也承担着价值观导向的责任，对用户进行长期引导，弘扬主流价值观是一个负责任的推荐系统应当具备的能力，这需要研究人员的不懈探索。隐私保护问题是推荐系统的一大难点，如何在个性化服务和隐私保护之间做好平衡是未来的一大研究趋势。

6.3.3. 检索式对话系统

6.3.3.1. 任务定义

在人工智能领域，构建一个能够与人类进行自然而有意义的对话的智能对话系统，一直是一个吸引人但具有挑战性的任务。近年来，社交网络服务的蓬勃发展积累了大量的人类在 Web 上的对话数据，从而鼓励研究人员研究数据驱动的方法来建立开放

领域的对话系统。现有研究大致可分为基于生成的方法或基于检索的方法。前者通过自然语言生成技术直接合成回复。后者从预构建的索引中检索大量候选响应，然后选择一个合适的作为回应。基于检索的模型重用人类对话，从中选取合适的回复，在回复的流畅度和信息量方面要优于基于生成的模型。本章将对检索式对话系统的 Response Selection 模型的体系结构进行简要介绍，并依据匹配方法和模型框架对该课题的最新进展进行分类总结。

6.3.3.2. 任务目标

检索式对话的整个流程可概括为：Query 理解→检索召回→候选回复排序。Response Selection 问题即候选回复排序问题。它可定义为：给定一个话语序列$C=u_1, u_2,...,u_n$作为对话历史，如何建立一个匹配模型$s(C,r)$或是$s(C,S)$来判断候选回复r是C的合适回答的可能性。$s(C,r)$和$s(C,S)$分别表示现有的两种匹配模型(Context-to-Response 和 Context-to-Session)，其中S表示候选回复r所在的会话。匹配模型的学习是一个监督式任务，通常采用使用二元标签y来表示对话历史C与回复r之间的匹配程度。

6.3.3.3. 任务进展

近年来大部分检索式对话的研究集中在(1)Response Selection 匹配方法的研究上，一些学者通过研究(2)学习策略来改进模型的性能。此外，也有大量的研究通过(3)引入外部知识来获取更好的回复。

Response Selection 匹配方法

根据现有的研究本章将匹配模型分为 Context-to-Response 和 Context-to-Session(或 Query-to-Session)两类，由于先前的工作多集中在 Context-to-Response 匹配方法上，Context-to-Session 匹配并没有太多的框架。这里仅对使用 Context-to-Response 匹配方法的模型进行不同框架的分析，依据 Tao 等人[103]的调研工作，主要将其分为：基于表征的匹配框架和基于交互的匹配框架，以及特别地，由于其强大的表征学习和理解能力，基于预训练模型的框架单独归为一类。

·Context-to-Response 匹配方法

基于表征的框架。该框架通常先将话语序列输入表示层，再通过一个聚合函数来获得 context-level 的向量，最后将其与回复 r 进行匹配打分。为了从语篇层面捕捉到重要的全局主题感知线索，Yi 等人[104]提出了新的多轮对话建模的主题感知解决方案，通过无监督方式对主题感知话语进行分割和提取，在话语层面捕捉显著的话题

转移，从而有效地跟踪多轮对话中的话题流动。Jia 等人[105]通过利用基于对话依存关系的对话抽取方法来将对话历史变为多个子线程，并通过 Thread-Encoder 模型学习各子线程的表征向量，最后通过注意层来获取与候选回复的匹配分数，实验结果表明了依存关系对对话上下文理解的作用。

基于交互的框架。基于交互的框架一般先将 query 中的每个话语 u 与回复 r 进行交互匹配，并提取出交互匹配向量，再通过聚合各匹配向量来获得匹配分数。为了挖掘对话中的潜在信息，如用户意图和对话主题等，Deng 等人[106]提出一种包含潜在交互模型的 Intra-/Inter-交互网络来对话语和会话进行多个层面的交互，其主要贡献是提出了层次化模型来捕捉多个层次的匹配信息，以及开发了两种类型的隐层多视角子空间聚类模块来对话语和回复隐层特征的连贯性进行建模。

基于预训练的框架。由于预训练模型展示的强大语言表征和理解能力，一些研究将其用于 Re-sponseSelection 任务。其通过将拼接的话语序列和回复同时输入预训练的多层自注意网络(如 BERT)来完成表示、交互、聚合操作。Humeau 等人[107]提出了 Poly-encoders 架构来权衡效率和性能，通过一个额外学习的注意力机制来学习更多的全局特征，实现自我注意，实验证明该架构比 Bi-encoders 具有更好的性能，比 Cross-encoders 具有更快的速度。多轮对话通常包含着多个说话者对象，探索对话中的说话者信息是 ResponseSelection 任务来说是有意义的工作。Gu 等人[108]提出了一种说话者感知模型 SA-BERT 使模型能够感知说话者的变化信息。Liu 等人[109]通过掩码式解耦融合网络(MDFN)对话语感知信息和说话者感知信息进行解耦，使每个词分别只关注于当前的句子、其他的句子、发生者话语与接收者话语的词，解决了对话中的角色转换和远程文本噪声问题。现有的研究更多地关注话语与反应之间的匹配，基于习得特征计算匹配得分，导致模型推理能力不足。Liu 等人[110]通过一个具有序列推理模块和图推理模块的集成网络对预训练模型进行微调来增强模型的推理能力，并在多轮对话推理基准数据集 Mu-Tual[111]上进行了实验，结果表明其性能显著优于强基线方法，达到接近人类水平的性能。为了处理多数模型在任务导向对话集上无法发挥性能的问题，Wu 等人[112]提出了面向任务的对话 BERT(TOD-BERT)模型。针对对话历史和回复之间的不连贯和不一致问题，Xu 等人[113]提出了一种基于预训练的 context-to-response 匹配模型，该模型带有 4 个辅助自我监督任务，通过与这些辅助任务联合训练，使用模型更好地学习对话数据中包含地任务相关知识，以产生更好的回复选择特征。Wang 等人[114]将 ResponseSelection 作为一个动态话题跟踪任务，通过自监督学习将主题信息纳入预训练，提出了在多方对话中选择应答的 Topic-BERT 方法。

·Context-to-Session 匹配方法

先前的工作在 Context-to-Response 匹配方法上已有了很大的进展，但这些工作都忽略了回复的上下文信息。这些信息可以为选择最合适候选回复提供丰富的信息。当查询信息和回复上下文信息越相似时，候选回复命中的可能性就越高。因此，Fu 等人[115]提出了 Context-to-Session 匹配方法，通过 Context-to-Response 匹配方法与 Context-to-Context 匹配方法(查询的上下文和回复的上下文)集成模型来获取最终的匹配分数。并在进一步的工作[116]中，首次将回复的历史信息和未来信息同时加入考虑，提出了一种对话流感知 Query-to-Session 匹配模型(DF-QSM)，实验在三个基准数据集上都表现出比强基线显著提升的性能。

Response Selection 的学习策略

现有的许多的研究都致力于通过多样的神经单元架构来搭建匹配模型。但如何更好的学习模型，以及如何寻找更好的模型也是非常值得关注的问题。一些工作从学习策略入手来优化模型的性能。由于随机负样本训练的模型在现实世界具体场景并不理想，Su 等人[46]提出了一种分层课程的学习框架，以“易到难”的方案来训练匹配模型。Whang 等人[117]通过一种自监督的话语操作策略(UMS)来帮助模型保持对话的一致性，实验对跨多种语言和模型进行广泛评估，发现 UMS 使模型在多个公共基准数据集上表现出显著的改进。Zhang 等人[118]探索了一种联合匹配方案，使得通过一次匹配就能完成对所有候选回复的预测，将训练时间减少了一半以上，同时他们的工作还提出了一种有效的基于置换的成本低但有效的数据增强方法，进一步地提升了模型的性能。匹配问题的二元标签可能会导致对回复质量多样性的忽视，为了解决这一问题，Lin 等人[119]以现成的回复检索模型和回复生成模型作为自动灰度数据生成器，通过构造不同类型的灰度数据和多级排序目标，使匹配模型更好地捕捉候选回复之间细粒度的质量差异，减少训练集与测试集扰动强度的差异。Ma 等人[120]提出了 PR-Embedding 会话词嵌入方法，使用机器翻译词对齐模型计算跨句共现来捕捉会话对<post，reply>之间的关系，帮助模型在单轮和多轮对话中选择更好的回复。

引入外部知识的 Response Selection 模型

知识是一种对输入信息及其周围环境的意识和理解。它可以从各种信息来源获得，包括但不限于关键字、主题、语言特征、知识库、基础文本和视觉信息。一些研究从模型的输入入手，增加额外的外部知识信息，然后用于加强 Response Selection 的过程。值得注意的是，该节只考虑增加了“外部”知识的模型，由自监督学习产生知识的模型并不归为该节。通过引入外部 document 知识，Gu 等人[121]提出对上下文和知

识进行预过滤，然后利用过滤后的上下文和知识与响应进行匹配。由于现实中的人类对话不仅是文本感知的，还会受图像和视频等的影响。因此，越来越多的工作将多模态运用于对话系统中，Shuster 等人[122]将图像引入对话，并公开 Image-Chat 数据集用于研究多模态对话。研究对话中的语义和隐含的情感信息都有助于选取更加合乎情理的回复，以此来提升回复的质量。目前已有许多的研究通过将情感因素(如用户情感、用户意图、个性化信息等)引入模型来提升回复质量。Qiu 等人[123]提出了一个情绪感知的过渡网络，来对对话的情绪流建模，并进一步设计了一个情绪可 Response Selection 的统一模型。Zandie 和 Mahoor[124]提出了一个可以使用情感、主题和 DA 对应的特定上下文信息的多头 Transformer 架构，来以合适的情绪选取回复。Yang 等人[125]分析了检索式对话的用户意图，提出了一种意图感知的神经排序模型。在共情会话模型在许多领域证明可以提升用户满意度和任务结果，Zhong 等人[126]引入人格角色来对共情会话模型进行改进，并提出了一个新的大规模多领域的基于人格角色的共情对话数据集 PEC。在 Gu 等人[127]的工作中，个性化信息被当作为保持对话系统一致性的先验知识，文章提出了 4 种人物角色融合策略，来探讨了个性化信息中人物角色(自身说话者与搭档说话者)的描述对 Response Selection 任务的影响，在三个代表性的模型上证实了自我/搭档的角色描述对 Response Selection 性能的提升。由于对话中的话题转移而产生的无用话语可能会导致匹配效率的下降，为了解决这一问题，Hua 等人[128]提出了 RSM-DCK 来检测对话上下文与知识集的相关模块。为了处理相同知识多次融入对话导致重复的事实，Sun 等人[129]设计了一种历史适应的知识整合机制来提高对知识的理解和运用，有效提高回复质量。

6.3.3.4. 发展趋势

本章综述了近两年来关于检索式对话中 Response Selection 任务中匹配方法的相关研究，并总结了一些 Response Selection 模型的前沿工作，包括模型的学习策略，外部知识的引入。通过大量的工作，我们在许多基准上都看到了基于检索的对话系统令人可喜的性能表现。虽然如此，该课题的研究仍然存在着许多挑战可以探索。如模型是否真的能有效理解对话历史？尽管现有的工作在选择相关回复方面有着不错的效果，但模型本身仍缺少对对话的理解能力。近年来开始有工作[105,110]尝试挖掘话语中的语义和时间依赖性，以增强模型的推理能力。但目前用于研究模型推理能力的基准数据集仅有 Mutual[112]，在未来仍需引入更多的基准数据集和新的模型。关于对话历史和回复的逻辑一致性也是非常值得研究的方向。现有的

ResponseSelection模型更多关注对话历史和回复的语义性，而忽略了其逻辑一致性。一些模型和学习策略[113,130]为处理该问题提供了一定的帮助，但未来还需要更多的工作来处理这种不一致问题。多轮对话选择模型的领域迁移和领域适应也是非常值得关注的问题，现有的研究通常关注于单一的领域或从固定来源获取人类的对话信息用于模型的训练与测试。然而人类的对话内容总是随着社会发展和语言变迁而改变的，这就导致模型在领域迁移时，并不能有效地理解对话内容，从而使性能大幅度下降。为处理这一问题，一些研究者基于预训练模型进行了相关的工作[29]。但目前可用的会话数据集还远远没有涵盖开放域会话中可能涉及的所有内容。因此，未来的工作可以考虑构建可持续的“进化”对话系统模型，根据各种社交平台上不断更新的开源对话数据进行自我学习和进化。

6.3.4. 问答系统

6.3.4.1. 任务定义

问答系统是服务于用户的信息查询需求的智能系统，是用户得以通过语言来访问海量的、结构化或非结构化、单模态或跨模态知识库的便捷接口。现有常见的问答系统的应用包括搜索引擎、智能客服、语音助手等，这些应用在我们的生活和工作中扮演着非常重要的角色。

6.3.4.2. 任务目标

研究问答系统的目标是让系统具备高水平的问题理解能力，精确的信息检索能力和灵活的复杂推理能力，从而能够准确地解答用户复杂的问题，以更好地满足用户的需求。一种最为自然直接的评估问答系统的方式是统计用户在使用系统之后的满意程度。然而，这种方式显然是昂贵且耗时的，不利于系统的快速迭代。早在 1967 年，Cleverdon 等人[131]就提出收集并重用同一个测试样例集合来测试并对比多个不同系统的表现;一个足够好的问答系统应该能够得到测试问题的预期结果，包括所使用相关文档以及所预测最终答案。这种评估方式一直沿用至今。

6.3.4.3. 研究进展

通常而言，只有工业界才能轻易获得大量真实的用户问题。而利用用户数据开展的研究将有助于直接提升问答系统的用户满意度。微软于 2016 年公开的

MSMARCO 问答数据集[132]以及谷歌于 2019 年公开的 Natural Questions 数据集[133]就是由真实的用户搜索组成；这些数据集吸引了广泛的关注，极大地促进了信息检索和答案抽取技术的发展。然而这些真实的用户查询主要是单跳的事实型问题，即回答问题所需的所有证据信息包含于单个句子或单个事实中，基本上不需要复杂的推理能力以整合多方信息。一个重要的原因在于，现有问答系统的应用——如搜索引擎等——尚不具备足够的智能让用户相信能够处理需要复杂推理的问题，导致复杂问题占据相对较小的比例[134]。为了提升系统的智能，问答系统这一研究领域不断地定位现有系统仍然欠缺的智能行为，构建相应的评测基准，并研究出能够通过这些基准的模型[135]。这些基准通常由针对特定推理能力而精心构造的问题组成，而不一定是真实场景中用户会问的"自然"的问题，例如，关注于多跳推理的 HotpotQA 数据集[136]，逻辑推理的 ReClor 数据集[137]，数值推理的 DROP 数据集[138]，以及常识推理的众多数据集[139]等。依托于这些评测基准的研究将对服务于真实用户的问答系统的构建具有很强的启发意义。

面向单跳事实型问题的开放域问答系统。近两年来，得益于 Natural Questions 等评测基准的提出，预训练模型相关技术的发展，以及 EfficientQA[140]等问答比赛的举办，面向单跳事实型问题的开放域问答系统取得了快速的发展。开放域问答系统通常由检索器、重排器、以及阅读器级连组成：检索器利用稀疏表征[141]或稠密语义表征[26]从大型文本知识库（可包含转换成非结构化文本形式的结构化知识[142]）中高效地检索出与问题最相关的若干证据段落；为进一步缩小证据的范围，一般会采用比检索器具有更强的表达能力的重排器[39]重新评估检索得到的证据段落与问题的相关度，并保留置信度最高的部分段落；最后采用抽取式或生成式阅读器[143]阅读所选证据，并结合问题给出答案。目前基于预训练大模型的开放域问答系统已经能够在 Natural Questions 上取得较高的答案准确度（人工评测）。然而当受限于存储空间而需要使用更小的模型或更少的检索语料时，问答系统的性能仍然有较大的提升空间[140]。另外，近期的研究表明[144]，现有的问答系统还不能很好地泛化到新颖的用户问题上，即当用户问题包含训练期间未见过的实体或呈现出新颖的模式时，系统的答案准确率会有显著的下降。

面向复杂推理问题的问答模型。包括多跳推理、逻辑推理、数值推理和常识推理：

I）多跳推理。多跳推理需要系统整合分散于多个句子、多个段落、甚至多个文档中的信息并推理得到答案。最具影响力的多跳推理数据集包括开放域问答数据集 HotpotQA[136]以及开放域事实验证数据集 HoVer[145]等。多跳问答系统一般由多跳检

索器和阅读器组成：由于多跳问题涉及多个证据，且问题与单个证据之间以及证据与证据之间通常只存在局部的语义重叠，多跳检索器[146-147]需要顺藤摸瓜式地定位证据链条；所抽取证据链条将提供给阅读器生成答案。由于 HotpotQA 等多跳推理数据集中的问题与证据链条之间存在较高的词汇和语义重叠度，目前基于预训练大模型的多跳问答系统在 HotpotQA 上已经取得了不错的证据定位准确率和答案准确率，并且几乎没有太多有意义的提升空间。然而，这并不意味着现有问答系统已经能够具备很好的多跳推理能力。事实上，在一些更具挑战性的问答数据集上，如 StrategyQA[148]等，问题与证据链条之间存在隐式的逻辑关联但词汇和语义的重叠度不高，现有模型在应对这类问题时的多跳推理表现仍然有待提高。

II）逻辑推理。逻辑推理是指，从一些既定事实出发，基于逻辑规则，推理出新的结论的过程。最具影响力的逻辑推理数据集有 LogiQA[149]、ReClor[137]、LSAT[150]等。逻辑推理一般依赖于两个模块——语义解析器和推理器，前者帮助模型理解文本包含的逻辑信息，后者基于逻辑信息和逻辑规则完成推理预测。根据这两个模块的实现方式，现有方法可以分为三类：（1）基于符号系统的推理，这类方法通常会设计一些规则将自然语言转化为逻辑表达式，然后基于符号操作实现可解释的、确定性的推理。例如 Zhong 等人[150]设计规则将自然文本转化为排列组合问题的约束条件，然后使用可满足性判断方法，找出满足约束条件的排列方式。（2）基于神经网络的推理，这类方法依赖神经网络的学习能力，期望模型能够通过数据隐式地学到逻辑推理能力。例如 Clark 等人[151]通过在合成数据上的实验，说明了 Transformer 能够在简单任务上实现隐式的逻辑推理。（3）神经符号系统，这类方法希望结合神经网络和符号系统的优点，神经网络具备强大的学习能力和一定程度的鲁棒性，符号系统的可解释性和组合泛化能力较强。Rocktäschel 等人[152]基于 Backward Chaining 框架，将部分符号操作替换成神经操作，在保有符号系统优点的情况下，提升了学习能力和泛化性。

III）数值推理。数值推理需要系统对输入文本中的相关数字进行运算才能得到答案。最具影响力的数值推理数据集包括 DROP[138]和 MathQA[153]等。这些数据集提供与问题相关的段落，侧重于考察模型对段落的理解和分析，并不需要模型具备检索能力。数值推理的一个挑战在于，当数据集不提供运算过程标注而只提供答案标注时，仅通过答案往往不能唯一确定正确的运算过程，而使用错误的运算过程作为训练数据则会误导模型[154]。现有工作通过使用去噪训练算法[154]或者给模型增加合适的归纳偏置[155]，能够一定程度上缓解误导性运算过程的问题，并且在 DROP 上已经取得了接近人类水平的表现。然而，近期的研究分析暴露

了这些模型的鲁棒性问题[156]，它们会依赖于一些病态的特征来定位答案。另外，当问题与证据的词汇重叠度较低且证据隐藏于更长的文档中时，如IIRC数据集[157]，现有模型仍然无法准确定位证据和执行正确的运算。

IV）常识推理。常识是指人类约定俗成、默认知晓的一类知识，在问答任务中，常识知识通常不会显式地出现在背景信息中，但是可能会影响答案的预测。最具影响力的常识推理数据集包括 CommonsenseQA[158]、SocialIQA[159]、CosmosQA[160]等。为了获取缺失的常识，可以选择从外部的常识知识库中检索相关信息，常用的知识库有Concept-Net[161]、Atomic[162]等。基于检索得到的知识，可以使用图网络或其他模型来整合信息并预测答案[163]。随着对预训练语言模型的研究，越来越多的证据表明其中蕴含着大量的常识知识，因此也可以将预训练模型用作知识库[164-165]。另一方面，一些工作认为预训练模型中已经包含了足够的常识知识，因此可以直接基于预训练模型训练一个常识问答系统，而不需要从知识库中检索知识。考虑到常识推理是人类的一种基本能力，对大多数人而言，不需要额外的学习就能做得很好，因此我们期望问答系统也可以在无监督的场景下实现常识推理能力，现有的无监督常识问答方法大多基于预训练语言模型给出的文本生成概率来实现[166-167]。

6.3.4.4. 任务影响

问答系统是自然语言处理技术的综合应用，涵盖了语言理解、语义解析、信息检索、复杂推理等多种技术领域。对问答系统的研究推动了各个子领域的发展，在一定程度上，问答系统的任务需求指导了各个子领域的研究方向。此外，作为最近的一个热门研究方向，问答系统可以作为一个基础模块用于其他任务，例如事实验证和对话系统，通过将任务分解成多个子问题，可以直接利用问答系统来回答每个子问题，进而完成整个任务。对社会而言，问答系统有很多实际的应用。我们常见的智能语音助手、智能家居设备，可以按照我们的自然语言指令完成相应的操作；许多搜索引擎除了检索功能外，还可以利用问答系统实现更精确的查询；面对大量文档，例如参考资料或说明书，问答系统可以帮助我们从大量冗余的信息中快速找到我们需要的内容。总的来说，在这个信息爆炸的时代，问答系统可以在很大程度上提升我们整合、查找、使用信息的效率。

6.3.4.5. 发展趋势

在过去的几年中，大多数研究针对的都是比较简单的问题，很少涉及复杂推理，

同时比较依赖标注数据。随着研究的不断推进，一个重要的趋势是更加重视解决真实场景中存在的挑战，包括但不限于复杂推理和低资源场景。当前前沿的方向有：（1）泛化性和鲁棒性。在实际应用场景中，用户输入的问题模式可能难以预测，当问题模式较为新颖或者问题涉及新的领域时，模型需要仍然具有较好且稳定的表现。（2）低资源问答系统构建[168]。在许多真实场景中，构建大规模标注数据耗时耗力，甚至难以实现，因此需要能够利用少量的任务资源高效地搭建问答系统。（3）构建具有复杂推理能力的问答系统。许多实际问题需要模型基于逻辑推理、数值推理、常识推理等能力去完成，如何实现这些能力、不同能力如何联合使用、推理模块的组合泛化与可扩展性如何实现，这些都是亟待解决的问题。（4）问题的模糊性与答案的不确定性[169]。许多现有数据集的问题形式都比较理想化，即问题清晰且答案确定，但在实际场景中，由于用户可能对相关信息缺乏了解，输入的问题可能是模糊的、信息不完整的，这时候需要模型能对多种可能性进行综合分析，或者通过交互式问答引导用户给出更具体的信息。

6.3.5. 表示与匹配

6.3.5.1. 任务定义

文本的表示（representation）和匹配（matching）是信息检索的核心概念，**文本表示**旨在将文本数据转化为数字化表达（如：向量）并应用于信息检索任务中，常用的表示模型包括独热表示、分布式表示等。在获得合适的文本表示之后，信息检索的众多任务（如：搜索、推荐等）都可抽象为文本**匹配**任务，文本匹配的目标是从大量的数据中自动提取出词语之间的语义关系，并应用匹配函数度量和输出匹配结果。如何针对不同信息检索任务找到适合的表达方式并利用匹配模型进行解决，已成为信息检索领域的热门研究课题。

6.3.5.2. 任务目标

信息检索中表示与匹配的目标为：构建单词、语句乃至文档级别的表示模型，实现匹配模型与算法服务于不同的检索任务。例如，互联网搜索可归约为用户搜索查询关键词与网页内容的相关性匹配问题，信息推荐可抽象为用户与物品的偏好匹配问题。

6.3.5.3. 任务进展

短文本匹配是近年来的研究重点，研究人员按已经进行了深入的探索并提出了一系列的方法，根据语义特征提取方式的不同，短文本匹配模型可以分成三类，包括：基于表示的模型、基于交互的模型和两者融合的模型。

短文本匹配模型。基于表示的模型假设文本之间的相关性取决于输入文本的各组成的意义。因此这类模型常包含复杂的结构表示文本，用相对简单的评价函数产生文本之间的相关性分数。例如 DSSM(Deep Structured Semantic Model)[170]用相同的方式表示两个输入文本，包括字母三元组（letter-trigram）映射和多层感知机（MLP）转换，应用余弦相似度函数来评估两个文本表示之间的相似度。类似地，卷积神经网络和循环神经网络也被应用于匹配中的文本表示，相关的模型包括 Arc-I[171]、CNTN[172]、CLSM[173]、LSTM-RNN[174]和 MV-LSTM[175]等。基于交互的匹配模型假设文本相关性在关键在于两段输入文本之间的关系而非它们各自的语义，直接从文本交互中不是从文本表征中学习往往会取得更好的效果。因而此类模型常定义文本交互函数和复杂的评价函数分别用来抽象出文本之间的交互关系和计算文本匹配分值，典型的模型包括 MatchPyramid[176], MatchSRNN[177]等。混合模型结合了上述两类模型的优点，例如 DUET[9]使用松散的混合策略，即分别使用一个基于表示的模型和一个基于交互的模型计算文本的匹配分值，再把它们相结合作为最终匹配结果。近年来，注意力机制提供了一种更加紧凑的混合策略，例如在 BERT[59]中，文本的表征通过文本之间的交互而获得，在经过若干次迭代后最终计算得出算文本的匹配分值。

面向搜索的查询-文档匹配。具体到搜索任务中，用户查询和文档的相关性匹配是典型的非对称短查询-长文本匹配问题，由于短文本匹配模型大多基于对称性假设，因而不能直接用于查询-文档匹配中。DRMM[178]利用直方图表示查询中的每个词和文档中每个词的局部交互信号，使用前馈神经网络和门控网络输出查询和文档整体相关性；K-NRM[179]用使用核池化层替代了 DRMM[180]中的直方图表示，是的模型能够端到端地学习网络参数。近年来，基于大规模语料库的预训练模型 BERT[59]被广泛应用查询-文档匹配任务，其使用堆叠的 transformer[180]模块直接对查询和文档进行全局交互计算，取得了良好的效果；为解决 BERT[59]在实际检索中效率问题，ColBERT[70]提出将查询和文档的编码表示部分放在离线阶段进行。

长文档匹配。和短文本相比，长文档通常包含更加丰富复杂的语义和篇章信息，因而长文档间的匹配更具挑战性，也逐渐受到了研究人员的关注。当前长文本匹配任务

有两类代表性解决方案：层次化模型和局部交互模型。受到文档内容和形式通常具有层次化结构的启发，层次化模型将长文档按照一定的层级进行划分，例如SMASH[181]使用双向循环神经网络和注意力机制按照词、句子、段落到篇章的顺序对长文档编码，CDA[182]进一步探索了文档间的注意力机制对长文档匹配的作用。局部交互模型则假设长文档内部分内容的交互结果即可决定文档整体的匹配结果，例如 CIG[183]利用 TextRank[184]提取长文档对的关键词，以关键词所在的句子作为结点，以关键词是否在同一个句子中共现连边构建概念图(concept graph)，再使用图卷积神经网络计算匹配分值，Match-Ignition[185]也采用了类似的方式，筛选出图上 PageRank[186]权重最高的 k 个结点所对应的句子用以计算文档整体匹配分数。

6.3.5.4. 任务影响

匹配已成为信息检索中的一项关键技术，在搜索中，匹配函数常用来度量文档与查询的相关性，在推荐中，匹配模型用来衡量用户对物品的偏好。基于人工或者自动标记的数据，机器学习方法已被广泛应用于匹配函数的构建中，被称为“匹配学习”。近年来，随着海量数据、强大计算资源的出现和先进深度学习技术的发展，学术界和工业界致力于开发面向检索任务的深度语义匹配算法模型，得益于深度学习方法强大的学习能力和对匹配模式的表示与泛化能力，深度语义匹配已成为信息检索领域中的最先进和广泛应用的技术之一。与此同时，深度语义匹配方法的应用领域也不限于信息检索，其可被抽象为一般的学习框架，广泛应用于自然语言处理和跨模态任务中，如：语句复述、自然语言推理、机器阅读理解、跨模态检索等。

6.3.5.5. 发展趋势

随着信息的多源化和海量化发展，当前信息检索中表示与匹配研究的一大趋势是向更广的应用领域过渡，包括由短文本匹配向长文档匹配过渡、由文本匹配向跨模态和多模态数据匹配（如：使用文本检索图像）过渡。这一发展趋势需要寻找更加精准有效的多模态数据表示方法和挖掘不同来源表示信息相关性的匹配模型，以弥合各模态间的“异质性鸿沟”。面向垂直领域寻求能同时满足多种评价指标的匹配模型是当前研究的另一大趋势，例如在保持高精度匹配的同时，保证模型的无偏性、匹配结果的可解释性和公平性等。此外，为了构建更加鲁棒精准的匹配模型，近年来研究者们开始从传统的统计预测过渡到因果推断，以应对信息检索日志数据中存在的各类干扰因素，这一趋势使得因果纠偏与反事实学习技术成为提升表示与匹

配的稳健性和非 IID 场景下泛化性的重要途径。

6.3.6. 个性化检索

6.3.6.1. 任务定义

搜索是人们日常获取信息最常用的途径之一——用户主动输入查询，搜索引擎返回与查询相匹配的搜索结果。然而，由于大部分搜索关键词本身就有多重含义，具有不同兴趣爱好的用户在输入这些查询时所表达的查询意图也是不一样的。例如，查询“Cherry Reviews”中的“Cherry”一词既可能指代“Cherry flower”（樱花），也有可能指代“Cherry keyboard”（樱花牌键盘）。目前的搜索引擎通常不区分用户真实意图，将结果混合在一起返回给用户，导致搜索结果列表难以准确地满足用户的信息需求。个性化搜索为解决这一问题提供了有效的解决方案。针对每一个用户，个性化搜索通过分析该用户的查询历史来挖掘他的兴趣爱好，建模用户画像，明确当前查询所表达的个性化查询意图，最后基于用户画像和个性化查询意图来对候选文档进行重排序，返回搜索结果。

6.3.6.2. 任务目标

个性化搜索的主要目标是对用户的知识背景、兴趣爱好、查询习惯、当前意图等多方面的信息进行建模，并利用这些信息来消除或降低用户当前查询的歧义性，预测用户查询的真实意图，基于真实意图对结果进行排序。最终目标是对于不同用户，搜索引擎可以根据他们兴趣的不同返回满足其真实需要的个性化排序列表，从而提高用户使用搜索引擎的满意度。

和传统的搜索相比，个性化搜索引入了用户兴趣建模，用户获得的是匹配其兴趣爱好的质量更优的排序结果。这一点在本质上和个性化推荐系统是类似的。但当相比于个性化推荐系统，个性化搜索中引入了查询这一复杂因素，因此个性化搜索算法和系统在设计上更为复杂。

6.3.6.3. 任务进展

传统的个性化搜索技术主要通过从用户查询日志中抽取有效特征来对用户兴趣进行建模，例如利用用户点击行为、文档关键词、文档话题等特征来刻画用户兴趣。然而，这些方法通常假设在某一时刻用户的兴趣是固定的，事实上，用户的兴趣

是动态变化的，因此这些方法整体上还有很大改善空间。近些年来，随着深度学习的出现，个性

化搜索的进展主要可以归纳为以下几个阶段，如图 1 所示。

动态建模用户画像。用户兴趣的动态性主要体现在两个方面。首先，用户的兴趣会随着时间发生变化，近期的一些查询行为对于建模用户画像往往更加重要。其次，用户的历史行为对于建模用户兴趣的价值会随着当前查询词的不同而改变。基于这些想法，HRNN+QA 模型[187]被提出，基于深度学习自动捕捉用户兴趣的动态变化规律。然而，该模型只考虑了用户行为的序列信息，却忽视了时间信息，PSTIE 模型[188]首次显式地考虑了用户历史行为的时间间隔信息，更加细致地建模用户画像。

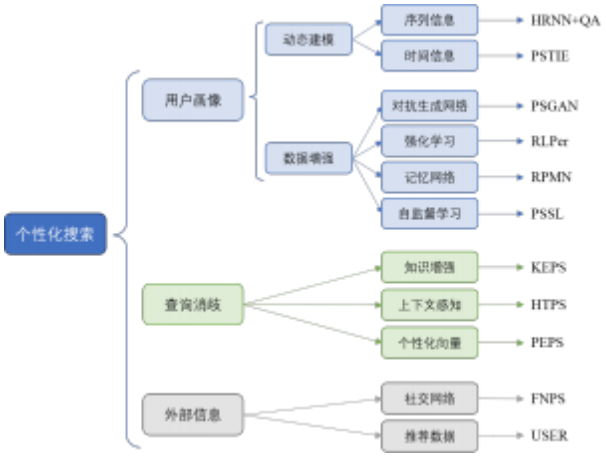


图 1 个性化搜索任务进展

然而，这些深度学习的方法通常会遇到数据稀疏性与数据噪声的问题。为了解决这个问题，多种数据增强的算法被提出。PSGAN 模型[189]提出利用生成对抗网络挑选高质量的数据来训练模型，RPMN 模型[190]提出使用记忆网络来挖掘用户潜在的重新查找行为，从而捕捉细粒度的用户偏好。RLPer 模型[191]提出强化学习框架，通过追踪用户搜索过程来动态学习并实时调整用户兴趣。PSSL 模型[38]提出使用自监督学习进行数据增强，通过对比学习的框架学习高质量的数据表示。

显式查询消歧。上述基于用户画像的方法通过总结用户历史行为来实现个性化，但他们并没有从本质上对当前歧义查询进行消歧，因此构建的用户画像仍然会保留语义偏差。为了加强对查询的理解，KEPS 模型[192]融入知识库信息，通过关联实体来强化对查询的语义理解。HTPS 模型[193]借助上下文感知表示学习的方式来编码历史，使用层次化 Transformer 结构对当前查询进行消歧。PEPS 模型[194]为每个用户训练了个性化的词嵌入表示，彻底抛弃用户画像，转变为个性化的语言模型。

融合外部信息。除了上述算法的创新，越来越多可以辅助个性化搜索的外部信息逐渐被利用起来。例如：FNPS 模型[37]提出利用社交网络信息强化个性化搜索，通

过挖掘朋友的兴趣增强对当前用户兴趣的建模，在这种情况下，即使当前用户缺乏有用的历史信息，也可以为其返回个性化的排序结果。USER 模型[195]融合个性化搜索和推荐，借助用户的搜索数据与浏览数据捕捉更加全面的用户兴趣，同时强化搜索与推荐的质量。

6.3.6.4. 任务影响

通过分析用户的兴趣爱好和个性化查询意图，个性化搜索能够为用户返回更准确的搜索结果，更好地满足用户的信息需求。普通的搜索排序通常只关注查询和文档在广义上的相关性，比如关键词匹配等。然而，在个性化搜索系统中，用户也是很核心的一部分。在进行文档排序时，模型除了考虑文档和查询的相关性，还会同时关注用户的兴趣偏好和个性化查询意图。因此，在个性化搜索系统中，不同的用户都能有个性化的搜索结果，获得更好的搜索体验。

6.3.6.5. 发展趋势

随着深度学习的引入，近年来个性化搜索发展迅速。总的来说，发展趋势可以总结为以下几点：

用户画像由静态变为动态。在传统的个性化搜索中，用户的兴趣画像是静态的——仅仅由该用户的搜索历史集合决定，既没有考虑用户历史行为和兴趣爱好的时间序列性和动态变化性，也和用户当前输入的查询无关。而近年来的模型更关注于捕捉用户兴趣的动态变化以及根据用户当前输入的查询来动态建模相关的用户画像。

2.从精准的关键词匹配到平滑的查询意图匹配。搜索历史中出现过的相似查询以及点击过的相关文档能为当前查询的个性化提供有效的辅助信息。传统的个性化搜索工作主要通过精准的关键词匹配来挖掘重新查找行为，这样会忽略掉很多意图相近但文本相差较大的查询，比如“Java Language”和“Python Program”。现有工作则利用了深度神经网络表示学习的能力，充分挖掘语义和查询意图的相关性。

3.从间接的用户画像到显式的查询消歧、查询表示。个性化搜索的根本目标是消除查询的歧义，明确用户的查询意图。之前的工作倾向于创建用户画像来辅助个性化排序。逐渐地，随着表示学习的发展，现有工作开始转向直接消除用户查询的歧义，得到更准确的查询意图表示。

个性化搜索仍然会不断向前发展，以下几个领域逐渐成为具有潜在价值的研究

方向。

（1）个性化搜索融合其他数据信息。其他有效信息逐渐被引入个性化搜索来提升用户建模质量，比如用户在推荐系统中的行为、用户的社交网络、知识图谱、预训练语言模型等。

（2）个性化搜索带来的隐私保护问题。比如，个性化过程中可能涉及的隐私保护问题，用户行为或数据有偏的问题等。

（3）个性化搜索的关联任务。个性化搜索的模型可以应用到个性化对话、个性化产品搜索、会话式搜索等相关任务上，也可以从类似任务中获取想法。

（4）个人搜索。目前，个性化搜索主要停留于网页文档的搜索，进一步地，可以拓展到个人搜索——为用户创造一个智能信息检索助手，满足用户对任意信息的搜索需求。

6.3.7. 量子信息检索

6.3.7.1. 任务定义

量子信息检索（QIR）指借助量子理论（Quantum Theory,QT)的数学框架来发展信息检索（Information Retrieval,IR）的基础理论与模型架构，以期达到接近人类认知的信息处理水平[196]。量子信息检索的研究可主要分为两个子领域：（1）基于量子理论的表示和排序；（2）基于量子认知理论的用户交互。与量子计算的区别在于，量子信息检索并不涉及物理层面的量子态计算。

6.3.7.2. 任务目标

信息检索的目标是查找与用户需求相关的信息。由于互联网在线信息的激增、用户与系统交互的复杂程度提升，信息检索的理论与系统必须不断改进，以满足用户更高的信息需求。从本质上，信息检索系统的任务可以简化为两个方面。一是如何有效地表示和排序正在创建的各种非结构化信息，二是如何使系统更好地理解用户的复杂信息需求和信息寻求行为。

针对信息检索过程中经典理论无法解决的问题，量子信息检索的目标是通过量子理论的数学框架优化信息检索的基础理论与建模技术，使其信息表示与交互过程更加贴合用户认知。在信息表示与排序层面，借助量子理论的基本表示与数学框架优化信息检索过程中的文本表示，旨在进一步实现文档和查询等信息对象的

抽象化和语境化，充分捕获上下文语境信息的同时实现最优排序。在用户交互层面，量子理论已成功应用于建模和预测非理性决策[197]，并解释人类判断中的认知偏差，量子信息检索期望借助于量子理论与人类认知的契合推动用户交互过程的智能化发展。

6.3.7.3. 任务进展

量子信息检索的发展主要集中与量子理论启发的文本表示与排序、用户交互两个子领域的研究。下面分别概述这两方面的研究进展：

基于量子理论启发的文本表示与排序对查询、文档等信息检索元素进行量子化表示，继而优化文档排序方法。（1）van Rijsbergen 开创性地提出将传统信息检索模型(布尔模型、向量空间模型、概率检索模型)统一在 Hilbert 向量空间中的量子力学形式化框架中，并试图探索用户交互以及信息检索元素（文档、查询等）的形式化描述[198]。该工作启发研究者们通过量子视角研究以用户为中心的检索模式，继而涌现出许多量子信息检索的工作；（2）Sordoni 等人提出受量子力学的启发提出量子语言模型(QLM)[199]。该工作在希尔伯特空间下提出经典概率的量子推广形式并将查询和文档中的复合项表示为叠加事件。Zhang 等人对量子语言模型进行发展，提出基于量子多体的语言建模[200]、端到端的神经量子语言模型[201]、张量空间语言模型[202]等启发性的工作；（3）量子理论启发的排序研究分为基于量子概率和量子测量排序原理。传统的概率排序原理假设“文档与信息需求的相关性不依赖于其他文档”，然而这种假设并不符合实际检索需求，Zuccon 等人、Zhao 等人的工作分别用双缝实验[203]、光子极化实验[204]与检索过程进行类比，建模检索过程中文档之间的类量子干涉现象。

基于量子认知理论用户交互研究包括用户认知建模、查询扩展等，在建模用户动态与上下文信息需求等方面获得阶段性进展。（1）基于量子干涉与投影测量，Jiang 等人从概率相关性的角度分析了神经匹配模型的不足，推导出基于用户认知的检索过程存在类量子干涉项，类量子干涉项用于建模检索过程中匹配单元之间的相互作用所产生的额外证据[205]；（2）量子干涉理论被应用于解决查询扩展中存在的查询漂移问题，Zhang 等人的工作提出基于用户历史信息建模隐式需求的量子干涉检索模型，有效建模隐式信息（历史信息）与显式信息（查询）的关联[206]；（3）量子理论为经典框架无法解释的次序效应提供全新的理论解释与建模视角。在 Wang 等人所进行的用户实验中，当系统向用户展示一对文档进行查询，用户对文档的相关性判断受文档呈现顺序的影响（即用户首先看到一个更相关的文档后，用户对下一个文档的相关

性判断会降低）[207]，这为信息检索中的宏观类量子现象提供有力的实验证明。

6.3.7.4. 任务影响

量子信息检索领域的蓬勃发展对于相关领域产生重要影响：（1）对于计算机领域，基于量子理论的信息检索研究提出了新颖的理论用于建模用户认知，进一步提升了检索模型的有效性和实用性；（2）对于量子研究的交叉领域，量子信息检索为进一步推动量子人工智能、量子机器学习的研究奠定了坚实的理论基础与研究动机。

6.3.7.5. 发展趋势

目前，量子信息检索已经取得阶段性的进展，研究热度不断提升，未来的发展趋势可能涉及以下三个方面：（1）理论层面：进一步探索量子理论与信息检索的内在联系，研究酉演化、自旋、反对称性等性质与自然语言特性是否有合适的对应关系；（2）算法层面：当前量子信息检索的研究成果均面向检索式任务，如何构建面向生成式任务的量子信息检索算法是一项具有挑战的课题；（3）推广层面：当前的量子信息检索仅仅局限于启发式研究，然而如何在量子模拟器等量子硬件环境中部署量子信息检索模型是量子计算时代来临之际亟需研究的问题。

6.3.8. 用户模型

6.3.8.1. 任务定义

用户使用包括搜索引擎、推荐系统在内信息检索系统的过程是一个人-机协同交互的过程。以使用网络搜索引擎完成搜索为例，用户会首先根据自身信息需求，组织查询，提交到网络搜索引擎进行搜索。搜索引擎会根据用户查询，检索相关文档，生成搜索结果页面返回给用户。然后，用户会浏览和检验搜索结果页面，选择有用的结果进行点击。最后，用户会判断其信息需求是否得到满足，以及是否要进行查询改写，提交新的查询开启下一轮搜索。在上述过程中，用户的查询、浏览、点击等行为对完成搜索起到了重要的作用。若缺乏对用户行为和意图的分析、理解和建模，搜索引擎将很难有效的帮助用户完成搜索。因此，近年来信息检索领域研究的重点从传统的“以系统为中心”的模式逐渐向“以用户为中心”或“以交互为中心”发展。越来越多的研究者开始认识到用户的重要性，并逐渐将用户与信息检索的交

互过程作为该领域中最重要的研究对象之一。他们希望通过分析用户与系统的交互过程，更深入的理解用户的行为模式，进而指导信息检索系统的改进；或利用用户交互数据，优化点击率预测、结果排序、性能评价、查询推荐等信息检索系统关键任务的性能。

在“以用户为中心”或“以交互为中心”的研究中，用户模型的设计、构建和应用处于核心地位。广义上说，一切能够描述、刻画、解释、预测、模拟用户在使用信息检索系统时行为的定性或定量的、具体或抽象的模型均属于用户模型的范畴。而针对用户与系统交互过程中的某一类具体的行为，我们可以设计和构建相应的用户模型。例如，我们可以针对用户在搜索结果页面上的浏览和点击行为构建点击模型；针对用户的查询改写行为，构建查询预测模型；针对用户的阅读一条搜索结果并做出相关性判断的行为，构建用户阅读行为模型和相关性判断模型；以及面向搜索评价的需求，构建用户浏览模型和满意度模型。我们还可以整合多个用户模型，对完整的用户-系统交互过程进行建模，模拟真实用户行为，对交互式信息检索系统进行评价和优化。

6.3.8.2. 任务目标

信息检索领域用户模型相关研究的主要目标可以概括为以下三点：

一是**加深对用户行为的认识**：通过构建用户模型，我们可以分析用户在与信息检索系统交互时的行为模式和规律，加深我们对用户以及信息检索过程的理解，进而回答一些信息检索乃至人工智能和认知科学领域内的本质性问题。

二是**准确的预测用户的行为**：通过构建用户模型，尤其是定量的、可计算的用户模型，尽可能准确的预测用户在不同的环境和条件下会做出的行为（例如：准确预测用户查询改写行为，以提供查询推荐；预测用户面对搜索结果列表时的满意度，以评价搜索系统）。

三是**指导和支持系统的改进和优化**：基于对用户行为模式的深入认识，指导信息检索系统的改进（如：基于顺序检验和相关性独立假设，提出经典的概率排序原则（Probability ranking principle, PRP））；或利用可计算的用户模型，从用户交互数据中挖掘有用的信息，优化信息检索系统（如：利用点击模型，从点击数据中挖掘相关性反馈，改进搜索结果排序）。

6.3.8.3. 任务进展

针对上述三个研究目标，近年来用户模型研究的主要进展总结如下：

首先，在加深对用户行为的认识方面，一系列基于用户行为分析和建模的工作，对信息检索的一些本质性、关键性问题进行了卓有成效的探索。例如，通过使用眼动仪记录用户在进行相关性判断时的阅读行为，来自清华大学的研究者提出了一个面向相关性判断任务的阅读行为模型[208]，并发现了人类在进行文档相关性判断时所依赖的一系列启发式规则（Heuristic），从一个新的角度加深了我们对相关性这一信息检索核心概念的认识[209]；通过使用功能性核磁共振成像技术（fMRI）分析用户在进行搜索时大脑的活动，来自格拉斯哥大学的研究者发现在信息需求产生时大脑活跃的区域与没有信息需求时活跃的区域存在明显差别[210]，开辟了一条从大脑和神经活动的层面探究搜索过程的研究路径[211-213]；通过在用户模型和评价指标间建立联系，一系列工作[214-218]系统性地分析了现有离线搜索评价指标背后隐含的假设，共同为搜索评价指标的设计建立了一套较为完整的理论框架，并在该框架下，通过扩展用户模型假设[218-219]及拟合用户模型参数[5,220-222]等方式，设计了一系列能更准确的估计用户满意度的搜索评价指标。

其次，在准确预测用户行为方面，伴随着深度学习技术的快速发展，研究者们开始使用各类基于深度神经网络的用户模型，显著提升了用户模型在预测用户行为方面的准确度。例如，在查询预测模型方面，来自蒙特利尔大学的研究者首先提出了一个基于层次化循环神经网络的模型，相较于传统模型，大幅提升查询预测的准确率和查询推荐的性能[223]。后续的研究进一步将会话中的查询预测和结果排序建模为一个多任务学习问题进行联合优化[224-227]，并采用 Transformer[226]、图神经网络[225]等新技术，有效的提升了会话搜索中的查询推荐和搜索结果排序性能。在点击模型方面，来自阿姆斯特丹大学的研究者首先提出了基于循环神经网络的神经点击模型（Neural Click Model）[228]。由于神经网络模型具有较强的通用性（可以方便的使用不同输入作为特征），后续研究进一步将搜索结果页面上的多模态信息[229]和搜索环境中丰富的上下文信息[230-231]引入神经点击模型的构建中，进一步提升了点击预测性能。

最后，在指导和支持系统的改进和优化方面，信息检索系统的应用场景不断扩展、系统本身的迭代更新、和用户-系统交互范式的演进会使得已有的用户模型不再适用，因此，研究者需要针对新场景、新系统设计和构建新的用户模型，进而指导和支持新系统的改进和优化。例如，针对移动搜索、图片搜索、购物搜索等新场景，来自清华大学的研究者分别设计了考虑结果异质性和点击必要性的移动搜索点

击模型[232-233]，和适用于图片搜索、购物搜索等二维网格界面的点击模型[234]和搜索评价指标[235]。针对探索式搜索、会话式搜索等新的交互范式，研究者开始尝试构建会话级别的用户模型和评价指标[219-220,236]。进一步的，包括 UIUC 的 Cheng-Xiang Zhai 教授在内的一批研究者提出采用模拟生成用户行为（users imulation）的方式对交互式信息检索系统进行统一的评价与优化[229,237-240]。

6.3.8.4. 任务影响

信息检索领域用户模型方面的一系列研究对整个领域的发展产生了重要的影响。具体来说，在研究方法层面，越来越多的研究者开始重视用户在信息检索过程中起到的作用，开始将用户与系统的交互过程作为一个重要的研究对象。同时，通过分析用户行为，构建用户模型，进而指导和支持系统的改进已经成为信息检索领域内的一个新的研究范式[241-242]，得到了较为广泛的应用。

在具体应用层面，基于用户模型设计的各种评价指标和包括用户模拟在内的评价方法被广泛的应用于对各类信息检索系统性能的评价；点击模型成为获取大规模相关性标注数据的重要方法，有效的支撑基于深度神经网络的检索模型（NeuralIR）的训练[243-244]，同时其对位置偏执的建模方式，也是无偏排序学习（Unbiased Learning to Rank）[245-246]和在线排序学习（Online Learning to Rank）[243-244,247-250]等新型研究领域的基础。

6.3.8.5. 发展趋势

最后，我们对信息检索领域用户模型相关的研究的发展趋势做出展望。首先，如前文所述，由于信息检索是个用户与检索系统协同交互的过程，对信息检索系统的改进离不开对用户的分析和建模，同时系统演进会反过来影响用户行为的模式。因此，随着短视频、信息流等系统的广泛使用，以及未来对话式搜索和推荐（conversational search and recommendation）出现，我们需要在未来工作中对用户如何与这些系统进行交互进行分析，进而构建适用于这些场景的用户模型。

同时，人类的认知和决策过程是非常复杂的，我们还远不能说现有的用户模型能够较为精确和完整的刻画和模拟人类用户实际的认知和行为过程。因此，在未来工作中，我们需要结合新的研究方法，进一步探究用户行为背后的原理和机制，以改进现有用户模型。例如，基于人工智能、博弈论和行为经济学领域相关方法，将用户视为一个具有有限理性的智能体，建模其在搜索和推荐场景下的策略性行为

（strategic behavior）：基于认知科学的研究范式，使用眼动仪、脑机接口等设备，收集用户在使用信息检索系统时的信号，进而对用户的认知与决策过程进行更深入的分析，探究其行为背后的原理与机制。

6.3.9. 用户交互

6.3.9.1. 任务定义

随着信息检索领域研究的重点从“以系统为中心”的模式转向“以用户为中心”的发展，信息检索研究越来越关注用户交互行为和用户的认知。目前绝大多数检索系统都记录用户与系统的交互行为，以分析用户使用信息检索系统的行为和偏好，以及用户对信息检索结果和检索系统的满意程度等。最广受关注的搜索交互行为是用户的查询式构建行为和搜索结果的点击和选择行为，但是随着人机交互技术和智能搜索技术的快速发展，信息检索系统已经不局限于支持用户单个查询式请求的功能，而是逐渐发展成为更全面地支持用户完成工作生活中的各类任务的信息系统和智能助手，因此用户与检索系统的交互行为也更加丰富多样，对用户交互行为的分析也需要多维度多层次展开。

6.3.9.2. 任务目标

信息检索领域对于用户交互行为的相关研究主要目标可以概括为以下三点：

一是加深对用户搜索交互行为与认知过程的理解：用户使用信息检索系统的目的是为了满足自己的信息需求，完成相应的任务，所以信息检索系统的设计和优化也应该建立在对用户搜索中的交互行为和及其认知过程的分析基础上。只有全面了解和理解用户的需求和认知特征，才能设计出高效、满意的信息检索系统。

二是利用交互行为构建更优的隐性相关反馈模型：信息检索领域对用户交互行为的关注还源于隐性相关反馈技术，即通过对用户与检索系统的交互行为的记录和分析，构建相关性预测模型、用户特征模型和搜索情境模型等，并基于此实现搜索结果的个性化和情境化。

三是提出更符合用户行为与认知的检索评价指标：信息检索系统的评估不应只注重系统的准确性和效率，还应该充分考虑用户信息搜索的行为特征、最终目标，以及用户对信息检索系统的评价维度，提出能够符合用户行为与认知特征的检索评价指标。

6.3.9.3. 任务进展

针对上述三个研究目标，近年来用户模型研究的主要进展总结如下：

首先，对用户搜索交互行为的理解应基于任务完成的全流程的视角，而不能局限于单个查询式请求。如研究人员依据 Marchionini 的信息搜寻过程模型[251]，将交互行为分为四种类型：计划和理解任务的行为、需求表达的行为、筛选和评估搜索结果的行为、信息使用和任务完成行为。这四种行为也是信息搜索过程的不同阶段，用户在完成任务的过程中这些行为会重复出现且不断迭代，直至任务完成。其中我们经常分析的查询式构建行为属于用户需求表达的行为，除此之外，用户浏览和选择系统推荐的查询式、使用系统的高级搜索和分面搜索中的筛选功能等行为也都属于用户需求表达的行为范畴。搜索结果的点击和选择属于筛选和评估搜索结果的行为，除此之外，用户在搜索过程中也会将相似或相关的多个信息项进行比较和区分，或将重要的信息内容记录或存储下来。最近几年在“搜索即学习”的视角下，学者开始广泛关注支持用户完成学习型任务时的多种交互行为，如学习行为、记录行为等，改进和设计相应的搜索系统功能以支持用户完成学习型任务。

除了分析单个的交互行为，根据研究的需要，学者也尝试将一系列的交互行为组成模块，采用滑动时间窗口建模等方法[252]对交互行为序列展开分析。信息行为领域的[253]曾提出，用户的搜索交互行为可以按照从低到高的层次将其分成：动作（move）、战术（tactic）、战略（strategem）和策略（strategy）四个层次。其中动作是指单个可以观测或识别出的动作或想法，如输入一个检索式或单次点击；战术是指为了完成某个子（小）目标而进行的一系列的动作，如搜索路径（search trails,如[254]）或查询式区间行为（query interval interactions,如[255-256]）；战略指的是为了完成更大的目标而进行的一系列战术，可能是一个搜索会话（session）；策略指的是为了完成整个搜索任务而进行的一系列动作、战术和战略，可能包含若干个搜索会话。对于搜索策略的研究可以使我们从整体上理解用户对整个搜索进程中的情绪、认知和行为的变化特征和对系统的评估方式，进而改进对检索系统的评估[218]。随着检索系统的智能化和多元化发展，用户的交互行为也会变得越来越多样化和碎片化，针对交互行为的分析也面临着更多的挑战。因此从低到高的多层次的分析有助于我们更加有条理地分析和理解用户的搜索交互行为，进而优化和改进检索系统[257]。

其次，信息检索领域对用户交互行为的分析主要应用在隐性相关反馈技术，即通过对用户与检索系统的交互行为的观察和分析，构建相关性预测模型、用户特征模型和搜索情境模型等，并基于此实现搜索结果的个性化和情境化。学者将可作为隐性相关反馈的交互行为按照不同的维度进行分类。如根据搜索阶段可分成：当前搜索前

的交互、当前搜索结果页面上的交互、具体内容页面上的交互、搜索后的信息使用等四个阶段；从信号类型可以分成：关注度、行为、内容三个方面。在这些行为中，点击（click through）、页面停留时间（dwell time）、光标活动行为（cursor movements）等广受关注[258-260]。但很多研究也证实，在基于这些交互行为构建相关反馈模型时，应结合情境因素，尤其是搜索任务类型等，能够有效地提升模型的准确度和搜索效果[261]。最近几年，随着搜索引擎的普及和认知科学研究工具的快速发展，更多具体的交互行为得以观察和记录，如触屏行为、脑电波、眼动行为、情绪识别、皮电刺激反应等，也都可以作为隐性相关反馈或用户建模的依据[262-264]。研究人员除了依据这些交互行为进行相关反馈和用户建模，也尝试根据这些交互行为评估检索系统的用户满意度、沉浸感、焦虑及困难程度等。[265]指出，在使用交互行为作为检索系统评估指标时，一定要结合用户所处的情境，具体情况具体分析，切不可仅根据单个行为的数量和频次直接判断用户对系统的满意度。如较多的检索式或较长的搜索时间可能意味着用户查找和获取信息遇到了困难，但在有些情况下也可能是用户沉浸感和兴趣高的表现。

最后，学者对用户交互行为的分析促进了对现有的信息检索系统评价指标的反思，并在研究中逐步结合用户的成本-收益、有限理性、情感、认知偏见、系统偏差对用户评价系统的影响等，并提出了许多新的系统评估指标[215-216,220,266]。这些融入了用户认知和行为特征的评价指标的提出，使得对信息检索系统的评价越来越接近真实用户的评价，也从一定程度上反映了检索系统对用户需求满足和任务支持。

6.3.9.4. 任务影响

近些年，对交互行为的研究对推进整个信息检索领域的发展起到了关键的作用，推荐系统、对话系统、问答系统、个性化检索、检索评价等多个子领域的发展都离不开对用户的需求、认知、情感和行为的分析。从研究范式而言，将田野观察和访谈、用户实验等用户行为的研究方法引入信息检索研究领域，通过分析用户交互行为，构建用户模型，进而改进和优化信息检索系统的研究范式流行起来。这也使得信息检索系统更有人情味，在人们的日常工作和学习中发挥更大的作用。

从社会发展的而言，建立在用户交互行为分析基础上的信息检索系统提升了个体用户的信息搜索和获取的满意度，智能技术和多场景下的信息系统的开发设计，也促进和保障了整个社会有效获取信息资源。

6.3.9.5. 发展趋势

虽然近些年在用户交互行为方面加大了重视的程度，也取得了长足的进展。但是仍存在诸多问题有待解决，也为未来的交互行为研究在信息检索设计和优化提出了新的挑战。首先，当前的信息检索系统对于用户的计划和理解任务行为、信息的使用和任务完成的行为还没有较充分的支持。如何让检索系统在未来人们工作和生活中发挥更大的作用是我们的目标。最近几年有不少学者的关注到了检索系统对用户任务层面的支持，这也将是未来一段时间的研究重点。

其次，对用户交互行为的分析也不能单纯地停留在单个行为的层面，要结合用户的认知和情感等多维度综合分析。这样有助于我们更全面、更真实地了解用户的信息需求及其当前所处的情境，也能够在结合交互行为评估信息检索系统或构建个性化检索、情境化检索的模型时更加充分和有效。

最后，随着智能技术和新的交互技术、交互模式的出现，用户与系统的交互行为也会越来越丰富多样。我们需要在未来研究中进一步探索用户与智能检索技术的交互行为和认知体验，设计能够让用户更智慧、更高效、更快乐的检索系统。

我们有理由认为未来的搜索无处不在，它不仅能够帮助人们获取信息，更能够帮助用户完成工作和生活中的各类任务。因此未来的搜索系统应结合用户情境和需求的分析，提供更加丰富的功能，从全流程支持用户完成任务；允许用户依据自己的兴趣随意浏览，获得知识的同时收获快乐。随着未来搜索系统的智能化和多场景化发展，用户可以在任何场景下通过多种渠道和方式与信息检索系统交互（如语音助手、增强现实技术、可视化系统等）。最后，检索系统也应考虑到技术的公平和不同群体，尤其是弱势群体（如儿童、弱视、残障等人群）的信息搜寻与获取的可能性和公平性。

6.3.10. 检索评价

6.3.10.1.任务定义

信息检索评价是指采用科学的评价方法、依据合理的评价指标对检索系统的各方面表现进行客观评价，从而帮助研究和开发人员进一步完善检索系统的过程。对检索系统的评价总体上包括多个维度，如检索的**效果（Effectiveness）**、检索的**效率（Efficiency）**、检索系统的**可用性（Usability）**。其中，检索效果主要考虑检索系统返回的结果是否能够满足用户信息需求，帮助用户完成检索任务；检索效率主要考虑检索过程的时间、空间等资源的开销以及系统的响应速度等；检索系统可用性主要考虑

检索系统用户界面的友好度、易用性等，用户是否能够容易地学习如何使用检索系统。由于信息检索的核心任务是帮助用户获取满足他们需求的信息，因此，对于检索效果的评价一直以来都是信息检索评价领域研究的重点。

6.3.10.2.任务目标

总的来说，信息检索评价的主要目标是评价检索系统能够在多大程度上帮助用户满足信息需求，为检索系统的改进提供指导，进而缩小系统表现和用户需求之间的差距，提升用户满意度。为了达成这一目标，我们在评价时需要选择科学的评价方法，设计合理的评价指标，这也是信息检索评价领域面临的关键科学问题和关注的重点研究内容。

6.3.10.3.任务进展

信息检索的评价方法主要分为离线（Offline）评价和在线（Online）评价。离线评价历史悠久，源于 Cranfield 评价范式，通常是使用一套可复用的评测集合来评价不同检索系统在评测集合上的表现差异。评测集合包括三部分：文档集合（也被称为语料库）、信息需求集合（通常由一些具有代表性的查询构成）、以及相关性标注集合（衡量文档与查询的相关性匹配程度）。除了评测集合，选择有效的评价指标也是离线评价的关键一环，评价指标反映的正是检索系统在给定的文档集合和信息需求集合上检索返回的结果相关性的整体表现。当前的研究主要也是围绕评测集合的构建以及评价指标的设计开展，例如：

1）相关性（Relevance）一直都是信息检索的核心概念之一，相应地，查询-文档对的相关性标注是评测集合的重要组成。近年来，随着研究者们开始对相关性这一概念进行重新思考，围绕评测集合中相关性标注的研究也受到了广泛关注。有学者聚焦于相关性的标注方式：Maddalena 等人[267]将心理物理学中的量值估计（Magnitude Estimation）方法用于文档的相关性标注；而 Roitero 等人[268]则是尝试了一种细粒度的百级标注方法（S100），兼顾了相关性标注的灵活性与鲁棒性。也有学者关注不同标注模式下相关性标注集合的质量差异：Chu 等人[269]基于信息熵的概念，提出了成对判别力（Pairwise Discriminative Power）这一指标来衡量相关性标注集合在区分不同检索系统时的判别能力。

2）评价指标（Evaluation Metric）是离线评价方法对不同检索系统在评测集合上表现优劣的直观反馈。由于用户这一角色受到越来越多的关注，过去几年，基于用户

模型的评价指标也成为研究的一大热点：如 Azzopardi 等人[270]和 Zhang 等人[219]分别将信息觅食理论中的边际收益准则和认知心理学中的近因效应引入用户模型的构建中，设计了新的评价指标。

与离线评价不同，在线评价通常不是基于静态的评测集合进行评价，而是在实际生产环境中依据用户使用检索系统时的反馈信息来比较不同系统之间的表现差异。通常，这种反馈信息是隐式（Implicit）的，如点击、查询改写、停留时长等交互行为。在线评价方法通常被认为更适用于工业界，因为学术界往往缺乏实际生产环境和大规模的真实用户。目前，有一些成熟的在线评价方法被大公司广泛应用，如 A/B 测试（A/Btest）和混排（Interleaving）实验，对于系统的迭代更新起到了重要的推动作用。同时，在线评价最近的许多研究也是由工业应用的需求所驱动的，例如：由于 A/B 测试往往耗时耗钱，且存在损害用户体验的风险，基于用户历史日志和因果推断的反事实评价（Counterfactual Evaluation，也被称为离线 A/B 测试）成为当前的一大研究热点。

6.3.10.4.任务影响

信息检索评价一直都是信息检索领域研究的核心问题之一。信息检索领域最高荣誉 Salton 奖的获得者 TefkoSaracevic 就曾指出：“在信息检索的研究和发展中，评价始终是一支主要的力量，它是如此的重要，以至于对系统的新设计及其评价是合二为一的。”[271]当前，信息检索的产品形态及其应用场景相较于传统的搜索引擎都发生了巨大的改变，与之相应的信息检索评价技术也亟需发展。围绕信息检索的新变化构建新的评价体系，将会成为推动信息检索领域持续发展的重要动力。

6.3.10.5.发展趋势

从整体的发展趋势来看，近年来，用户这一角色在信息检索评价中受到的关注越来越多。围绕用户在评价中所发挥的作用，以下问题已经或即将成为接下来信息检索评价研究的重点方向：

1）随着交互式检索的发展，如何对多轮的、会话级的人机交互检索过程进行评价，是当前信息检索评价领域面临的重要挑战；

2）随着认知科学研究工具的进步，更好地理解用户的检索认知过程以对用户进行建模，并将模拟用户应用于评价，将会成为连接离线与在线评价的桥梁；

3）随着用户个性化技术在检索产品的普遍应用，如何对个性化检索进行评价，如

何平衡短期和长期的评价指标等，将成为信息检索评价面临的新课题。

6.4.领域产业发展现状及趋势

6.4.1. 信息检索相关产业概述

信息检索相关技术一直以来在产业界有着广泛的应用。早期信息检索技术，随着雅虎、谷歌和百度等通用搜索引擎的兴起，为用户提供了在海量搜索中查询相关网络内容的业务。近年来，由于互联网用户能够访问的信息资源规模的快速增长，以及信息资源平台之间相对隔离的现状，传统的通用搜索引擎已经逐渐不能满足用户的信息获取需求。近年来，各类垂直信息资源平台纷纷推出独立的搜索或推荐系统（如微信搜一搜、头条搜索以及各类电子商务搜索与推荐平台等）以便利用户的信息访问。同时，推荐系统、问答和对话系统、用户模型和交互模型等相关技术研究路线的快速发展为新的产品和业务落地提供了基础和条件，大量新的信息检索相关业务在传统的通用搜索引擎基础上得到了快速发展，一系列相关的新型产业也在传统搜索引擎以外获得了落地应用。接下来将从信息检索产业分类、信息检索产业现状、和未来发展趋势 3 个方面简要阐述当前信息检索产业状况。

6.4.2. 信息检索产业分类

按照目前信息检索产业相关的服务内容和服务对象，可以将信息检索涉及的工业界应用分为 4 类主要范畴：网络搜索引擎、电子商务、社交网络和社交媒体、和其他相关产业应用。

6.4.2.1. 网络搜索引擎

网络搜索引擎是传统信息检索领域的直接落地应用，即根据用户的输入检索词为用户提供信息检索服务。在原有传统信息检索引擎的基础上，随着自然语言处理技术和深度学习技术的发展近年来在业务领域获得了新的发展。一方面，传统通用搜索引擎，诸如谷歌、Bing、百度和搜狗等仍然为广大用户提供了传统的搜索业务，同时由于信息平台的相对隔离现状，垂直搜索引擎在各个具体领域内得到了快速的发展和应用。

6.4.2.2. 电子商务

近年来，随着用户数量的不断增加，电子商务服务日益成为互联网服务中的最重要产业之一。电子商务应用按照交易对象不同，可以分为企业-消费者（B2C）、企业-企业（B2B）、消费者-消费者（C2C）和企业-政府（B2G）四种模式。电子商务应用的核心业务即电商检索和电商推荐服务，同时智能问答和对话系统也在电商服务中发挥了重要作用。

6.4.2.3. 社交网络与社交媒体

社交网络和社交媒体通过构建线上社交关系将人和信息进行了重构和连接。社交网络和社交媒体在2010年前后随着脸书、推特、微博和微信的发展，近年来，社交网络和社交媒体也开始发生变化，一方面用户数量的快速增加让多样化用户行为下的业务模式更加多元，同时多模态结合下的自媒体和视频短媒体，诸如抖音、快手等应用也获得了快速发展。

6.4.2.4. 其他相关产业应用

诸如医疗大数据、金融科技、智慧司法、物理网等相关产业近年来发展迅速，随着数据和用户规模的迅速增长，在信息检索领域技术的推动下，出现了一系列的相关应用业务。

6.4.2.5. 信息检索产业现状

6.4.2.5.1. 网络搜索引擎

传统的检索技术关注于用户相关查询词与网络文档的匹配和排序问题。随着信息检索技术的发展，相关业务也在发生着变化，在满足用户传统的信息获取需求的同时，用户在搜索引擎中的异构和会话检索行为已经被应用于用户画像和意图预测，从而优化检索效果；同时基于深度学习技术和强化学习技术的在线排序优化技术、个性化检索技术也正在多个通用搜索引擎中得到广泛的应用。近年来，预训练技术在工业界已经获得了广泛的应用和发展。基于预训练的匹配模型和检索模型也正在被应用于百度等搜索引擎的搜索和排序的优化中，基于BERT、GPT-2/3、ERNIE等预训练模型下的排序或者检索模型在线上系统中获得了显著性的效果提升。除此以外，面向用户排序公平性

的相关排序优化算法也正在被微软、谷歌、百度等公司应用在其排序系统中；同时，除传统的通用搜索引擎以外，诸如今日头条等其他领域企业也开始涉足网络搜索业务并获得了巨大的发展。

推荐系统在越来越多的网络搜索引擎中的信息过滤相关业务场景中也发挥着的巨大的作用，在推荐系统技术的支撑下诸如百度 Feed 流等根据用户兴趣实时推送信息等业务场景近年来得到了长足发展。

6.4.2.5.2. 电子商务

近年来，个性化检索和推荐系统在电子商务产业界获得了极为广泛的应用。电子商务相关产业的核心问题集中分布于电子商务检索和电子商务推荐过程中。电子商务检索作为垂直细分领域的商品检索系统，与传统的网络检索系统相比，存在如何解决用户检索词和候选商品语义鸿沟以及用户画像等问题，同时在保证性能的前提下需要在海量在架商品查询中保证系统的实时反馈效率。电子商务推荐则在不同推荐位根据用户画像实时为用户推荐相关商品。与传统推荐系统相比，电子商务推荐系统往往被分为 2 阶段流程：召回优化和排序优化。前者根据用户兴趣从海量商品中获得召回子集，后者则利用多个排序模型针对第一阶段的结果进行重排序。亚马逊、阿里巴巴、京东、拼多多、美团等相关工业界研发人员在这两个相关任务上进行了大量的实践和探索，取得了一系列的显著的研发成果。随着深度学习技术和知识图谱的发展，传统基于 GBDT 等算法的召回优化和排序优化策略也在发生着变化。在电商推荐过程中的领域知识，诸如商品的关系、品牌、店铺、类别等信息，特别是商品互补和可替代关系的识别，作为电商知识图谱的典型应用，在商品召回领域得到了大量关注。个性化推荐技术和强化学习的推荐算法在阿里巴巴、京东等电商平台上落地后，显著性的提升了电商服务的各项指标。在自然语言处理技术和多媒体技术的支撑下，近年来，智能客服也在电商领域发挥了越来越重要的作用，从而节约了大量的人力成本并提升了用户购物效率。

6.4.2.5.3. 社交网络和社交媒体

社交网络和社交媒体在 2010 年前后随着脸书、推特、微博和微信的推出得到了巨大的发展并吸引了越来越多的用户。社交网络和社交媒体已经日益成为了大部分人生活必不可少的一部分。近年来，在异构图模型、信息扩展和推荐系统相关应用在社交网络和社交媒体各个平台上得到了应用。从传统社交媒体上基于关注/被关

注关系发展来的简单的协同过滤推荐方法发展成基于用户画像和用户多维度兴趣的个性化社交媒体和社交网络推荐服务。同时面向垂直领域，领域知识的引入让诸如微博和微信等平台上的检索功能更加强大。随着移动互联网和诸如 5G 技术的发展，多媒体媒体，诸如抖音和快手等平台，得到了快速的发展并获得了海量的用户群体。其平台上包含的大量多模态和跨模态信息处理也需要多媒体检索相关技术发挥作用。基于哈希算法的多媒体检索技术上近年来得到了广泛的应用，并取得了显著性的效果。

6.4.2.5.4. 其他产业应用

信息检索技术也在其他相关产业，例如医疗大数据、在线广告、金融科技（Fintech）、智慧司法和物联网等领域获得了越来越多的关注和应用落地。中国的医疗健康产业近年来随着相关政策性文件的推出，获得了巨大的变革和发展。其中医疗检索、医疗问答和公共卫生智能监测等业务和应用近年来得到了长足的进展。金融科技中涉及到了基于海量信息的检索和匹配问题，同时面向金融的各类预测问题也与信息检索相关技术高度相关。智慧司法中的智能法务，例如合同内容理解和合同要素抽取、司法摘要、类案检索以及自动判决预测等任务也吸引了业界关注和多项落地示范应用。移动物联网的发展为连接虚拟和现实世界提供了“万物互联”的场景，信息检索技术也开始在相关产业的领域发挥作用。

6.4.3. 未来发展趋势

随着超大规模预训练模型在语义理解上所展示良好效果，考虑到工业界大运算量和海量用户数据的前提下，未来信息检索产业会更加依赖大规模预训练语言模型在检索相关业务上获得更好的效果。同时，多模态的数据内容也为未来的产业发展提供了大量的素材和条件，近年来多模态跨模态等应用开始在相关产品上获得应用和落地，在此基础上的统一化的信息获取工具（如对话式搜索系统和智能信息助手等）将提供与传统检索和推荐方法不一样的新型信息检索业务服务。

6.5.总结与展望

以上，我们对信息检索领域近期的热点技术进行了概要性的综述，并对学术界和产业界未来一段时间可能的发展趋势做了展望。需要指出的是，我们所进行的综述与展望仅是基于已有趋势的剖析，其目的也是为信息检索研究领域的学者提供参考，而不是对未来技术的“预测”。

2004年，信息检索领域的顶尖学者们首次组织了国际信息检索研究战略研讨会（Strategic Workshop on Information Retrieval in Lorne[3]，SWIRL），研讨会根据信息检索领域技术发展的情况不定期召开，目前共召开过三届（2004，2012，2018）。第三届研讨会上，学者们组织对以往研讨会未能成功预测的技术趋势（What do you think previous SWIRL attendees did not predict about the future of Information Retrieval）进行了投票，基于深度神经网络的检索方法（NeuralIR）和机器学习在检索中的普遍应用（ML Domination）高居票数前两位。这说明，对技术发展趋势的误判是常态，准确预知趋势才是偶然。事实上，由于信息检索技术本身的发展方兴未艾，互联网搜索引擎、推荐系统、电子商务平台等密集应用信息检索技术的产品形态进化更是一日千里，任何这种形式的“预测”是不负责任也很难准确的。

尽管对于技术发展趋势的预测是十分困难的任务，但对于信息检索研究力量的变化预测则相对容易。近年来，来自中国的研究学者越来越成为国际信息检索领域的骨干组成力量。自2018年开始至今，信息检索领域旗舰国际会议SIGIR中来自中国的长文投稿量和录用量均超过美国跃居全球各国首位。近年来，中国学者在信息检索领域各重要学术会议（如SIGIR，WSDM，CIKM，theWebConf）和学术期刊（如ACMTOIS，IRJ，FnTIR）中开始承担包括大会主席、程序委员会主席、主编在内的主要职务。SIGIR会议继2011年以后于2020年再次在国内召开，清华大学张敏副教授成为ACMSIGIR执行委员会首位来自中国的成员，ACMSIGIR首个国内分支机构Beijing Chapter成立，这些在信息检索发展历史上具有重要意义的事件都反映了学术界与产业界“中国力量”的崛起。

面向未来，一方面，作为国际信息检索领域日益崛起的主要力量，我们需要承担起更多的引领性责任，在关系信息检索技术发展的关键核心性问题上、在学术共同体制定话题的过程中发出中国声音、作出中国贡献；另一方面，我们更需要秉持“四个面向”的科技创新方向，引导全球力量对中文环境下的挑战性问题开展研究，推动信息检索前沿技术在关系国民经济发展、公共安全等在内的战略核心领域更好应用。在这个过程中，中文信息学会信息检索专委会作为国内信息检索学者主要的学术社区必将发挥更加重要的作用。

6.6.参考文献

[1] Cuadra C, Katter R. Experimental studies of relevance judgements: Final report. i: Project summary[M]//NSF Report No. TM- 3520/001/00. System Development Corporation Santa Monica, CA, 1967.

[2] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613- 620.
[3] Robertson S E, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]// SIGIR'94. Springer, 1994: 232-241.
[4] Harter S P. Psychological relevance and information science[J]. Journal of the American Society for information Science, 1992, 43(9): 602-615.
[5] Liu M, Mao J, Liu Y, et al. Investigating cognitive effects in session-level search user satisfaction[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 923-931.
[6] Fan Y, Guo J, Lan Y, et al. Learning visual features from snapshots for web search[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 247-256.
[7] Wu Z, Mao J, Liu Y, et al. Investigating passage-level relevance and its role in document-level relevance judgment[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 605-614.
[8] Jones G J, Belkin N J, Kando N, et al. Third workshop on evaluation of personalisation in information retrieval (wepir 2020) in memoriam seamus lawless[C]//Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. 2020: 488- 491.
[9] Lee M K, Jain A, Cha H J, et al. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation[J]. Proceedings of the ACM on HumanComputer Interaction, 2019, 3(CSCW): 1-26.
[10] Islam R, Keya K N, Zeng Z, et al. Debiasing career recommendations with neural fair collaborative filtering[C]//Proceedings of the Web Conference 2021. 2021: 3779-3790.
[11] Wu C, Wu F, Wang X, et al. Fairness-aware news recommendation with decomposed adversarial learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 4462-4469.
[12] Wan M, Ni J, Misra R, et al. Addressing marketing bias in productrecommendations[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 618-626.
[13] Steck H. Calibrated recommendations[C]//Proceedings of the12th ACM conference on recommender systems. 2018: 154-162.
[14] Ekstrand M D, Tian M, Azpiazu I M, et al. All the cool kids,how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness[C]//Conference on fairness, accountability and transparency. PMLR, 2018: 172-186.
[15] Morik M, Singh A, Hong J, et al. Controlling fairness and biasin dynamic learning-to-rank[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 429-438.
[16] Zehlike M, Castillo C. Reducing disparate exposure in ranking:A learning to rank approach[C]//Proceedings of The Web Conference 2020. 2020: 2849-2855.
[17] Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness[C]//Proceedings of the 3rd innovations in theoretical computer science conference. 2012: 214-226.

[18] Serbos D, Qi S, Mamoulis N, et al. Fairness in package-to-grouprecommendations[C]//Proceedings of the 26th international conference on world wide web. 2017: 371-379.
[19] Li Y, Chen H, Xu S, et al. Towards personalized fairness based oncausal notion[J]. arXiv preprint arXiv:2105.09829, 2021.
[20] Ge Y, Liu S, Gao R, et al. Towards long-term fairness in recommendation[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 445-453.
[21] Li Y, Chen H, Fu Z, et al. User-oriented fairness in recommendation[C]//Proceedings of the Web Conference 2021. 2021: 624- 632.
[22] Sen P, Ganguly D, Verma M, et al. The curious case of ir explainability: Explaining document scores within and across ranking models[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2069-2072.
[23] Singh J, Anand A. Posthoc interpretability of learning torank models using secondary training data[J]. arXiv preprint arXiv:1806.11330, 2018.
[24] Singh J, Anand A. Exs: Explainable search using local modelagnostic interpretability[C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019: 770-773.
[25] Rahimi R, Kim Y, Zamani H, et al. Explaining documents' relevance to search queries[J]. arXiv preprint arXiv:2111.01314, 2021.
[26] Karpukhin V, Oğuz B, Min S, et al. Dense passage retrieval for open-domain question answering[J]. arXiv preprint arXiv:2004.04906, 2020.
[27] Xiong L, Xiong C, Li Y, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval[J]. arXiv preprint arXiv:2007.00808, 2020.
[28] Fan Y, Guo J, Lan Y, et al. Modeling diverse relevance patternsin adhoc retrieval[C]//The 41st international ACM SIGIR confer ence on research & development in information retrieval. 2018: 375-384.
[29] Wu Z, Mao J, Liu Y, et al. Leveraging passage-level cumulativegain for document ranking[C]//Proceedings of The Web Conference 2020. 2020: 2421-2431.
[30] Nogueira R, Cho K. Passage reranking with bert[J]. arXivpreprint arXiv:1901.04085, 2019.
[31] Dai Z, Callan J. Deeper text understanding for ir with contextualneural language modeling[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 985-988.
[32] Zhang W, Zhao X, Zhao L, et al. Deep reinforcement learning for information retrieval: Fundamentals and advances[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2468- 2471.
[33] Wang J, Yu L, Zhang W, et al. Irgan: A minimax game for unifyinggenerative and discriminative information retrieval models[C]// Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017: 515- 524.
[34] Li X, de Rijke M, Liu Y, et al. Learning better representations for neural information retrieval with graph information[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 795-804.

[35] Reinanda R, Meij E, de Rijke M, et al. Knowledge graphs: Aninformation retrieval perspective[M]. Now Publishers, 2020.
[36] Liu Z, Xiong C, Sun M, et al. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval[J]. arXiv preprint arXiv:1805.07591, 2018.
[37] Zhou Y, Dou Z, Wei B, et al. Group based personalized search byintegrating search behaviour and friend network[J]. 2021.
[38] Zhou Y, Dou Z, Zhu Y, et al. Pssl: Self-supervised learning forpersonalized search with contrastive sampling[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 2749-2758.
[39] Nogueira R, Jiang Z, Lin J. Document ranking with apretrained sequence-to-sequence model[J]. arXiv preprint arXiv:2003.06713, 2020.
[40] dos Santos C, Ma X, Nallapati R, et al. Beyond [cls] through ranking by generation[C]//Proceedings ofthe 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 1722-1727.
[41] MacAvaney S, Nardini F M, Perego R, et al. Efficient document reranking for transformers by precomputing term representations[C]//Proceedings of the 43rd International ACM SIGIRConference on Research and Development in Information Retrieval. 2020: 49-58.
[42] Ma X, Guo J, Zhang R, et al. Prop: Pre-training with representative words prediction for ad-hoc retrieval[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 283-291.
[43] Ma X, Guo J, Zhang R, et al. Bprop: Bootstrapped pre-training with representative words prediction for ad-hoc retrieval[J]. arXiv preprint arXiv:2104.09791, 2021.
[44] Su L, Guo J, Fan Y, et al. Controlling risk of web question answering[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 115-124.
[45] Qin K, Li C, Pavlu V, et al. Improving query graph generation for complex question answering over knowledge base[C]// Proceedings ofthe 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 4201-4207.
[46] Su L, Zhang R, Guo J, et al. Beyond relevance: Trustworthy answer selection via consensus verification[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 562-570.
[47] Li S, Cao J, Sridhar M, et al. Zero-shot generalization in dialogstate tracking through generative question answering[J]. arXiv preprint arXiv:2101.08333, 2021.
[48] Su L, Guo J, Zhang R, et al. Continual domain adaptation for machine reading comprehension[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 1395-1404.
[49] Qu Y, Ding Y, Liu J, et al. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5835-5847.
[50] Lin Y, Xu B, Feng J, et al. Knowledge-enhanced recommendationusing item embedding and path attention[J]. Knowledge-Based Systems, 2021, 233: 107484.

[51] Guo R, Zhao X, Henderson A, et al. Debiasing grid-based product search in ecommerce[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2852-2860.
[52] Yang M, Dai Q, Dong Z, et al. Top-n recommendation withcounterfactual user preference simulation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 2342-2351.
[53] Truong Q T, Salah A, Tran T B, et al. Exploring cross-modalityutilization in recommender systems[J]. IEEE Internet Computing, 2021.
[54] Pretet L, Richard G, Peeters G. Cross-modal music-video rec-ommendation: A study of design choices[J]. arXiv preprint arXiv:2104.14799, 2021.
[55] Ren G, Ni X, Malik M, et al. Conversational query understandingusing sequence to sequence modeling[C]//Proceedings ofthe 2018 World Wide Web Conference. 2018: 1715-1724.
[56] Lin S C, Yang J H, Nogueira R, et al. Query reformulation using query history for passage retrieval in conversational search[J]. arXiv preprint arXiv:2005.02230, 2020.
[57] Krasakis A M, Aliannejadi M, Voskarides N, et al. Analysing theeffect of clarifying questions on document ranking in conversational search[C]//Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. 2020: 129-132.
[58] Wang Z, Ai Q. Controlling the risk of conversational search viareinforcement learning[C]//Proceedings of the Web Conference 2021. 2021: 1968-1977.
[59] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deepbidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[60] Radford A, Narasimhan K, Salimans T, et al. Improving languageunderstanding by generative pre-training[J]. 2018.
[61] Baevski A, Zhou H, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. arXiv preprint arXiv:2006.11477, 2020.
[62] Chen T, Kornblith S, Norouzi M, et al. A simple frameworkfor contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.
[63] Lin J, Nogueira R, Yates A. Pretrained transformers for text ranking: Bert and beyond[J]. Synthesis Lectures on Human Language Technologies, 2021, 14(4): 1-325.
[64] Chang W C, Yu F X, Chang Y W, et al. Pre-training tasksfor embedding-based large-scale retrieval[J]. arXiv preprint arXiv:2002.03932, 2020.
[65] Qiao Y, Xiong C, Liu Z, et al. Understanding the behaviors of bertin ranking[J]. arXiv preprint arXiv:1904.07531, 2019.
[66] Zheng Z, Hui K, He B, et al. Bert-qe: contextualizedquery expansion for document re-ranking[J]. arXiv preprint arXiv:2009.07258, 2020.
[67] Dai Z, Callan J. Context-aware term weighting for first stagepassage retrieval[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1533-1536.
[68] Hofstätter S, Mitra B, Zamani H, et al. Intra-document cascading:Learning to select passages for neural document ranking[J]. arXiv preprint arXiv:2105.09816, 2021.
[69] Li C, Yates A, MacAvaney S, et al. Parade: Passage representation aggregation for document reranking[J]. arXiv preprint arXiv:2008.09093, 2020.

[70] Khattab O, Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert[C]// Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020: 39-48.
[71] Gao L, Dai Z, Callan J. Understanding bert rankers under distillation[J]. arXiv preprint arXiv:2007.11088, 2020.
[72] Lin S C, Yang J H, Lin J. In-batch negatives for knowledgedistillation with tightly-coupled teachers for dense retrieval[C]//Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). 2021: 163-173.
[73] Chen X, He B, Hui K, et al. Simplified tinybert: Knowledge distillation for document retrieval[C]//European Conference on Information Retrieval. Springer, 2021: 241-248.
[74] Ma Z, Dou Z, Xu W, et al. Pre-training for ad-hoc retrieval: Hyperlink is also you need[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 1212-1221.
[75] Cai Y, Fan Y, Guo J, et al. Semantic models for the first-stage retrieval: A comprehensive review[J]. arXiv preprint arXiv:2103.04831, 2021.
[76] Craswell N, Mitra B, Yilmaz E, et al. Ms marco: Benchmarking ranking models in the large-data regime[J]. arXiv preprint arXiv:2105.04021, 2021.
[77] Zou L, Zhang S, Cai H, et al. Pre-trained language modelbased ranking in baidu search[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 4014-4022.
[78] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborativefiltering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. 2001: 285-295.
[79] Koren Y, Bell R, Volinsky C. Matrix factorization techniques forrecommender systems[J]. Computer, 2009, 42(8): 30-37.
[80] Rendle S, Freudenthaler C, Gantner Z, et al. Bpr: Bayesianpersonalized ranking from implicit feedback[J]. arXiv preprint arXiv:1205.2618, 2012.
[81] Kabbur S, Ning X, Karypis G. Fism: factored item similarity models for top-n recommender systems[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 659-667.
[82] Koren Y. Factorization meets the neighborhood: a multifacetedcollaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008: 426-434.
[83] Zhou G, Zhu X, Song C, et al. Deep interest network forclick-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1059-1068.
[84] Wang X, He X, Wang M, et al. Neural graph collaborative filtering[C]//Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval. 2019: 165-174.
[85] Ying R, He R, Chen K, et al. Graph convolutional neural networksfor web-scale recommender systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 974-983.

[86] He X, Deng K, Wang X, et al. Lightgcn: Simplifying andpowering graph convolution network for recommendation[C]// Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020: 639- 648.
[87] Wu J, Wang X, Feng F, et al. Self-supervised graph learning forrecommendation[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 726-735.
[88] Li Y, Chen H, Sun X, et al. Hyperbolic hypergraphs for sequentialrecommendation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 988-997.
[89] Rendle S. Factorization machines[C]//2010 IEEE Internationalconference on data mining. IEEE, 2010: 995-1000.
[90] Juan Y, Zhuang Y, Chin W S, et al. Field-aware factorization machines for ctr prediction[C]//Proceedings of the 10th ACM conference on recommender systems. 2016: 43-50.
[91] He X, Chua T S. Neural factorization machines for sparse predictive analytics[C]//Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017: 355-364.
[92] Xiao J, Ye H, He X, et al. Attentional factorization machines:Learning the weight of feature interactions via attention networks [J]. arXiv preprint arXiv:1708.04617, 2017.
[93] Lian J, Zhou X, Zhang F, et al. xdeepfm: Combining explicitand implicit feature interactions for recommender systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1754-1763.
[94] Song W, Shi C, Xiao Z, et al. Autoint: Automatic feature interaction learning via self-attentive neural networks[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1161-1170.
[95] Li Z, Cui Z, Wu S, et al. Fi-gnn: Modeling feature interactions viagraph neural networks for ctr prediction[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 539-548.
[96] Liu B, Xue N, Guo H, et al. Autogroup: Automatic featuregrouping for modelling explicit high-order feature interactions in ctr prediction[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 199-208.
[97] Liu B, Zhu C, Li G, et al. Autofis: Automatic feature interactionselection in factorization models for click-through rate prediction [C]//Proceedings ofthe 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2636-2645.
[98] Gao C, Lei W, He X, et al. Advances and challenges in conversational recommender systems: A survey[J]. arXiv preprint arXiv:2101.09459, 2021.
[99] Zhang Y, Chen X. Explainable recommendation: A survey and new perspectives[J]. arXiv preprint arXiv:1804.11192, 2018.
[100] Afsar M M, Crump T, Far B. Reinforcement learningbased recommender systems: A survey[J]. arXiv preprint arXiv:2101.06286, 2021.
[101] Chen X, Yao L, McAuley J, et al. A survey of deep reinforcementlearning in recommender systems: A systematic review and future directions[J]. arXiv preprint arXiv:2109.03540, 2021.

[102] Wang K, Zou Z, Deng Q, et al. Reinforcement learning with a disentangled universal value function for item recommendation [J]. arXiv preprint arXiv:2104.02981, 2021.
[103] Tao C, Feng J, Yan R, et al. A survey on response selection forretrieval-based dialogues[C]//IJCAI, 2021.
[104] Xu Y, Zhao H, Zhang Z. Topic-aware multi-turn dialogue modeling[C]//The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). 2021.
[105] Jia Q, Liu Y, Ren S, et al. Multi-turn response selection using dialogue dependency relations[J]. arXiv preprint arXiv:2010.01502, 2020.
[106] Deng Y, Zhang W, Lam W. Intra-/inter-interaction network withlatent interaction modeling for multi-turn response selection[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 4981-4992.
[107] Humeau S, Shuster K, Lachaux M A, et al. Poly-encoders: Transformer architectures and pre-training strategies for fast and accu-rate multi-sentence scoring[J]. arXiv preprint arXiv:1905.01969, 2019.
[108] Gu J C, Li T, Liu Q, et al. Speaker-aware bert for multi-turnresponse selection in retrieval-based chatbots[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2041-2044.
[109] Liu L, Zhang Z, Zhao H, et al. Filling the gap of utterance-awareand speaker-aware representation for multi-turn dialogue[C]//The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI- 21). 2021.
[110] Liu Y, Feng S, Wang D, et al. A graph reasoning networkfor multi-turn response selection via customized pre-training[J]. arXiv preprint arXiv:2012.11099, 2020.
[111] Cui L, Wu Y, Liu S, et al. Mutual: A dataset for multi-turn dialoguereasoning[J]. arXiv preprint arXiv:2004.04494, 2020.
[112] Wu C S, Hoi S, Socher R, et al. Tod-bert: pre-trained natural language understanding for task-oriented dialogue[J]. arXiv preprint arXiv:2004.06871, 2020.
[113] Xu R, Tao C, Jiang D, et al. Learning an effective context-responsematching model with self-supervised tasks for retrieval-based dialogues[J]. arXiv preprint arXiv:2009.06265, 2020.
[114] Wang W, Joty S, Hoi S C. Response selection for multi-partyconversations with dynamic topic tracking[J]. arXiv preprint arXiv:2010.07785, 2020.
[115] Fu Z, Cui S, Shang M, et al. Context-to-session matching: Utilizing whole session for response selection in information-seeking dialogue systems[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining.2020: 1605-1613.
[116] Fu Z, Cui S, Ji F, et al. Query-to-session matching: Do not forget history and future during response selection for multi-turn dialogue systems[C]//Proceedings of the 29th ACM InternationalConference on Information & Knowledge Management. 2020: 365-374.
[117] Whang T, Lee D, Oh D, et al. Do response selection models reallyknow what ’s next? utterance manipulation strategies for multiturn response selection[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 14041-14049.
[118] Zhang L, Ma D, Li S, et al. Do it once: An embarrassingly simplejoint matching approach to response selection[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 4872-4877.

[119] Lin Z, Cai D, Wang Y, et al. The world is not binary: Learning torank with grayscale data for dialogue response selection[J]. arXiv preprint arXiv:2004.02421, 2020.
[120] Ma W, Cui Y, Liu T, et al. Conversational word embedding for retrieval-based dialog system[J]. arXiv preprint arXiv:2004.13249, 2020.
[121] Gu J C, Ling Z H, Liu Q, et al. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots[J]. arXiv preprint arXiv:2004.14550, 2020.
[122] Shuster K, Humeau S, Bordes A, et al. Image chat: Engaging grounded conversations[J]. arXiv preprint arXiv:1811.00945, 2018.
[123] Qiu L, Shiu Y, Lin P, et al. What if bots feel moods?[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1161- 1170.
[124] Zandie R, Mahoor M H. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems[C]// The Thirty-Third International Flairs Conference. 2020.
[125] Yang L, Qiu M, Qu C, et al. Iart: Intent-aware response ranking with transformers in information-seeking conversation systems[C]//Proceedings of The Web Conference 2020. 2020: 2592- 2598.
[126] Zhong P, Zhang C, Wang H, et al. Towards persona-based empathetic conversational models[J]. arXiv preprint arXiv:2004.12316, 2020.
[127] Gu J C, Liu H, Ling Z H, et al. Partner matters! an empirical studyon fusing personas for personalized response selection in retrieval- based chatbots[J]. arXiv preprint arXiv:2105.09050, 2021.
[128] Hua K, Feng Z, Tao C, et al. Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 525-534.
[129] Sun Y, Hu Y, Xing L, et al. History-adaption knowledge incorporation mechanism for multi-turn dialogue system[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 8944-8951.
[130] Su Y, Cai D, Zhou Q, et al. Dialogue response selection with hierarchical curriculum learning[J]. arXiv preprint arXiv:2012.14756, 2020.
[131] Cleverdon C. The cranfield tests on index language devices[C]//Aslib proceedings. MCB UP Ltd, 1967.
[132] Nguyen T, Rosenberg M, Song X, et al. Ms marco: A humangenerated machine reading comprehension dataset[C]//CoCo@ NIPS. 2016.
[133] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions:a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.
[134] Ng A. Deep learning: What's next[C]//Proceedings of the 2016International Conference on Autonomous Agents & Multiagent Systems. 2016: 1-1.

[135] Rodriguez P, Boyd-Graber J. Evaluation paradigms in questionanswering[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9630-9642.

[136] Yang Z, Qi P, Zhang S, et al. Hotpotqa: A dataset for diverse,explainable multi-hop question answering[J]. arXiv preprint arXiv:1809.09600, 2018.

[137] Yu W, Jiang Z, Dong Y, et al. Reclor: A reading comprehension dataset requiring logical reasoning[J]. arXiv preprint arXiv:2002.04326, 2020.

[138] Dua D, Wang Y, Dasigi P, et al. Drop: A reading comprehensionbenchmark requiring discrete reasoning over paragraphs[J]. arXiv preprint arXiv:1903.00161, 2019.

[139] Bisk Y, Zellers R, Gao J, et al. Piqa: Reasoning about physicalcommonsense in natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 7432- 7439.

[140] Min S, Boyd-Graber J, Alberti C, et al. Neurips 2020 efficien-tqa competition: Systems, analyses and lessons learned[J]. arXiv preprint arXiv:2101.00133, 2021.

[141] Mao Y, He P, Liu X, et al. Generation-augmented retrieval for open-domain question answering[J]. arXiv preprint arXiv:2009.08553, 2020.

[142] Oguz B, Chen X, Karpukhin V, et al. Unified open-domain question answering with structured and unstructured knowledge[J]. arXiv preprint arXiv:2012.14610, 2020.

[143] Cheng H, Shen Y, Liu X, et al. Unitedqa: A hybrid approach for open domain question answering[J]. arXiv preprint arXiv:2101.00178, 2021.

[144] Liu L, Lewis P, Riedel S, et al. Challenges in generalization in open domain question answering[J]. arXiv preprint arXiv:2109.01156, 2021.

[145] Jiang Y, Bordia S, Zhong Z, et al. Hover: A dataset for many-hop fact extraction and claim verification[J]. arXiv preprint arXiv:2011.03088, 2020.

[146] Qi P, Lee H, Sido T, et al. Answering open-domain questionsof varying reasoning steps from text[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 3599-3614.

[147] Li S, Li X, Shang L, et al. Hopretriever: Retrieve hopsover wikipedia to answer complex questions[J]. arXiv preprint arXiv:2012.15534, 2020.

[148] Geva M, Khashabi D, Segal E, et al. Did aristotle use a laptop? aquestion answering benchmark with implicit reasoning strategies [J]. Transactions of the Association for Computational Linguistics, 2021, 9: 346-361.

[149] Liu J, Cui L, Liu H, et al. Logiqa: A challenge dataset for machinereading comprehension with logical reasoning[J]. arXiv preprint arXiv:2007.08124, 2020.

[150] Zhong W, Wang S, Tang D, et al. Ar-lsat: Investigating analyticalreasoning of text[J]. arXiv preprint arXiv:2104.06598, 2021.

[151] Clark P, Tafjord O, Richardson K. Transformers as soft reasonersover language[J]. arXiv preprint arXiv:2002.05867, 2020.

[152] Rocktäschel T, Riedel S. End-to-end differentiable proving[J].arXiv preprint arXiv:1705.11040, 2017.

[153] Amini A, Gabriel S, Lin P, et al. Mathqa: Towards interpretablemath word problem solving with operation-based formalisms[J]. arXiv preprint arXiv:1905.13319, 2019.

[154] Shao Z, Shang L, Liu Q, et al. A mutual information maximizationapproach for the spurious solution problem in weakly supervised question answering[J]. arXiv preprint arXiv:2106.07174, 2021.

[155] Chen K, Xu W, Cheng X, et al. Question directed graph attention network for numerical reasoning over text[J]. arXiv preprint arXiv:2009.07448, 2020.

[156] Al-Negheimish H, Madhyastha P, Russo A. Numerical reasoningin machine reading comprehension tasks: are we there yet?[J]. arXiv preprint arXiv:2109.08207, 2021.

[157] Ferguson J, Gardner M, Hajishirzi H, et al. Iirc: A dataset of incomplete information reading comprehension questions[J]. arXiv preprint arXiv:2011.07127, 2020.

[158] Talmor A, Herzig J, Lourie N, et al. Commonsenseqa: A questionanswering challenge targeting commonsense knowledge[J]. arXiv preprint arXiv:1811.00937, 2018.

[159] Sap M, Rashkin H, Chen D, et al. Socialiqa: Commonsense reasoning about social interactions[J]. arXiv preprint arXiv:1904.09728, 2019.

[160] Huang L, Bras R L, Bhagavatula C, et al. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning [J]. arXiv preprint arXiv:1909.00277, 2019.

[161] Speer R, Chin J, Havasi C C. 5.5: An open multilingual graphof general knowledge[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (December 2016). 4444- 4451.

[162] Sap M, Le Bras R, Allaway E, et al. Atomic: An atlas of machinecommonsense for if-then reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 3027- 3035.

[163] Lin B Y, Chen X, Chen J, et al. Kagnet: Knowledge-awaregraph networks for commonsense reasoning[J]. arXiv preprint arXiv:1909.02151, 2019.

[164] Bosselut A, Rashkin H, Sap M, et al. Comet: Commonsense transformers for automatic knowledge graph construction[J]. arXiv preprint arXiv:1906.05317, 2019.

[165] West P, Bhagavatula C, Hessel J, et al. Symbolic knowledge distillation: from general language models to commonsense models [J]. arXiv preprint arXiv:2110.07178, 2021.

[166] Tamborrino A, Pellicano N, Pannier B, et al. Pre-training is (al-most) all you need: An application to commonsense reasoning[J]. arXiv preprint arXiv:2004.14074, 2020.

[167] Niu Y, Huang F, Liang J, et al. A semantic-based method for un-supervised commonsense question answering[J]. arXiv preprint arXiv:2105.14781, 2021.

[168] Khashabi D, Min S, Khot T, et al. Unifiedqa: Crossing format boundaries with a single qa system[J]. arXiv preprint arXiv:2005.00700, 2020.

[169] Min S, Michael J, Hajishirzi H, et al. Ambigqa: Answering ambiguous open-domain questions[J]. arXiv preprint arXiv:2004.10645, 2020.

[170] Huang P S, He X, Gao J, et al. Learning deep structured semanticmodels for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.

[171] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences[C]//Advances in neural information processing systems. 2014: 2042-2050.

[172] Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering[C]//Twenty-Fourth international joint conference on artificial intelligence. 2015.

[173] Shen Y, He X, Gao J, et al. A latent semantic modelwith convolutional-pooling structure for information retrieval [C]//Proceedings of the 23rd ACM international conference onconference on information and knowledge management. 2014: 101-110.

[174] Palangi H, Deng L, Shen Y, et al. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 694-707.

[175] Wan S, Lan Y, Guo J, et al. A deep architecture for semanticmatching with multiple positional sentence representations[C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 30. 2016.

[176] Pang L, Lan Y, Guo J, et al. Text matching as image recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 30. 2016.

[177] Mitra B, Diaz F, Craswell N. Learning to match using local anddistributed representations of text for web search[C]//Proceedings of the 26th International Conference on World Wide Web. 2017: 1291-1299.

[178] Guo J, Fan Y, Ai Q, et al. A deep relevance matching model forad-hoc retrieval[C]//Proceedings of the 25th ACM international on conference on information and knowledge management. 2016: 55-64.

[179] Xiong C, Dai Z, Callan J, et al. End-to-end neural ad-hoc ranking with kernel pooling[C]//Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval. 2017: 55-64.

[180] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[181] Jiang J Y, Zhang M, Li C, et al. Semantic text matching for long-form documents[C]//The World Wide Web Conference. 2019: 795-806.

[182] Zhou X, Pappas N, Smith N A. Multilevel text alignment withcross-document attention[J]. arXiv preprint arXiv:2010.01263, 2020.

[183] Liu B, Niu D, Wei H, et al. Matching article pairs withgraphical decomposition and convolutions[J]. arXiv preprint arXiv:1802.07459, 2018.

[184] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

[185] Pang L, Lan Y, Cheng X. Matchignition: Plugging pagerankinto transformer for long-form text matching[J]. arXiv preprint arXiv:2101.06423, 2021.

[186] Page L, Brin S, Motwani R, et al. The pagerank citation ranking:Bringing order to the web.[R]. Stanford InfoLab, 1999.

[187] Ge S, Dou Z, Jiang Z, et al. Personalizing search results using hierarchical rnn with query-aware attention[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 347-356.

[188] Ma Z, Dou Z, Bian G, et al. Pstie: Time information enhanced personalized search[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 1075-1084.

[189] Lu S, Dou Z, Jun X, et al. Psgan: A minimax game for personalized search with limited and noisy click data[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 555-564.

[190] Zhou Y, Dou Z, Wen J R. Enhancing refinding behavior withexternal memories for personalized search[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 789-797.

[191] Yao J, Dou Z, Xu J, et al. Rlper: A reinforcement learning modelfor personalized search[C]//Proceedings of The Web Conference 2020. 2020: 2298-2308.

[192] Lu S, Dou Z, Xiong C, et al. Knowledge enhanced personalized search[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 709-718.

[193] Zhou Y, Dou Z, Wen J R. Encoding history with context-awarerepresentation learning for personalized search[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1111-1120.

[194] Yao J, Dou Z, Wen J R. Employing personal word embeddingsfor personalized search[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1359-1368.

[195] Yao J, Dou Z, Xie R, et al. User: A unified information search and recommendation model based on integrated behavior sequence [C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 2373-2382.

[196] Uprety S, Gkoumas D, Song D. A survey of quantum theory inspired approaches to information retrieval[J]. ACM Computing Surveys (CSUR), 2020, 53(5): 1-39.

[197] Bruza P D, Wang Z, Busemeyer J R. Quantum cognition: a newtheoretical approach to psychology[J]. Trends in cognitive sciences, 2015, 19(7): 383-393.

[198] Van Rijsbergen C J. The geometry of information retrieval[M].Cambridge University Press, 2004.

[199] Sordoni A, Nie J Y, Bengio Y. Modeling term dependencies withquantum language models for ir[C]//Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 653-662.

[200] Zhang P, Su Z, Zhang L, et al. A quantum many-body wave function inspired language modeling approach[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 1303-1312.

[201] Zhang P, Niu J, Su Z, et al. End-to-end quantum-like languagemodels with application to question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.

[202] Zhang L, Zhang P, Ma X, et al. A generalized language model intensor space[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 7450-7458.

[203] Zuccon G, Azzopardi L A, Van Rijsbergen K. The quantum probability ranking principle for information retrieval[C]//Conference on the Theory of Information Retrieval. Springer, 2009: 232-240.

[204] Zhao X, Zhang P, Song D, et al. A novel re-ranking approachinspired by quantum measurement[C]//European Conference on Information Retrieval. Springer, 2011: 721-724.

[205] Jiang Y, Zhang P, Gao H, et al. A quantum interference inspiredneural matching model for ad-hoc retrieval[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 19-28.

[206] Zhang P, Li J, Wang B, et al. A quantum query expansion approachfor session search[J]. Entropy, 2016, 18(4): 146.

[207] Wang B, Zhang P, Li J, et al. Exploration of quantum interferencein document relevance judgement discrepancy[J]. Entropy, 2016, 18(4): 144.

[208] Li X, Liu Y, Mao J, et al. Understanding reading attention distribution during relevance judgement[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 733-742.

[209] Li X, Mao J, Wang C, et al. Teach machine how to read: reading behavior inspired relevance estimation[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 795-804.

[210] Moshfeghi Y, Triantafillou P, Pollick F E. Understanding information need: An fmri study[C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016: 335-344.

[211] Moshfeghi Y, Pollick F E. Search process as transitions betweenneural states[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 1683-1692.

[212] Moshfeghi Y, Triantafillou P, Pollick F. Towards predicting a realisation of an information need based on brain signals[C]//The World Wide Web Conference. 2019: 1300-1309.

[213] Pinkosova Z, McGeown W J, Moshfeghi Y. The cortical activity ofgraded relevance[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 299-308.

[214] Carterette B. System effectiveness, user models, and user utility: a conceptual framework for investigation[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval. 2011: 903-912.

[215] Moffat A, Thomas P, Scholer F. Users versus models: What observation tells us about effectiveness metrics[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 659-668.

[216] Moffat A, Bailey P, Scholer F, et al. Incorporating user expectations and behavior into the measurement of search effectiveness [J]. ACM Transactions on Information Systems (TOIS), 2017, 35 (3): 1-38.

[217] Azzopardi L, Zuccon G. An analysis of the cost and benefitof search interactions[C]//Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. 2016: 59-68.

[218] Zhang F, Liu Y, Li X, et al. Evaluating web search with a bejeweled player model[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 425-434.

[219] Zhang F, Mao J, Liu Y, et al. Cascade or recency: Constructingbetter evaluation metrics for session search[C]//Proceedings ofthe 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 389-398.

[220] Liu M, Liu Y, Mao J, et al. Towards designing better sessionsearch evaluation metrics[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1121-1124.

[221] Wicaksono A F, Moffat A. Metrics, user models, and satisfaction[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 654-662.

[222] Zhang F, Mao J, Liu Y, et al. Models versus satisfaction: Towardsa better understanding of evaluation metrics[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 379-388.

[223] Sordoni A, Bengio Y, Vahabi H, et al. A hierarchical recurrentencoder-decoder for generative context-aware query suggestion[C]//proceedings of the 24th ACM international on conference on information and knowledge management. 2015: 553-562.

[224] Ahmad W U, Chang K W, Wang H. Multi-task learning for document ranking and query suggestion[C]//International Conference on Learning Representations. 2018.

[225] Jiang J Y, Wang W. Rin: Reformulation inference networkfor context-aware query suggestion[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 197-206.

[226] Qu C, Xiong C, Zhang Y, et al. Contextual re-ranking with behavior aware transformers[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1589-1592.

[227] Chen J, Mao J, Liu Y, et al. A hybrid framework for session contextmodeling[J]. ACM Transactions on Information Systems (TOIS), 2021, 39(3): 1-35.

[228] Borisov A, Markov I, De Rijke M, et al. A neural click model forweb search[C]//Proceedings of the 25th International Conference on World Wide Web. 2016: 531-541.

[229] Zhang J, Mao J, Liu Y, et al. Context-aware ranking by constructing a virtual environment for reinforcement learning[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1603-1612.

[230] Borisov A, Wardenaar M, Markov I, et al. A click sequence modelfor web search[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 45-54.

[231] Chen J, Mao J, Liu Y, et al. A context-aware click model for websearch[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 88-96.

[232] Mao J, Luo C, Zhang M, et al. Constructing click models formobile search[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 775- 784.

[233] Zheng Y, Mao J, Liu Y, et al. Constructing click model for mobilesearch with viewport time[J]. ACM Transactions on Information Systems (TOIS), 2019, 37(4): 1-34.

[234] Xie X, Mao J, de Rijke M, et al. Constructing an interaction behavior model for web image search[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 425-434.

[235] Xie X, Mao J, Liu Y, et al. Grid-based evaluation metrics for webimage search[C]//The World Wide Web Conference. 2019: 2103- 2114.

[236] Wicaksono A F, Moffat A. Modeling search and session effectiveness[J]. Information Processing & Management, 2021, 58(4): 102601.

[237] Zhang Y, Liu X, Zhai C. Information retrieval evaluation as searchsimulation: A general formal framework for ir evaluation[C]// Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. 2017: 193-200.

[238] Maxwell D, Azzopardi L. Agents, simulated users and humans:An analysis of performance and behaviour[C]//Proceedings of the 25th ACM international on conference on information and knowledge management. 2016: 731-740.

[239] Zhai C. Interactive information retrieval: Models, algorithms,and evaluation[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2444-2447.

[240] Ferrante M, Ferro N. Exploiting stopping time to evaluate accumulated relevance[C]//Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. 2020: 169-176.

[241] Liu Y, Mao J. ” revisiting information retrieval tasks with user behavior models” by yiqun liu and jiaxin mao with martin veselyas coordinator[J]. ACM SIGWEB Newsletter, 2020(Autumn): 1- 8.

[242] Li X, Liu Y, Mao J. Understanding the role of human-inspired heuristics for retrieval models[J]. Frontiers of Computer Science, 2022, 16(1): 1-11.

[243] Dai Z, Xiong C, Callan J, et al. Convolutional neural networksfor soft-matching n-grams in ad-hoc search[C]//Proceedings ofthe eleventh ACM international conference on web search and data mining. 2018: 126-134.

[244] Zheng Y, Fan Z, Liu Y, et al. Sogou-qcl: a new dataset with clickrelevance label[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1117-1120.

[245] Joachims T, Swaminathan A, Schnabel T. Unbiased learning-to-rank with biased feedback[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017: 781-789.

[246] Wang X, Golbandi N, Bendersky M, et al. Position bias estimationfor unbiased learning to rank in personal search[C]//Proceedings ofthe Eleventh ACM International Conference on Web Search and Data Mining. 2018: 610-618.

[247] Ai Q, Yang T, Wang H, et al. Unbiased learning to rank: Online oroffline?[J]. ACM Transactions on Information Systems (TOIS), 2021, 39(2): 1-29.

[248] Oosterhuis H, de Rijke M. Differentiable unbiased online learningto rank[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 1293- 1302.

[249] Wang H, Kim S, McCord-Snook E, et al. Variance reduction ingradient exploration for online learning to rank[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 835-844.

[250] Jia Y, Wang H, Guo S, et al. Pairrank: Online pairwise learningto rank by divide-and-conquer[C]//Proceedings of the Web Conference 2021. 2021: 146-157.

[251] Marchionini G. Information Seeking in Electronic Environments[M/OL]. Cambridge University Press, 1995. DOI: 10.1017/cbo9 780511626388.

[252] Liu C, Song X, Liu H, et al. Modeling Knowledge Change Behaviors in Learning-related Tasks[C/OL]//CEUR Workshop Proceedings: volume 2699. Galway, Ireland, 2020. http://ceur-ws.or g/Vol-2699/paper18.pdf.

[253] Bates M J. Information search tactics[J/OL]. Journal of theAmerican Society for Information Science, 1979, 30(4): 205-214. https://onlinelibrary.wiley.com/doi/10.1002/asi.4630300406.

[254] Koskela M, Luukkonen P, Ruotsalo T, et al. Proactive Information Retrieval by Capturing Search Intent from Primary Task Context[J/OL]. ACM Transactions on Interactive Intelligent Systems, 2018, 8(3): 1-25. DOI: 10.1145/3150975.

[255] Chen J, Mao J, Liu Y, et al. Towards a Better Understanding of Query Reformulation Behavior in Web Search[C/OL]// Proceedings of the Web Conference 2021. New York, NY, USA: ACM, 2021: 743-755. https://dl.acm.org/doi/10.1145/3442381.3 450127.

[256] Mitsui M, Liu J, Shah C. How much is too much?: Whole session vs. first query behaviors in task type prediction[C]//The 41st International ACM SIGIR Conference on Research Development in Information Retrieval. ACM, 2018: 1141-1144.

[257] White R W. Interactions with search systems[M]. CambridgeUniversity Press, New York, 2016.

[258] González-Ibáñez R, Esparza-Villamán A, Vargas-Godoy J C, et al.A comparison ofunimodal and multimodal models for implicit detection of relevance in interactive IR[J/OL]. Journal of the Association for Information Science and Technology, 2019, 0(0): 1-13. DOI: 10.1002/asi.24202.

[259] Liu Z, Mao J, Wang C, et al. Enhancing click models withmouse movement information[J/OL]. Information Retrieval Journal, 2017, 20(1): 53-80. http://link.springer.com/10.1007/s10791 -016-9292-4.

[260] Vakkari P, Völske M, Potthast M, et al. Modeling the usefulness ofsearch results as measured by information use[J/OL]. Information Processing Management, 2019, 56(3): 879-894. DOI: 10.1016/j. ipm.2019.02.001.

[261] Liu J, Liu C, Belkin N J. Personalization in text information retrieval: A survey[J/OL]. Journal of the Association for Information Science and Technology, 2019: asi.24234. DOI: 10.1002/as i.24234.

[262] Eugster M J, Ruotsalo T, Spapé M M, et al. Predicting term-relevance from brain signals[C/OL]//SIGIR ’14: Proceedings of the 37th International ACM SIGIR Conference on

Research andDevelopment in Information Retrieval. New York, NY, USA: ACM, 2014: 425-434. DOI: 10.1145/2600428.2609594.

[263] Moshfeghi Y, Pollick F, Triantafillou P, et al. Towards Predicting a Realisation of an Information Need Based on Brain Signals [C/OL]//WWW '19: The World Wide Web Conference. New York, NY, USA: ACM, 2019: 1300-1309. DOI: 10.1145/3308 558.3313671.

[264] Wu Y, Liu Y, Tsai Y R, et al. Investigating the role of eye movements and physiological signals in search satisfaction prediction using geometric analysis[J/OL]. Journal ofthe Association for Information Science and Technology, 2019, 70(9): 981-999. https://onlinelibrary.wiley.com/doi/10.1002/asi.24240.

[265] Kelly D. Methods for evaluating interactive information retrievalsystems with users[J]. Foundations and trends in Information Retrieval, 2009, 3(1-2): 1-224.

[266] Liu J, Han F. Investigating Reference Dependence Effects onUser Search Interaction and Satisfaction[C/OL]//Proceedings of the 43rd International ACM SIGIR Conference on Research andDevelopment in Information Retrieval. New York, NY, USA: ACM, 2020: 1141-1150. https://dl.acm.org/doi/10.1145/33972 71.3401085.

[267] Maddalena E, Mizzaro S, Scholer F, et al. On crowdsourcing relevance magnitudes for information retrieval evaluation[J]. ACM Transactions on Information Systems (TOIS), 2017, 35(3): 1-32.

[268] Roitero K, Maddalena E, Demartini G, et al. On fine-grained relevance scales[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 675- 684.

[269] Chu Z, Mao J, Zhang F, et al. Evaluating relevance judgments withpairwise discriminative power[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 261-270.

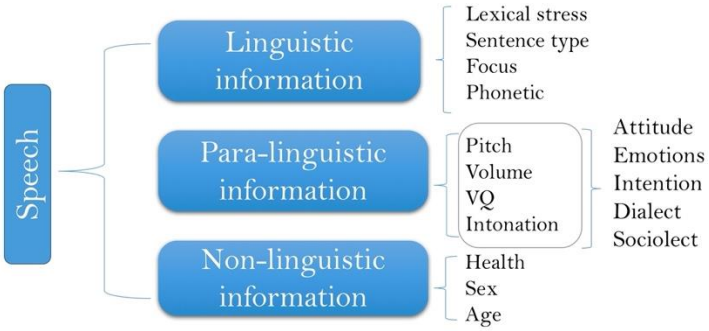
[270] Azzopardi L, Thomas P, Craswell N. Measuring the utility of search engine result pages: an information foraging based measure[C]//The 41st International ACM SIGIR conference on research & development in information retrieval. 2018: 605-614.

[271] Saracevic T. Evaluation of evaluation in information retrieval[C]//Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. 1995: 138-146.

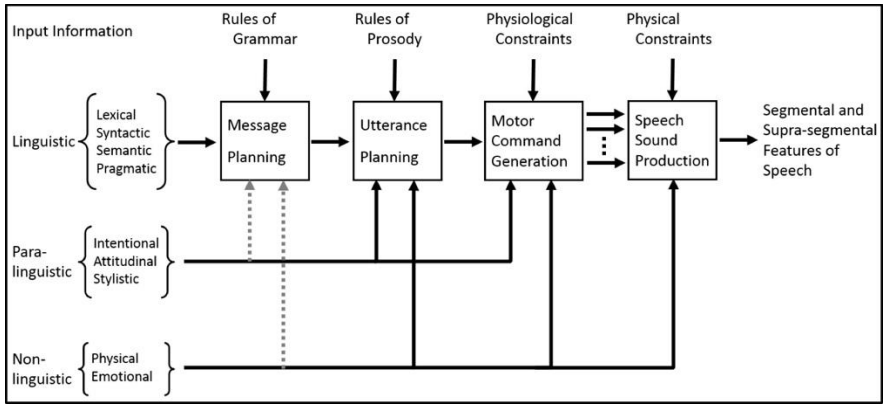
第七章　语音信号技术研究进展、现状及趋势

7.1.研究背景与意义

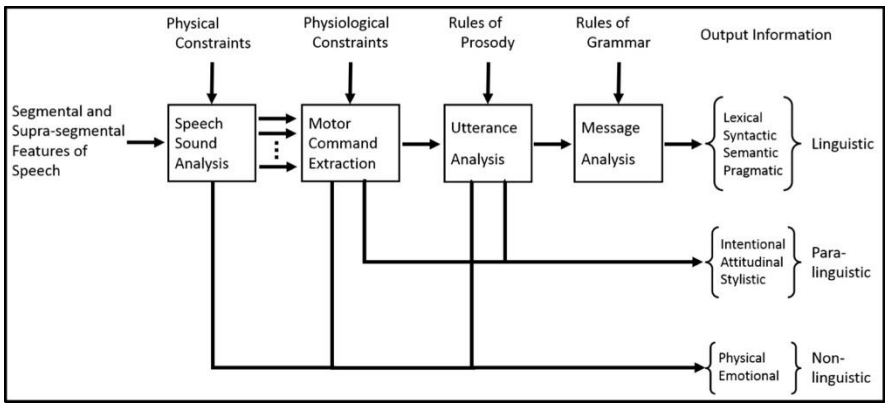
语音信号形简意丰，“形简”指形式简单为一维信号，“意丰”指信息丰富，包含很多信息。语音信号处理（Speech Signal Processing）是一门多学科的综合技术，主要包括语音识别、声纹识别、语音情感识别、语音会话理解、语音合成和转换等。从信号处理的角度看，语音信号的解码问题可以统称为“识别问题”。语音识别的目的是解码出语音信号中的发音内容信息，即要表达的句子。这些句子中，既包括语法和语义的约束，也包含大量的知识、经验、文化、习俗等背景。发音人通过发音器官的动作，将句子单元顺序编码在声门产生的载波信号中，形成形式简单但表义丰富的语音信号。特别重要的是，这一编码过程不仅编码了要表达的句子，同时编码了重音、语调、身体状况、背景环境等各种信息，因此语音信号具有极高的复杂度。语音识别是这一编码过程的逆过程，即将句子单元从语音信号中顺序提取出来。基于发音过程的随机性以及语义本身的复杂性，语音识别需要处理多种来源的不确定性，并引入各个层次的知识进行约束，以完成从一个高熵信号中提取出目标语句的任务。而语音信号的编码问题，可以统称为“合成问题”。语音合成又称文语转换，旨在实现将输入文本转换为流畅自然的输出语音，是实现智能人机语音交互的关键技术。当前的主流语音合成技术使用基于神经网络的序列到序列模型，已经可以端到端地生成较高自然度的语音。



语音信号中所含信息的分层表示



语音的信号的形成（产生），对应合成过程



语音信号的分解（分析），对应识别过程

本报告主要围绕语音信号的“识别”和“合成”两大维度，从研究背景与意义、领域发展现状与关键科学问题，领域关键技术进展及趋势，领域产业发展现状及趋势、总结及展望进行详细综述。

7.1.1. 语音识别

会话(Conversation)是人类最基本的信息交互方式。从传统的面对面自然会话发展到目前互联网环境下通过键盘输入形成的网络会话(Internet Relay Conversation, IRC)，人们的会话交流方式发生了翻天覆地的变化，期间积累的各种类型的会话数据也在成倍增长。根据记录设备的不同，这些会话数据可以包括语音、文本，甚至图像、视频等多模态信息。其中不同会话者与话语消息之间多方(multi-party)和多轮(multi-turn)复杂多通道交互流程，与以单模态单通道的文本数据为基础的独白(monolog)语篇描述结构有着本质的区别。如何从这些复杂的多模态多通道会话数据中挖掘出有用的知识，对会话理解技术研发提出了新的挑战。

7.1.2. 声纹识别

随着互联网的高速发展，远程身份认证成为各种网络应用的必备需求。传统的身份认证方式，如密码，安全问题，U 盾等，存在容易遗忘、容易丢失的问题。近年来，随着生物识别技术的发展，基于生物识别的认证技术逐渐成为新型远程身份认证方式而进入大众的生活，而其中声纹识别（也即说话人识别）是一种非常特殊的生物特征。

与人脸、指纹等不同，语音信号具有交互性、便捷性、变化性和丰富性的特点。交互性是指语音是可双向传递，既可接收信息，又可发出信息。便捷性是指语音信号可在空间中 360 度全方位无死角传播，使用时无需接触，无需对准摄像头，使用方便。变化性是指语音信号的高可变性与唯一性的完美统一。说话人每次说的话均存在变化，而同一个人的语音对应的人的特性又总是保持不变。丰富性是指语音具有“形简意丰”的特点，一维信号内同时蕴含着内容、情感、身份、年龄等丰富的信息，可同时执行语音识别、声纹识别、意图理解等多项任务。

承载与语音信号之中的声纹信息是一种具有生理特性的行为特征。说其具有生理特性，是因为先天发音器官（如舌头、牙齿、口腔、声带、肺、鼻腔等）差异将直接体现在语音当中；说其具有行为特性，是因为语音包含了后天发音与言语习惯的特殊征象。因此，可以说，任何两个人的声纹都不尽相同

与其他生物特征识别技术如指纹识别、掌纹识别、虹膜识别等一样，说话人识别有不会遗忘、无须记忆的优点。与此同时，说话人识别所用的采集设备成本很低，对麦克风和手机、电话录音等都没有特殊的要求，用户使用时也不用刻意接触采集设备，用户的接受程度普遍较高。因此，声纹识别是一种具有独特优势的身份认证方式，可在多个领域发挥作用。

7.1.3. 语音情感识别

语音是人类交流思想、情感、态度和认知状态最常见、最自然的方式，而语音情感识别(Speech Emotion Recognition)是通过对语音进行分析,以预测说话人的情绪状态。尤其是在人机交互中，情感信息能够帮助机器正确理解人的真实意图，从而为人类提供舒适的交互体验,其中特别是在安全医疗健康、娱乐和教育领域。例如，基于 ARM (Advanced RISC Machine)的自动柜员机，通过语音情感识别开发的沟通渠道或平台，帮助自闭症谱系障碍儿童有效沟通和游戏平台，通过在虚拟世界中玩来学习情感表达开发了一种智能家庭生活支持机器人系统相关场景越来越受研究者的关注，新型的人机交互场景逐渐的成为研究的主流热点；综上所知语音情感识别的研究对于增强计算机的智

能化和人性化,开发新型人机环境,以及推动心理学等学科的发展,有着重要的现实意义。

7.1.4. 语音会话理解

会话(Conversation)是人类最基本的信息交互方式。从传统的面对面自然会话发展到目前互联网环境下通过键盘输入形成的网络会话(Internet Relay Conversation, IRC)，人们的会话交流方式发生了翻天覆地的变化，期间积累的各种类型的会话数据也在成倍增长。根据记录设备的不同，这些会话数据可以包括语音、文本，甚至图像、视频等多模态信息。其中不同会话者与话语消息之间多方(multi-party)和多轮(multi-turn)复杂多通道交互流程，与以单模态单通道的文本数据为基础的独白(monolog)语篇描述结构有着本质的区别。如何从这些复杂的多模态多通道会话数据中挖掘出有用的知识，对会话理解技术研发提出了新的挑战。

7.1.5. 语音合成（含转换）

语音合成（Speech Synthesis）通常指文语转换（Text-to-Speech，TTS），其目的在于将文本转换为流畅自然的语音，是实现智能人机语音交互的关键技术之一。语音合成一般包括文本分析与语音信号合成两部分，其中前者主要负责对输入的文本进行自然语言理解和分析得到文本特征，而后者基于文本特征预测得到声学特征并进而生成语音信号。

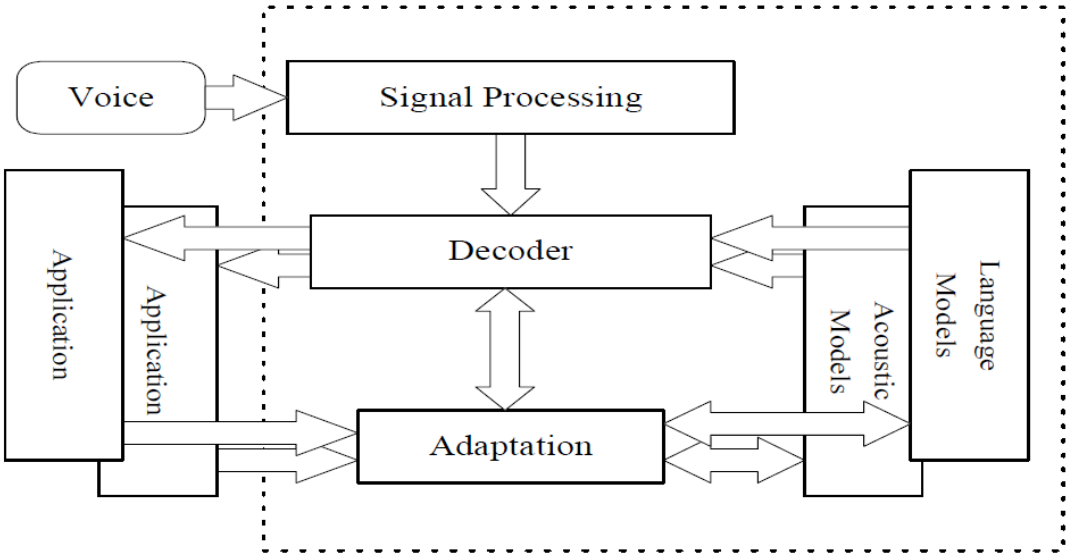
语音转换（Voice Conversion）则是一种将源说话人的语音转换成目标说话人的语音而不改变其说话内容的技术，其在实现个性化语音交互方面有着重要应用。一方面，语音转换系统可以串联在语音合成系统之后以构建个性化语音合成系统；另一方面，语音转换系统可以直接用于人声转换，例如在线语言教学中将用户发音转换成个性化的声音。语音转换对语音信号表征学习及加深人类对语音信号的理性理解也有重要意义。

近年来，以谷歌、苹果、微软、亚马逊为首的国际互联网巨头积极推进语音交互技术的研发；在国内，以百度、腾讯、讯飞、阿里巴巴、搜狗、京东为代表的互联网企业也纷纷布局智能语音市场并发布了众多语音交互产品。和谐自然的语音合成和语音转换技术是实现智能人机语音交互的关键。

7.2.领域发展现状与关键科学问题

7.2.1. 语音识别

语音识别研究主要包括如下三方面内容：语音信号的表示，即特征抽取；语音信号和语言知识建模；基于模型的推理，即解码。语音信号的复杂性和多变性使得这三方面的研究都面临相当大的挑战。下图给出一个语音识别系统的典型架构。



语音识别系统结构 (Huang, X., 1996)

7.2.1.1. 语音特征抽取

语音识别的一个主要困难在于语音信号的复杂性和多变性。一段看似简单的语音信号，其中包含了说话人、发音内容、信道特征、口音方言等大量信息。不仅如此，这些底层信息互相组合在一起，又表达了如情绪变化、语法语义、暗示内涵等丰富的高层信息。如此众多的信息中，仅有少量是和语音识别相关的，这些信息被淹没在大量其它信息中，因此充满了变动性。语音特征抽取即是在原始语音信号中提取出与语音识别最相关的信息，滤除其它无关信息。

语音特征抽取的原则是：尽量保留对发音内容的区分性，同时提高对其它信息变量的鲁棒性。历史上研究者通过各种物理学、生理学、心理学等模型构造出各种精巧的语音特征抽取方法，近年来的研究倾向于通过数据驱动学习适合某一应用场景的语音特征。特别是深度神经网络(DNN)的兴起后，基于大数据的特征学习方法取得了长足进展。

7.2.1.2. 模型构建

语音识别中的建模包括声学建模和语言建模。声学建模是对声音信号（语音特征）的特性进行抽象化。自上世纪 70 年代中期以来，声学模型基本上以统计模型为主，特别是隐马尔可夫模型/高斯混合模型(HMM/GMM)结构[1]。最近几年，深度神经网络成为声学模型的主流结构[2]。

声学模型需要解决两个基本问题：（1）如何对语音信号的短时平稳性及长时序列性时行建模;（2）如何对模型进行优化，即模型训练。同时，在实际应用中，还需要解决一系列具体问题，如：（1）如何从一个领域快速自适应到另一个领域；（2）如何处理噪音、信道等产生的干扰；（3）如何处理混合发音，解决鸡尾酒会问题；（4）如何解决方言、口音上的变异；（5）如何对低资源语言进行建模；（6）如何利用大量无标注数据提高建模质量；（7）如何利用多模态，特别是视频信息进行建模。

语言建模是对语言中的词语搭配关系进行归纳，抽象成概率模型。这一模型在解码过程中对搜索空间形成约束，不仅减小计算量，而且可以提高解码精度。传统语言模型多基于 N 元文法 (n-gram)[3]，近年来基于递归神经网络（RNN）的语言模型发展很快[4]，在某些识别任务中取得了比 n-gram 模型更好的结果。

语言模型要解决的主要问题是如何对低频词进行平滑。不论是 n-gram 模型还是 RNN 模型，低频词很难积累足够的统计量，因而无法得到较好的概率估计。平滑方法借用高频词或相似词的统计量，提高对低频词概率估计的准确性。除此之外，语言建模研究还包括：（1）如何对字母、字、词、短语、主题等多层次语言单元进行层次性建模；（2）如何对应用领域进行快速自适应；（3）如何提高训练效率；（4）如何有效利用大量带噪声文本，等等。

7.2.1.3. 解码

解码是利用语音模型和语言模型中积累的知识，基于输入语音信号，推理出相应发音内容的过程。早期的解码器一般为动态解码，即在开始解码前，将各种知识源以独立模块形式加载到内存中，动态构造解码图。现代语音识别系统多采用静态解码，将各种知识源统一表达成有限状态转移机（FST），并将各层次的 FST 嵌套组合在一起，形成解码图[5]。解码时，一般采用 Viterbi 算法在解码图中进行路径搜索。为加快搜索速度，一般要对搜索路径进行剪枝，去除概率较低的路径，以加快解码效率。

对解码器的研究包括但不限于如下内容：（1）如何加快解码速度，特别是在应用神经网络语言模型进行一遍解码时；（2）如何实现静态解码图的动态更新，如加入新词；

（2）如何有效利用高层语义信息； （3） 如何估计解码结果的信任度；（4）如何实现多语言和混合语言解码；（5）如何对多个解码器的解码结果进行融合。

7.2.2. 声纹识别

目前，声纹识别受到广大研究者的关注，组织了许多与声纹识别相关的竞赛与测评来推进声纹识别的发展。随着声纹识别技术的发展，此类竞赛或测频也从最开始的仅关注声纹识别系统的性能，开始引入跨信道、跨语言、短语音等复杂情况下的测试。例如美国国家标准与技术研究院从 1996 年开始， 每年定期举办说话人识别评测 NIST SRE[23][24]。VoxSRC 竞赛基于 VoxCeleb[25]数据库举办，后者为英国牛津大学在 2017 年开源的全球明星声纹数据库。该项竞赛中直接考验声纹识别技术在噪音复杂、环境不受限等场景下的性能表现。

目前，声纹识别存在以下关键科学问题：

1. 如何对说话人的本质差异进行表征与建模

声纹识别的前提，需要回答说话人之间的本质差异到底为何。说话人之间的本质差异可分为两个部分，一为生理差异，即由于发音器官的构造、形状的差异而导致的语音声色上的差异；二为行为差异，即由于说话人后天学习的不同导致的说话习惯、方式、口音等方面的差异。此外，由于语音信号中存在多种信息，如内容、语调等。声纹只是其中的一种信息。因此，如何在变化的语音信号中，准确的对说话人的本质差异进行表征和建模，成为说话人识别中的一个关键问题。

2. 如何构建提高声纹识别算法在实际复杂应用环境下的鲁棒性

声纹识别算法在实际应用过程会遇到多种问题，如噪声问题、跨信道问题，时变问题，短语音问题等。噪声问题，是指在实际应用场景中，说话人的语音中往往包含各种各样的噪声，如白噪声、背景噪声等。这些噪声在一定程度上淹没了语音信号中所含有的说话人特征信息，减少了说话人模型的分辨特性。跨信道问题，是指在实际应用中，语音信号可以从不同的终端通过各式各样的录音设备采集得到，例如不同的手机、麦克风、录音笔等等。录音设备的不同会直接导致语音信号在频谱上发生畸变，从而严重影响语音的声学特征和说话人模型对说话人个性的表征能力。时变问题是指，说话人随着年龄变化而导致的声音发生的变化。此类变化将导致早期构建的模型性能降低甚至失效。短语音问题时指，在较短甚至超短语音条件下的说话人识别，这个问题的解决将直接改善实际应用中的用户体验。因此，如何在此类复杂情况下，仍能鲁棒的进行识别，成为声纹识别中的一个关键问题。

3. 如何防止声纹识别系统受到蓄意攻击

若要使声纹识别系统进行广泛的应用，必须要提高声纹识别系统防假冒闯入攻击的能力。目前，声纹识别系统假冒攻击的方式主要有声音模仿、语音合成、声音转换、录音重放、对抗样本等五种类型。研究表明，声音模仿对声纹识别系统的影响不大，但语音合成、声音转换、录音重放以及对抗样本能以较高的成功率欺骗传统声纹识别系统，显著提升系统闯入率，对声纹识别的应用产生巨大的威胁。因此，如何防止声纹识别系统受到攻击，成为一个关键问题。

7.2.3. 语音情感识别

7.2.3.1. 发展现状

随着语音情感识别在模式识别领域的持续火热，过去常用的方法是采用人们通过提取语音中携带情感倾向的声学线索，设计了 LLDs 特征(low-level descriptors) 和统计特征(functions) ，并采用不同的机器学习识别算法来构建语音情感识别系统。随着深度神经网络的兴起，现主流的方案聚焦于采用深度神经网络（CNN、RNN）等自动提取情感有关的特征。并在此基础上，国内专家学者针对不同的问题展开了大量研究。

在语音情感数据集的扩展方面，国内诸多等院校针对情感数据覆盖不全面，规模小等问题相继推出了不同语言下的大型语音情感数据集 EMOVIE、LSSED，为提高现有算法的泛化性拟合应用于实际场景提供了数据基础；在情感表征能力提升方面，研究工作[65]提出了基于自注意的视频流与所提出的音频流相融合来实现视听信息融合。改进了情感相关表示向量的自适应策略(AG-FBP)、自适应和多级 FBP (AM-FBP)，动态计算两种模式的融合权值，并实现了全局表征和中间表征的结合。研究工作[66]提出了一种有监督的 NMF 模型 DSNMF，以提高 SER 的能力。DSNMF 作为一种特征学习模型，通过结合识别信息和相似度约束，获得数据的低维表示，增强了特征识别，取得了较好的性能。针对现有数据集规模不足问题，迁移学习被广泛的应用，[67]提出了迁移子空间学习算法(FSTSL)来解决特征匹配和特征选择问题，巧妙地将转移子空间学习和特征选择结合到一个统一的框架中学习鲁棒的低维语料库不变特征表示。

基于神经网络的端到端模型也逐渐成为当前的主流技术，针对语音情感的特点许多神经网络架构被提出，研究工作[68]通过堆叠多个 transformer 层，增强了全局特征表征效果，在常用的 IEMOCAP 数据集上获得了目前最高的 92%的效果。[69]利用多级跨模态情绪蒸馏(Multi-level Cross-modal Emotion extraction, MCED)方法，通过从预先训练的文本情绪模型中转移情绪知识来训练没有标记的语音情绪数据的语音情绪模型，

缓解了语音情感数据集标注成本高和标注歧义问题。面对跨语料库，嘈杂环境等复杂多样的语音情感场景下，工作[70]联合分布自适应回归(JDAR)方法联合考虑训练语音信号和测试语音信号之间的边际概率分布和条件概率分布来学习回归矩阵，从而在学习的回归矩阵所跨的子空间中缓解它们的特征分布差异。[71][72]引入域对抗训练，实现数据集无关表示，并采用预训练自编码器克服了训练数据不足的问题。

从目前主流的研究进展中可以发现，近年来的深度学习方法的准确率在逐年提高。在单模态语音情感识别方法中，[68]在常用的 IEMOCAP 数据集上达到了目前最高的92.00%的未加权准确率。[73],提出的基于 triplet 损失和数据增强的方法实现了 78.30%的准确率,[69]采用多尺度区域注意力机制和数据增强方案,实现了 77.54% 的准确率，在多模态语音情感识别方法中，[74]获得了充分利用了模态内和模态间的互信息，得到了 83.80%的准确率，[75]研究了多模态融合方法中信息冗余和信息互补的重要性,取得了 80.83%的准确率。

尽管各科研院所针对不同的问题提出了相应的机器学习方法，在性能表现上有明显的提醒，但是针对语音情感识别的几个关键性问题仍然存在。

基于真实场景下的情感数据资源不足。情感数据采集困难导致当下用于研究的数据规模较小，目前常用的数据集多通过人为表演和音视频录制或者启发式情感激发方法获得，缺乏真实场景下对外部刺激做出情绪样本。另外，在不同的场景下，人们的情感表达的强烈程度不同，不同场景情感样本类别可能会出现强烈的不平衡现象。这都阻碍了语音情感识别在真实场景下的应用。

情感数据标注困难。情感作为人类主观心理状态对外部刺激做的有规律的反映。在对情感数据标注时，标注者受自身情绪、经历，性格，年龄等诸多心理和生理因素影响，由此造成的，数据标注不一致，不可靠问题严重阻碍了语音情感识别的研究进展。

缺乏相对同一、有效的情感表征。尽管针对不同的问题提出了不同的解决方案，但是，情感作为世界上人类共有的心理和生理反映，其存在于语音信号中的形式可能存在相似性。探索情感在语音信号中的存在形式仍然发展缓慢，众多方案仅是通过不同的机器学习算法来拟合有区分性的特征，而非情感的真实存在形式。

7.2.3.2. 关键科学问题

1. 语音情感在机器学习中的形式化描述。情感本质上是人类对外部刺激做出的心理和生理反映，人脑复杂的神经系统能够较为容易的感知语音中的情感，将这种感知外化成准确地描述符则存在相当大的争议。情感作为复杂的心理和生理特征，其外部表达是各种因素复合的结果，基于离散类别的描述符（如高兴，愤怒等）无法表达情感的复

杂性。而基于连续的语音属性特征描述符，虽然从多个维度表达了情感的，建立语音信号情感本质特征与计算机形式化表达之间的映射关系。

2. 语音情感信息的表征。采用不同的方法抽取语音情感表征信息，探索情感表征的方法，如声学特征、统计特征以及深度神经网络自动抽取的特征。

3. 情感表征基础上的有效区分情感差异性的模型。如探究传统机器学习算法对表征的分类和识别效果，以及深度学习算法对于情感表征自动抽取的能力的研究

7.2.4. 语音会话理解

近年来，随着“预训练+微调”范式的提出及其在独白语言理解任务上取得的巨大进展，许多研究人员开始尝试将类似的处理方法移植到会话理解任务上。但最初的实验结果不是很理想，主要原因在于最初针对独白语言理解问题设计的深度学习和建模机制不能很好适应会话数据的描述特点。例如：会话中的多方和多轮交互描述特点，会形成复杂的会话者(speaker)和话语消息(utterance)交互结构，对其中不同话语语义内容和交际功能的有效识别提出了新的挑战。会话过程中的话题延续机制非常灵活，话题转换情况非常频繁，并由于交互方式的不同会形成复杂的话题纠缠(entanglement)现象，对其中话题结构的准确识别带来了极大困难。会话中的话语消息内容比较简洁，存在大量基于语境的内容省略情况，针对口语语音的转录文本中可能还会有停顿、重复、改错等特殊口语描述现象，需要设计特殊的语言建模结构对其进行特殊处理。形成准确流畅的会话过程，需要各个会话参与者随时把握对手方的内在心理状态，预判可能的情感变化趋势，分析隐含的交际意图，为此需要探索许多新的针对会话理解的新建模机制。

针对以上问题，近期研究人员主要在以下几个方面进行了深入探索：1） 会话表示学习问题，研究将会话结构描述特点有效融入现有的“预训练+微调”范式的新方法，形成具有更强描述能力的话语嵌入计算机制。2） 会话理解新任务探索，针对某类特殊的会话描述结构，设计并构建新的标注数据资源，探索新的机器学习方法，不断提升该项任务的处理性能。

7.2.5. 语音合成及转换

语音合成框架通常包含文本分析和语音信号合成两个模块[100]。文本分析模块包括对输入文本进行正则化、分词和词性分析、字音转换及多音字消歧、韵律结构分析等，最终生成语言学内容文本特征序列。语音信号合成模块包括时长预测、声学参数预测、

声码器合成等。其中时长预测模块预测每个语言学单元的发音时长，根据所预测的时长对语言学内容序列进行转换，使之与目标声学特征参数序列具有相同的长度；声学参数预测的作用是将对齐的语言学内容序列映射到声学特征参数序列；声码器合成则是将声学特征参数序列恢复成语音波形信号。

传统的拼接式合成（Concatenative Speech Synthesis）、基于隐马尔可夫模型（Hidden Markov Model，HMM）的参数化合成等方法[101]，通过对语音合成中的多个关键模块分别建模实现了较好的语音合成效果，但仍不尽如人意。近年来，随着深度学习技术的发展，基于深度神经网络（Deep Neural Network，DNN）和深度置信网络（Deep Belief Network，DBN）的强大建模能力[102][103][104]，语音合成的音质和自然度得到了显著的提升。而循环神经网络（Recurrent Neural Network，RNN）及其各种变体的应用[105][106][107]，更好地建模了当前语言学单元受上下文信息的影响，实现了自然度更高的语音合成效果。然而多个子模块的分别建模，需要根据复杂的专家知识且依赖于相应的数据标注，各个模块之间的误差累积也会使得整个系统的性能下降。

为解决上述不足，基于序列到序列建模（sequence-to-sequence）的语音合成框架直接对文本到语音的序列映射进行建模，实现了高音质、高自然度的语音合成。该框架通常包括编码器模块（Encoder）、解码器模块（Decoder）和注意力机制（Attention Mechanism）[108]。模型接收文本序列输入，通过编码器模块得到文本序列的编码表示，接着通过注意力机制为解码器的每个时间步的解码提供所需的文本编码的上下文信息，再通过解码器模块预测每个时间步的声学特征参数，最后通过相应的声码器（Vocoder）由声学特征参数生成语音波形。神经网络声码器（Neural Vocoder）[109][110]的提出则极大地提升了语音波形信号恢复的准确度，进一步提升了语音合成的性能。

语音转换有多种方式：根据源说话人或目标说话人的数量，可分为：一对一语音转换、多对一语音转换、多对多语音转换；根据是否需要平行语料库可分为：平行数据语音转换和非平行数据语音转换。得益于平行语料库所提供的训练数据上的优势，基于频谱映射的平行数据语音转换被大量应用，例如，基于样本稀疏表示的频谱映射方法[123]、基于高斯混合模型（Gaussian Mixture Model，GMM）的频谱映射方法[124][125]、基于双向长短时记忆网络（Bidirectional Long Short-Term Memory，BLSTM）的神经网络频谱映射方法[126][127]等。然而，在实际应用场景中平行语料较难获取，因此，基于识别-转换范式的非平行数据语音转换模型被广泛采用，例如，基于因素后验概率（Phonetic Posterior-Grams，PPGs）的语音转换[128]、基于量化变分自编码器（Vector Quantized Variational AutoEncoder，VQ-VAE）的语音转换[129][130]等。另外，在目标说话人语音数据不足的情况下，如何通过所给定的少量语音数据提取目标说话人的音色表征以及通过该音色表征控制语音转换模型的转换目标也是研究的热点问题，具体包

括基于说话人编码器（Speaker Encoder）的方法，其通过额外训练的说话人编码器学习目标说话人的音色表征[131][132][133]；以及基于模型适应性训练（Model Adaptation）的方法，其通过对平均音色语音转换模型进行适应性训练使模型学习到目标说话人的音色特性[134][135]。

语音信号中不仅包含了语言语义信息，同时也传达了说话人、语种、风格、情感、重音、语气等多种丰富的副语言信息。这些副语言信息承载了语音的言外之意，是语音表现力的重要方面，在传递信息时扮演着非常重要的作用，是和谐语音交互不可或缺的因素。而现有语音交互产品中的语音生成技术，在生成语音时缺乏对上述副语言信息表达的有效控制、不能充分利用和体现多种副语言信息的表现，导致合成语音大多局限于某种特定风格、难以满足人们和谐自然语音交互的表现力多样性需求。因此，如何进行个性化、表现力语音合成，实现多种副语言信息的解耦与可控生成是目前国内外语音合成领域研究的前沿热点问题。

7.3.领域关键技术进展及趋势

7.3.1. 语音识别

语音识别研究可追溯到 20 世纪 50 年代，例如贝尔实验室的 AUDREY 系统，用模拟电路实现了对 10 个数字的识别。从那以后，语音识别技术经历了模式识别、统计模型、深度学习等几个发展阶段。需要注意的是，语音识别技术包括特征提取、声学建模、语言建模、解码等几个方面，其中声学建模的发展最为显著。上述发展阶段基本上是以声学模型的发展而划分的。因而，本节主要关注声学模型技术 （特征提取在深度学习方法中成为声学模型的一部分），同时简述其它几种技术的发展现状。

7.3.1.1. 概率模型方法

语音识别技术发展初期以模式匹配方法为主，对不同词保留若干注册样本，将待测试语音信号与这些标准样本进行匹配，取距离最近的样本所对应的词作为该语音信号的发音。上个世纪 80 年代，Reddy、Jelinek、Baker 等研究者提出基于概率模型来描述这些不确定的发音。这一模型主要包括两个部分：在描述时序动态性上，认为一个发音单元（一般为音素）包括若干状态，同一状态内部的发音特性保持相对稳定，不同状态间以一定的概率进行跳转；在描述发音特征的不确定性上，通过概率模型描述某一发音状态内部的特征分布。应用最广泛的概率模型是 HMM/GMM 模型（如图 2 所示），其中 HMM 用来描述短时平稳的动态性，GMM 用来描述 HMM 每一状态内部发音特征的概率分布。

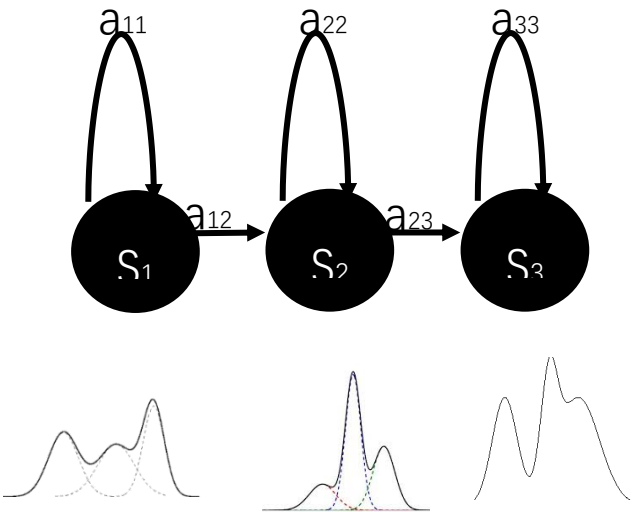


图 2 . HMM/GMM 模型

HMM/GMM 模型结构简单，有保证收敛的快速训练方法，可扩展性强，因此一直到 2011 年一直是语音识别领域的主流方法。基于 HMM/GMM 框架，研究者提出各种改进方法，如引入上下文相关性的动态贝叶斯方法、区分性训练方法、自适应训练方法、HMM/NN 混合模型方法等。这些方法都对语音识别研究产生了深远影响，并为下一代语音识别技术的发展做好了准备。

7.3.1.2. 深度学习方法

深度学习是“使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法”。深度学习在语音识别领域中的应用始于 2009 年，首先在 TIMIT 数据集上获得了成功[6]。之后，微软、IBM、谷歌等公司对深度学习模型进行了深入探索，尝试了各种深度学习模型在不同识别任务上的效果[7]。今天，深度学习技术已经成为语音识别中的主流方法，基于深度模型的语音识别系统不论是识别率还是鲁棒性都远好于基于 HMM/GMM 的系统。

（1） DNN/HMM 混合模型方法

2013 年以前，基于前馈网络的 DNN 是语音识别中应用最广泛的深度模型[2]。这一模型包括多个隐藏层，具有强大的特征学习能力。经过合理的初始化（如预训练），DNN 可通过随机梯度下降(SGD)算法进行优化。训练完成以后，DNN 即可用于替代 GMM 来计算语音特征在不同 HMM 状态下的概率，并基于传统 FST 框架进行解码。2013 年以后，研究者探索了各种神经网络模型，其中具有重要意义的是卷积神经网络（CNN）[8]和循环神经网络（RNN）[9]。

CNN 是一种比 DNN 更有效的特征提取模型，它利用语音信号中典型模式（如音素）的重复性和局部性，将 DNN 的全连接结构变成时频空间中的局部连接结构，相当于设计了一系列具有局部关注特性的滤波器，并通过训练学习滤波器的参数。这一思路将 DNN 模型的特征学习方式进一步结构化，不仅减小了参数量，得到的 CNN 模型也更加符合特征提取的结构化要求。

RNN 模型是一种状态累积的时序动态模型。通过时序建模，RNN 可以学习更长时的历史信息，进而提高模型的预测和分类能力。近年来，研究者探索出一系列更适合语音建模的 RNN 结构，如 LSTM， GRU，双向 LSTM 等。同时，人们发现将多层 RNN 迭加起来形成深层 RNN 结构，可进一步提高识别性能。

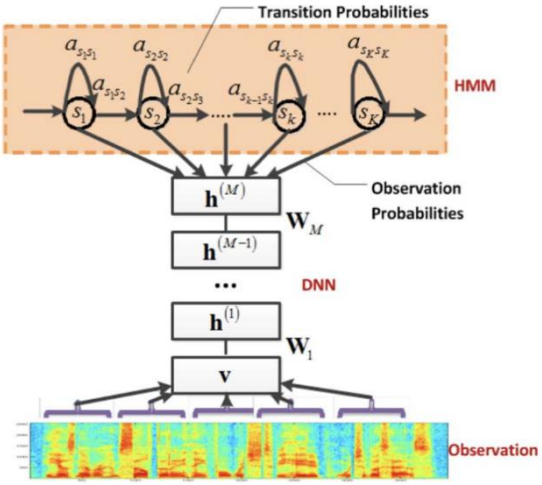


图 3 . DNN/HMM 混合模型

（2） 端到端模型

DNN/HMM 混合模型的依然依赖 HMM 模型，这一模型的潜在问题在于用离散的状态来描述动态的语音生成过程。2014 年以后，人们尝试去掉 HMM 模型，首先取得成功的是 CTC 方法[10]。这一方法对整个音素进行建模，因此不需要 HMM 模型。同时，CTC 在训练中考虑音素和语音信号之间的各种可能的对齐关系，因此不需要一个音素-语音对齐过程。总体来看，不论是训练还是解码，这一模型关注的是由一个输入语音序列到音素序列的转化过程，前者是输入端，后者是目标端。这种由输入端直接映射到目标端的模型称为“端到端模型”。CTC 模型是第一个获得成功的端到端语音模型，意味着统治语音识别研究近 40 年的 HMM 模型至少已经变成一个可选项。由传统模型相比，端到端模型避免了不同模块独立学习产生的次优化问题，同时也显著减轻了解码器的设计工作。端到端模型之所以成功，很大程度上归功于神经网络强大的学习能力，包括对特征的抽像学习能力和对时序相关性的学习能力。

CTC 之后，人们发现这一模型缺少对输出序列的建模能力，不同时刻的输出概率条件独立。为克服这一困难，研究者提出 RNN-T 模型[11]，在 CTC 框架基础上，引入相邻输出单元之间的概率模型，从而具有了更好的建模能力。事实上，这一输出单元之间的概率模型定义了一个音素串上的语言模型，这意味着如果训练数据足够大，RNN-T 可以同时学习语音模型和语言模型，因此解码时不再需要语言模型。因此，相比传统 CTC，RNN-T 模型是更完整的端到端模型。

2015 年，研究者提出另一种端到端语音识别框架，称为序列到序列（Sequence2Sequence）模型[12]。这一模型采用一种更自然的方式将语音信号映射到目标音素序列：首先设计一个基于 RNN 的编码器，将输入信号序列降采样成一个代表发音内容的抽象序列，再通过一个基于 RNN 的解码器，将这一抽象序列映射成目标音素串。这一模型具有如下特点：首无，通过注意力机制，在每一步解码时可以参考整条输入语音；其次，基于 RNN 的编码器可以对语音信号的长时信息进行建模；第三，基于 RNN 的解码器可以描述音素之间的相关性，事实上构建了音素上的语言模型。和 RNN-T 类似，如果训练数据足够充分，序列到序列模型将同时学习声学模型和语言模型，因此是一个完整的端到端模型。

7.3.1.3. 语言模型和解码器

在语言模型方面，传统的 n-gram 模型加上各种平滑算法在大词表语音识别中依然占有重要地位。近年来，基于深度神经网络的语言模型 (NNLM) 取得很大进展[4]，但 NNLM 在大词表任务中依然有很多优化工作要做。近年来，端到端模型取得很大进展，这一模型中，语言模型和声学模型深度绑定，因此不需要额外的语言模型。特别是当训练数据足够时，端到端模型可以独立用于语音识别任务。然而，当训练数据不足时，模型内嵌的语言模型将不足以描述语言生成过程；当出现领域失配时，内嵌的语言模型将产生领域偏差。这时，就需要融合外部语言模型才能得到较好的识别性能。通常的融合方式包括浅层融合、深层融合、冷融合等。

解码器方面，基于 FST 的静态解码方法依然适用，特别是当识别系统依赖大规模 n-gram 语言模型时， FST 解码效率更高。对于端到端系统，模型的输出多是字或词的片段，因此解码时只需简单的剪枝即可，不仅可以简化解码操作，还可以处理新词。

7.3.2. 声纹识别

7.3.2.1. 定义与目标

说话人识别（Speaker Recognition），又名声纹识别（Voiceprint Recognition, VPR），是根据语音中所蕴含的说话人个性信息，来对说话人身份进行自动鉴别的生物特征识别技术。与语音识别关注内容信息所不同，声纹识别关注的是说话人的个性化信息。

根据具体目标的不同，说话人识别可分为说话人辨认（Speaker Identification）、说话人确认（Speaker Verification）、说话人检测（Speaker Detection）、说话人追踪（Speaker Tracking）等几个类别。其中说话人辨认是一个“多选一”的选择问题，其需要从若干个参考说话人中，选择出当前语音的实际说话人。说话人确认则是一个“一对一”的判断问题，其需要判断两段语音是否属于同一个说话人。此外，根据语音文本内容的不同，说话人识别任务还可以分为文本相关、文本无关、文本提示等几类。

7.3.2.2. 进展与影响

说话人识别的研究最初着力于说话人特征设计，而后在说话人建模进行广泛研究。近年来，随着深度学习的发展，深度网络也应用于说话人识别中，开始出现端到端的说话人识别系统。

7.3.2.3. 说话人特征设计

为了进行说话人识别，需要对语音中的说话人个性信息进行表征。对此，研究者从语音短时频谱的静态特性、时序动态特性、听觉感知特性、声源声道特性入手，提出了众多声纹特征，目前应用最广的特征参数包括线性预测倒谱系数[26]（Linear predictive cepstrum coefficient, LPCC）、Mel 频率倒谱系数[27]（Mel-Frequency cepstrum coefficient, MFCC）和感知线性预测[28]（Perceptual linear predictive, PLP）等。

7.3.2.4. 话人建模方法

研究之初，说话人识别采用模板匹配方式进行，即从训练语句中提取特征，作为当前说话人的说话人参考模板；在测试阶段，从测试语音中提取出相同特征，并利用动态时间弯折[29]（Dynamic time warping, DTW）技术其与参考模板比对，并根据两者的

匹配程度来判断两句话是否为相同说话人。而这种方式只能做内容相同时的比对，且严重依赖特征提取，往往鲁棒性较差。2000 年，Reynolds 等人[30]提出高斯混合模型-通用背景模型（Gaussian mixture model-Universal background model, GMM-UBM）的说话人识别框架，这大大提升了说话人识别的性能，摆脱内容的限制，且提高了系统的鲁棒性。然而这种方式对于跨信道的鲁棒性仍然较差。而后，Kenny 等人[31]提出的联合因子分析（JFA）方法，其在跨信道的说话人识别领域取得了很大的成功。Dehak 等人[32]在此基础上提出了 i-vector 模型，保留了更多的说话人信息。为了提高区分性，多种后处理方法和模型被提升，如类内协方差归一化[33]（Within-class covariance normalization, WCCN），干扰属性投影[34]（Nuisance attribute projection, NAP）、概率线性区分性分析[35]（Probabilistic linear discriminant analysis, PLDA）等。

近年来，随着深度学习在语音识别等语音信号处理领域的快速发展和成功应用，深度学习方法也逐渐应用到声纹识别中，并取得了不俗的效果。一类方法尝试利用深度网络的学习能力，直接从原始频谱特征中基于数据驱动的方式学习说话人的表征，使得学习到的特征具有更强的任务相关性。根据网络结构、学习目标、学习方式的不同，研究者分别提出了 d-vector[36]、x-vector[37]等多种说话人表征。另一类方法尝试进行端到端识别建模，其尝试抛弃后端模型，直接由网络判断两段语音的相似度。Heigold 等人[38]针对说话人确认任务，设计了逻辑回归或者三元组损失（Triplet loss）等目标函数，实现了端到端的建模与打分。

7.3.2.5. 鲁棒性研究

为了提升说话人识别的噪声鲁棒性，研究者根据干扰因素的不同分门别类针对性展开研究，主要包含环境鲁棒性、说话人鲁棒性和应用鲁棒性三类。

对环境鲁棒性而言，研究通常聚焦在环境噪音、信道失配、多说话人等非说话人自身因素引起的问题，并在信号域、模型域和分数域上提出了前端降噪[39]、数据增强[40]、模型自适应[41][42]、话者分离[43]、分数规整[44]等相关解决方法。

对说话人鲁棒性而言，其通常是指由说话人自身的一些因素对声纹识别性能带来的影响，主要包括健康状况、年龄变化、情感波动、语速变化和语言变化等引起的问题。针对不同影响因素，研究者提出了相应的解决方法。例如，针对年龄变化，Wang 等人[45]提出了基于 F-ratio 准则的频带区分性特征提取算法；针对情感波动，Bie 等人[46]提出了基于 fMLLR 的情感特征空间补偿方法；针对语言变化，Lu 等人[47]提出了基于 JFA 的语言因子补偿算法。

对应用鲁棒性而言，其通常是指声纹识别技术在实际应用中所面临的问题，例如，

如何检测声音模仿、语音合成、声音转换、录音重放等假冒攻击手段[48]；如何缩短注册和验证语音时长，提升声纹识别的体验性[49]；如何理解被认证者的真实意图，实现认证的实人实意[50]。

7.3.2.6. 攻击研究

声纹识别系统容易遭受到多种攻击，其中合成音攻击（Logical Access）和录音重放攻击（Physical Access）是目前威胁最大的两类。对此，研究者从合成音攻击和录音重放攻击两个方面展开对应的防御措施研究。

在合成音检测方面，对于传统语音合成方法，研究者从其静态特性和动态特性两个方面进行攻击检测。在静态特性方面，Chen 等[51]在梅尔倒谱域进行观察，发现合成音的声道响应,在倒谱的高阶部分和真实语音具有显著差异。在动态特性方面,Leon 等[52]对语音激励信号的动态特性进行研究，发现合成语音的基频可能发生跳变，或变的过于平滑。Satoh 等[53]分析声纹识别系统对语音的每一帧的似然分，发现合成音的一阶似然分插值动态范围更小，从另一个角度说明了合成音具有更加平滑的动态特性。而对于基于深度神经网络的语音合成和声音转换技术[54]，研究者也采用了多种深度网络架构[55]，利用深度网络的强大建模能力，对合成音进行检测。

在录音重放检测方面，研究者根据录音与真实的差异性，从多个角度入手进行检测。从语音信号随机性的角度，研究者提出将验证语音与用户历史语音进行相似性匹配[56][57]；当发现存在相似性过高的历史语音时，则认为该语音有被录音重放的风险。从信号失真的角度，研究者认为语音信号经过物理设备的录放后，引入了不同程度的信号失真。为此，从频谱幅值[58]、相位[59]、调幅信号[60]、源激励信号[61]等多个角度入手，实现对录放信号的检测。此外，通过机器和人的差异性，研究者也尝试借助额外信息，如多普勒雷达探测[62]、声场检测[63]、喷麦检测[64]等方法，来区分录音重放和真实语音。

7.3.2.7. 发展

未来，说话人识别研究仍然需要攻克以下难题：

1. 提高说话人识别系统在复杂现实应用情况下的鲁棒性。

尽管当前在鲁棒性方面进行了一系列研究，而说话人识别系统在实际应用中仍然面临较大鲁棒性问题。如何进一步降低所依赖的语音长度、提高跨信道鲁棒性、是未来面临的主要难题。

2. 提升说话人识别系统防攻击能力

尽管当前已有一些防攻击策略，但攻击手段是不断更新的。语音合成算法的进步将导致合成语音的质量更高，更加逼真，检出将更加困难。而如何对高保真的录音重放进行检出，也是当前未解决的一大难题。此外，对于一些新的攻击方式，如对抗样本、海豚音攻击、激光照射攻击等，如何对其进行防范，也是一个难题。

3. 促进多模态信息融合

说话人识别作为单一身份认证来源，很难满足一些高安全性要求的应用场景。此时，多生物特征融合是一种更加有效的方式。此外，在面对复杂干扰的情况下，如鸡尾酒会效应、强噪声干扰等情况下，若能引入其他模态的信息，也将大大的提升模型识别的效果。因此，如何将说话人信息与其他信息相融合，这是未来的研究方向。

7.3.3. 语音情感识别

语音情感数据获取困难始终是制约语音情感识别发展的关键问题。长期以来，国内外各个科研院所所提出的先进的方法主要在 IEMOCAP、EMO-DB、RECOLAR、CASIA 等较为知名的数据集上进行研究，这些数据集由于规模、情感覆盖范围等因素一定程度上制约了新方法的创新和现有方法的改进。近年来科研人员发布了一个新的自然状态视频数据库 HEU Emotion[76]包含共包含 19004 个视频片段，9951 名参与者，以及上面提到的 LSSED（包含了 147025 个句子，总共 206 小时 25 分钟，由 820 名参与者构成）和 EMOVIE（包含 9724 个样本和音频文件的中文情感语音数据集及其情感标注）。针对现有汉语音位级情感数据库存在的差距，建立了音位级汉语视听情感数据库，该语料库将语音分为音素层次，涵盖了所有汉语音素，这是第一个音素级的汉语视听情感语料库，包括 35 人录制的 2480 个音频/视频和 74 类音素中的 115000 多个片段。语料库规模显著提高，真实世界场景下的情感表达更加丰富，为语音情感识别进一步发展提供了数据基础。

在考虑全局情感信息建模问题时，为解决对话语音情感识别任务，研究工作[77]提出了一个时频胶囊神经网络(TFCap)来建模全局信息，可以直接从光谱图中提取更稳定的全局时频信息。在话语间阶段，引入图卷积网络(GCN)来研究对话中话语间的关系。在研究工作[78]中,采用无监督方式对无监督 VQ-VAE 进行预训练，提取标记数据的潜在表示，并提出了一种 TACN 模型，用于对序列信息进行建模。研究工作[79]提出了一种新的对话情感检测模块，主要方式是设计了从源波形和合成波形中捕获情感特征的波形-注意模块，利用效能系数机制进行细粒度多模态信息融合。

对抗学习的域自适应方法成为近年来处理数据域不匹配问题的热门方法，研究工作

[80]针对训练(源)样本与测试(目标)样本分布不一致的问题，提出了一种基于非负矩阵分解的迁移子空间学习方法。该方法试图为源语料库和目标语料库找到一个共享的特征子空间，在子空间中尽可能消除源语料库和目标语料库之间分布的差异，排除它们各自的域相关成分，从而将源语料库的知识转移到目标语料库中。

在多模态情感识别研究中，工作[81]提出了一种新的基于权重共享的深度多模态变压器网络，该网络学习分离的时间特征并处理模式间的异步情感特征，探究了音频和文本特征所包含的情感具有内在的相互依赖性。研究工作[82]提出了一种多级跨模态情绪蒸馏(Multi-level Cross-modal Emotion extraction, MCED)方法，该方法通过从预先训练的文本情绪模型中转移情绪知识来训练没有标记的语音情绪数据的语音情绪模型。针对情感信息在语音中的持续时间较短的问题，多实例学习通过将语音分成多个片段，捕捉最显著的时刻来呈现情感信息，[83]提出了一种基于对抗性自动编码器的多实例学习分类器，该模型中的鉴别器可以将潜在表示映射到高维高斯分布，还采用了性别和情绪分类的多任务学习策略。

在对外部噪声环境对语音情感识别影响的研究中，[84]发现发现噪声对 SER 的影响受其 Mel 谱图的影响。噪声 Mel 谱图平均值的绝对值与识别精度呈正相关。噪声显著影响识别效果，某些类型的噪声具有与人类语音相似的 Mel 谱图，对 SER 的准确性影响不明显。

7.3.4. 语音会话理解

7.3.4.1. 会话表示学习

在会话表示学习方面,[85]通过融入多方会话 MPC 中类似“谁对谁说了什么”的会话者与话语消息之间的复杂结构关系构建了一个新的针对 MPC 的预训练模型 MPC-BERT。[95]通过设计两个新的语言模型训练目标：话语顺序重构和句子骨干规范化构建了一个融合更多会话特征的预训练模型 SPIDER。[89]构建了 DialoFlow 模型来动态建模会话历史与当前话语之间的复杂内容联系。[92]通过自动切分并抽取话题相关话语消息来构建话题相关的多轮会话模型 TADAM。[90]设计了一种基于话语消息的对比学习机制 DialogueCSE，可以有效获取不同话语消息的嵌入表示。

7.3.4.2. 会话理解新任务探索

在会话理解新任务探索方面，[99]设计了一套针对汉语日常会话的对话行为 DA 标注体系，并据此标注完成了包含 500 个汉语日常会话片段的 DA 标注库。在此基础上，

[97]初步验证了多任务学习架构对汉语会话片段的对话行为标记自动识别任务的有效性。[96]探索了从字幕流会话数据中自动切分出具有不同话题内容的会话片段的可行性。[98]针对会话角色识别任务，提出了一种基于多尺度自注意力增强的新方法。[86]设计了一种新的会话角色识别 SPD 任务，可以将若干特殊定制的会话者角色描述片段自动映射到个性角色库中的合适单元上。[93]结合心智理论思想来自动预测会话对手的个性化状态，以便实现更好的协商交易目标。[88]设计了一个心理知识感知的交互图模型，综合利用当前话语的过往行为和预期意图及效果来准确识别当前话语的情感标记。[87]提出了一个受话者感知模型来自动预判当前话语情感的保持和感染效果，从而可以为自动推断后续可能话语的情感标记提供更多支撑。[91]提出了一个会话结构感知图模型来融合会话者、话语谓词等关键信息以实现会话语义角色标注任务。[94] 提出了一个主题词引导的对话图注意网络来充分挖掘会话片段中各个话语消息之间的内在关联性以实现会话片段自动摘要任务。。

7.3.5. 语音合成（含转换）

7.3.5.1. 个性化语音生成

在个性化语音生成方面，基于序列建模的语音克隆和语音转换技术近年来受到广泛的关注。语音克隆（Voice Cloning）是指利用少量目标说话人语音数据，构建具有目标说话人音色特点的语音合成系统；而语音转换（Voice Conversion）则旨在保证输入的源说话人语音的文本内容不变的前提下，将语音的音色转换成目标说话人的音色。研究人员通过在模型中引入说话人嵌入向量（Speaker Embedding）表征，利用众多说话人的语音数据训练一个多说话人的语音合成模型[111]，并进而提出说话人自适应与说话人编码两种不同的语音克隆方法[112][113]。前者利用少量新的目标发音人的训练数据微调（Fine-tune）多说话人语音合成模型的参数，得到目标说话人的模型。后者通过构建说话人编码器（Speaker Encoder）从目标说话人训练数据中估计说话人嵌入向量，并与多说话人语音合成模型进行联合训练。针对语音转换，研究人员提出了基于说话人信息和文本内容信息解耦的语音转换方法[114]，其包含文本编码器（Text Encoder）、语音内容编码器（Recognition Encoder）、说话人编码器（Speaker Encoder）等多个编码器模块和解码器模块（Decoder），并利用多说话人数据进行编码器解码器的联合训练，在模型训练完成后分别提取源说话人语音的内容表征和目标说话人语音的说话人表征，进而输入到解码器中进行声学特征的预测。

7.3.5.2. 表现力语音合成

在表现力语音合成方面，有研究人员提出基于 Tacotron 框架的端到端情感语音合成方法[115]，通过预处理网络将中性、生气、害怕、高兴、悲伤、惊讶等 6 种情感标签进行编码，并将其引入到解码器（Decoder）网络中，实现指定情感的语音合成。在很多情况下，语音的情感韵律特性无法使用明确的表现力标签进行标注。为此，研究人员提出了一种从自然语音中提取韵律嵌入向量（Prosody Embedding）并通过韵律迁移实现表现力语音合成的方法[116]。在此基础上，研究者进一步提出了基于全局风格符号（Global Style Token, GST）[117]的风格化语音合成方法，其将韵律嵌入向量分解成固定个数的风格符号集合，实现了无监督的风格控制和迁移。

7.3.5.3. 多因素解耦与可控语音合成

在多因素解耦与可控语音合成方面，研究工作[118]通过说话人嵌入（Speaker Embedding）和重音信息嵌入（Emphasis Embedding）显式地引入说话人音色和重音的控制信息，能够实现将一个说话人的重音特性转移到另一个说话人上，同时通过对编码器输出的调制实现对合成语音重音强度的控制。在跨语言语音合成[119]中，通过说话人嵌入和语种嵌入（Language Embedding）显式地对说话人信息和语种进行编码控制，可实现语言特性的迁移（利用单一语言的数据集实现跨语言语音合成），同时实现口音（中式英语口音或英式中文口音）强度的可控生成。研究工作[120]通过变分自编码器（Variational Auto-Encoder，VAE）进行除说话人、语种之外的其他副语言信息的表征建模，并引入说话人对抗学习（Adversarial Learning）消除语言学内容中的说话人相关信息，通过对说话人信息和语种信息进行显示建模以控制合成语音的相关副语言表现。研究工作[121]基于 Tacotron 采用变分自编码器 VAE 的方法进行说话风格信息的解耦，实现了对生成语音的部分特性（基频）的可控性，表明 VAE 的隐变量能够实现对信息的解耦，并且能够直接对隐变量进行修改以控制生成语音的说话风格。而研究工作[122]提出了基于层级生成模型的可控语音合成系统，采用层级 VAE 进行语音信息的解耦合，能够实现较好的说话人、录音环境噪音及语音韵律的控制。

7.4.领域产业发展现状及趋势

7.4.1. 语音识别

7.4.1.1. 复杂网络结构

为了进一步提高对语音信号和语言现象的表达能力，研究者探索了一系列复杂网络结构。其中，Transformer 网络具有代表性。

Transformer 是近年来提出的一种新型深度神经网络模型。这一模型基于自注意力机制(Self Attention)，每一层特征对前一层特征做进一步总结与抽象。和传统前馈 DNN 模型相比，Transformer 具有时序建模能力；和传统 RNN 相比，Transformer 具有更宽的感受野，更适合长时序列建模，特别是非顺序的、跨段的长时建模。

2018 年，自动化所[13]、KIT/CMU[14]等研究组将 Transformer 结构应用于语音识别，在 WSJ 数据集取得了不错的效果。2020 年，Google 的研究者将 RNN-T 模型的 RNN 结构替换成 Transformer，得到了明显的性能提升[15]。进一步，Google 的研究者提出 Conformer 结构[16]，将 CNN 和 Self-Attention 进行结合，充分利用二者在局部建模和全局建模上的独特优势，得到了很好的效果。

2019 年，Google 研究者将 Transformer 应用于字符级语言模型 （Character Level LM）[17]。进一步，研究者提出 Transformer-XL 结构[18]，引入段级别递归机制和相对位置编码，从而可以学习更长时的信息。同年，研究者将 Transformer 语言模型应用于语音识别，对识别结果进行二次解码，或作为辅助信息提高端到端系统的识别性能[19]。目前，Transformer 结构作为语言模型在语音识别上的应用还在探索中。

7.4.1.2. 预训练模型

互联网上存在大量未标注数据，如果对这些数据进行合理应用，将有效提高声学模型的性能。近年来，基于自学习（Self Learning）的无监督学习方法得到广泛关注。所谓自学习，是指利用数据本身带有的天然属性作为学习信号，从而实现不需要人为标注的学习方法。例如，在一段语音信号中，较近距离内的发音内容相近，而较远距离的发音内容不同；再例如，一段语音信号加入噪声干扰后，发音内容不会改变。利用这些信息，研究者设计了语音预训练模型。Wave2Vec 是目前具有代表性的预训练模型[20]。这一模型的目的是对语音信号进行无监督的特征提取，使其在语音识

别等下游任务中带来性能提升。Wave2Vec 设计了对比损失函数（Contrastive Loss），基本思路是通过神经网络提取抽像特征，并使该特征符合自学习的基础假设，如相邻帧的特征更接近。最近，研究者提出 Wave2Vec 2.0 预训练模型[21]，将语音片段映射成离散的码字，不同码字对应不同的发音，从而可以自动学习出音素集。

基于这些预训练模型作为特征提取器，有望利用较少的数据构造更好的声学模型。多个研究组的研究结果显示，结合大规模复杂网络（如 Transformer 和 Conformer），预训练模型可显著提高语音识别的性能。最近，研究人员将 Wave2Vec2.0 和 BERT 模型结合，以增强特征的上下文表达能力，在 LibriSpeech 数据集上获得了目前最优的效果[22]。

7.4.1.3. 开源代码和数据

近年来语音识别技术的进展，很大程度上要归功于开源代码的普及和免费数据的大量涌现。

在开源代码方面，卡内基-梅隆大学的 Sphinx 系统和剑桥大学的 HTK 系统是早期开源系统的杰出代表，极大推动了 HMM 语音识别技术的进展。2012 年以来，由 Dan Povey 开创的 Kaldi 成为语音识别领域的标志性开源系统。和 HTK 相比，Kaldi 具有明显优势：鼓励全世界的研究人员加入到开发行列，因而更有活力；引入样例代码，和开源数据相配合，初学者可以根据样例完成一个标准的语音识别系统，极大降低了技术的准入门槛；允许商用，推动了技术社区和商业社区的互动，形成了良性循环。在 Kaldi 的引领下，一大批开源工具出现，包括 ESPNet, SpeechBrain，以及中文的 WeNet.

除了开源代码，大量免费数据也不断涌现，有力推动了产业进步。具有标志性的开源代码包括英文上的 LibriSpeech, 中文上的 THCHS30 和 AI-Shell 系列。最近，大规模开源数据开始涌现，有代表性的包括英文的 GigaSpeech 1 万小时，中文的 WeNetSpeech 1 万小时，多语种数据 M-AILABS 等。这些开源数据除了数量庞大，还有一个重要特性是多来源于互联网上的自媒体源，具有极高的复杂度，也更贴近实际应用场景。

7.4.2. 声纹识别

目前，声纹识别已应用于军事、国防、政府、金融等多个领域。在金融和社保领域，已经出现结合利用说话人确认技术来代替原有单一密码认证的新的身份认证方式。2018 年 10 月 9 日，中国人民银行正式发布了《移动金融基于声纹识别的安全应用技术规范》

金融行业标准。这预示着声纹识别成为移动金融领域唯一被金融监管部分认可的生物特征识别技术。在刑事侦察和技术侦领域，侦察人员通过采集犯罪现场的录音资料，可以对目标犯罪嫌疑人进行排查和取证；在国防安全领域，说话人识别技术可以直接帮助监听人员识别出是否有关键人员出现，有效进行敌我身份鉴别，继而完成侦听任务。未来，随着说话人识别技术的应用的推广，说话人识别技术将在更多的领域和产业中出现，在日常生活中发挥重要的作用。

7.4.3. 语音情感识别

随着语音情感识别技术的发展，相关产品的生态正在逐渐成长，已经有一些产品的落地和专利的申请，产品比如说小米公司的小爱同学，在不同场景下，可让小爱同学有 6 种不同情绪的音色，包括关心、开心、生气、惊讶、悲伤、害羞；专利中国科学院自动化研究所申请的基于微表情、肢体动作和语音的多模态情感识别方法，都展现出语音情感识别的研究前景，因为技术的不断走向成熟，未来将会有越来越多的技术产品服务于人类生活，在人机交互中表现的更加和谐。

7.5.总结及展望

7.5.1. 语音识别

语音识别技术已经逐渐走向成熟，在特定领域、特定环境下已经达到实用化程度。然而，在自由发音、高噪声、同时发音、远端声场等环境下，机器识别的性能还不能让人满意。本节对这一技术的未来发展做一展望，希望引起更多兴趣。

7.5.1.1. 远端语音识别

当前近端语音的识别性能基本可以满足很多应用场景的需求，但远端语音识别的性能依然不理想。当前远端语音识别多依赖各种麦克风阵列技术，包括各种 beamforming 技术和最近提出的基于 DNN 的信道融合技术。除了麦克风阵列，分布式麦克风技术也引起关注，但在理论和实践上还需进一步发展。相对人耳对远端声音的鲁棒性，远端语音识别性能的急剧下降可能意味着我们需要新的方法和思路，以便更深入地理解和描述声音信号的特性及其与声学模型的匹配性。

7.5.1.2. 多语种、小语言、方言识别

当前基于 DNN 的语音识别对资源丰富语言（如英语、汉语）的识别性能已经可满足实用性要求，但对小语种和方言这些资源稀缺语言的识别性能还比较差。如何利用多任务学习和迁移学习，实现对资源稀缺语言的“共享学习”，依然是比较困难的问题。特别是，如何实现多种语言在统一解码空间中解码，还需要一些探索。

7.5.1.3. 多任务协同学习

语音信号中包括说话内容、说话人、情绪 、信道等各种信息，这些信息混杂在同一信号中，在不同任务中的重要性各有不同。例如，语音识别希望只保留说话内容而去掉说话人信息，反之说话人识别希望保留说话人信息而去掉说话内容。如果将这两个任务放在一起协同学习，让每一任务可利用其它任务的信息，则有望同时提高各个任务的性能。这一协同学习也是人类学习的典型方式。

7.5.1.4. 语音-语义协同学习

语音识别的最终任务是让机器能理解人的语义，而非简单转换成文字。因此，语音识别最终要包含语义理解模块。当前语音识别和语义理解的研究还相对割裂。幸运的是，当前语义理解的主流方法同样基于 DNN/RNN 模型，这为两者的结合提供了基础。目前，端到端语音理解取得了一些进步，但还存在可扩展性差等实际问题。

7.5.1.5. 音视频多模态识别

每种技术都有其性能边界，语音识别同样如此。当噪声非常嘈杂时，单纯依赖声学信息已经很难得到合理的性能，此时引入多模态学习，将显著降低识别难度，有效提高识别性能。目前常见的是音视频多模态学习，基本方案是将音视频信息送入神经网络，利用神经网络的特征学习能力将二者进行融合。这一方法可取得一定效果，但鲁棒性低，当一路信息出现噪声时，总体识别性能反而会下降。如何有效利用多模态信息，是个重要的研究方向。

7.5.1.6. 神经网络持续学习

通常来说，一个网络优化以后，将很难再对新的数据进行学习。这显然不能满足实际应用的需要：我们希望对持续得到的新数据进行连续学习，使得模型可以持续更新，

并显不会忘记过去所学的内容。研究者提出了一些方法来解决这一缺陷，包括参数约束、隐空间再利用、Progressive 网络等。然而，这些方法的有效性还需要进一步验证。

7.5.2. 语音情感识别

情感识别是对人工智能技术探索的重要分支，其应用也是人机交互不可或缺的一部分，国内有很多单位学者都在开展研究，最近也取得了很大的进展，随着研究领域越来越广，角度越来越多，形成了很多的研究团队，国际上影响越来越大，解决了很多问题，但是仍然还有许多的缺点，比如对实验数据依赖性强、在不同情感数据库可移植性差等；进一步研究里面的基础理论，还需要很多的加强，情感的应用还需要跟其他的方向进行结合比如说：说话人识别深度的融合、多模态情感信息互相补充，当前并没有成熟的相关应用出现，所以对于加快人机交互的情感智能化进程应该着重的给予关注。

7.6.主要参考文献

[1] Frederick Jelinek, "Continuous speech recognition by statistical methods". Proceedings of the IEEE 64(4), 1976

[2] George E Dahl, Dong Yu, Li Deng, Alex Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, TASLP, 2011/4/5

[3] L. Bahl, F. Jelinek, R. Mercer: A maximum likelihood approach to continuous speech recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, No. 2, pp. 179－190, March 198

[4] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The journal of machine learning research, 2003, 3: 1137-1155.

[5] A. Salomaa, M. Soittola: Automata-Theoretic Aspects of Formal Power Series. Springer-Verlag, New York, NY, USA, 1978.

[6] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]//Nips workshop on deep learning for speech recognition and related applications. 2009, 1(9): 39.

[7] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research

groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[8] Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545.

[9] Sak H, Senior A, Rao K, et al. Fast and accurate recurrent neural network acoustic models for speech recognition[J]. arXiv preprint arXiv:1507.06947, 2015.

[10] Towards End-To-End Speech Recognition with Recurrent Neural Networks. A Graves, N Jaitly - ICML, 2014

[11] Sequence Transduction with Recurrent Neural Networks, Alex Graves, 2012

[12] Chan et al., Listen, attend and spell: A neural network for large vocabulary conversational speech recognition.

[13] Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5884-5888.

[14] Sperber M, Niehues J, Neubig G, et al. Self-attentional acoustic models[J]. arXiv preprint arXiv:1803.09519, 2018.

[15] Zhang Q, Lu H, Sak H, et al. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7829-7833.

[16] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.

[17] Al-Rfou R, Choe D, Constant N, et al. Character-level language modeling with deeper self-attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 3159-3166.

[18] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.

[19] Irie K, Zeyer A, Schlüter R, et al. Language modeling with deep transformers[J]. arXiv preprint arXiv:1905.04226, 2019.

[20] Schneider S, Baevski A, Collobert R, et al. wav2vec: Unsupervised pre-training for speech recognition[J]. arXiv preprint arXiv:1904.05862, 2019.

[21] Baevski A, Zhou H, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. arXiv preprint arXiv:2006.11477, 2020.

[22] W2V-BERT: COMBINING CONTRASTIVE LEARNING AND MASKED LANGUAGE MODELING FOR SELF-SUPERVISED SPEECH PRE-TRAINING

[23] SADJADI S O, KHEYRKHAH T, TONG A, GREENBERG C, REYNOLDS D, SINGER E, MASON L, HERNANDEZ-CORDERO J. The 2016 NIST Speaker Recognition Evaluation[C/OL]//Interspeech 2017. Stockholm, Sweden: ISCA, 2017: 1353-1357[2019-03-12]. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0458.html. DOI:10.21437/Interspeech.2017-458.

[24] SADJADI S O, GREENBERG C S, SINGER E, REYNOLDS D A, MASON L, HERNANDEZ-CORDERO J. The 2019 nist audio-visual speaker recognition evaluation[J]. Proc. Speaker Odyssey (submitted), Tokyo, Japan, 2020.

[25] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: a large-scale speaker identification dataset[J]. arXiv preprint arXiv:1706.08612, 2017.

[26] FURUI S. Cepstral analysis technique for automatic speaker verification[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(2): 254-272. DOI:10.1109/tassp.1981.1163530.

[27] RABINER L. Fundamentals of speech recognition[J]. Fundamentals of speech recognition, 1993.

[28] ALAM M J, KINNUNEN T, KENNY P, OUELLET P, O'SHAUGHNESSY D. Multitaper MFCC and PLP features for speaker verification using i-vectors[J]. Speech communication, 2013, 55(2): 237-251. DOI:10/f4h97n.

[29] MÜLLER M. Dynamic time warping[J]. Information retrieval for music and motion, 2007: 69-84. DOI:10/ftwjpf.

[30] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using

adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3): 19-41. DOI:10/ccm95g.

[31] DEHAK N, DUMOUCHEL P, KENNY P. Modeling prosodic features with joint factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2095-2103. DOI:10/fgp8b2.

[32] DEHAK N, KENNY P J, DEHAK R, DUMOUCHEL P, OUELLET P. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798. DOI:10.1109/tasl.2010.2064307.

[33] HATCH A O, KAJAREKAR S S, STOLCKE A. Within-class covariance normalization for SVM-based speaker recognition. [C]//Interspeech. . DOI:10.21437/interspeech.2006-183.

[34] SOLOMONOFF A, CAMPBELL W M, BOARDMAN I. Advances in channel compensation for SVM speaker recognition[C]//Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. IEEE, 2005: I-629. DOI:10.1109/icassp.2005.1415192.

[35] IOFFE S. Probabilistic linear discriminant analysis[C]//European Conference on Computer Vision. Springer, 2006: 531-542.

[36] VARIANI E, LEI X, MCDERMOTT E, MORENO I L, GONZALEZ-DOMINGUEZ J. Deep neural networks for small footprint text-dependent speaker verification[C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014: 4052-4056. DOI:10.1109/icassp.2014.6854363.

[37] SNYDER D, GARCIA-ROMERO D, SELL G, POVEY D, KHUDANPUR S. X-Vectors: Robust DNN Embeddings for Speaker Recognition[C/OL]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, 2018: 5329-5333[2019-06-03]. https://ieeexplore.ieee.org/document/8461375/. DOI:10.1109/ICASSP.2018.8461375.

[38] HEIGOLD G, MORENO I, BENGIO S, SHAZEER N. End-to-end text-dependent speaker verification[C]//2016 IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP). IEEE, 2016: 5115－5119. DOI:10/gfxcrd.

[39] FONT R. A denoising autoencoder for speaker recognition. results on the mce 2018 challenge[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6016－6020. DOI:10/gnhcdd.

[40] PARK D S, CHAN W, ZHANG Y, CHIU C-C, ZOPH B, CUBUK E D, LE Q V. SpecAugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[41] LEE K A, WANG Q, KOSHINAKA T. The CORAL+ algorithm for unsupervised domain adaptation of PLDA[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5821－5825. DOI:10/gmwbtz.

[42] KANG W H, MUN S H, HAN M H, KIM N S. Disentangled speaker and nuisance attribute embedding for robust speaker verification[J]. IEEE Access, 2020, 8: 141838－141849. DOI:10.1109/access.2020.3012893.

[43] ANGUERA X, BOZONNET S, EVANS N, FREDOUILLE C, FRIEDLAND G, VINYALS O. Speaker diarization: A review of recent research[J]. IEEE Transactions on audio, speech, and language processing, 2012, 20(2): 356－370. DOI:10.1109/tasl.2011.2125954.

[44] AUCKENTHALER R, CAREY M, LLOYD-THOMAS H. Score Normalization for Text-Independent Speaker Verification Systems[J]. 2000, 10: 13. .

[45] WANG L, WANG J, LI L, ZHENG T F, SOONG F K. Improving speaker verification performance against long-term speaker variability[J]. Speech Communication, 2016, 79: 14－29. DOI:10/f8nr75.

[46] BIE F, WANG D, ZHENG T F, TEJEDOR J, CHEN R. Emotional adaptive training for speaker verification[C]//2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, 2013: 1－4. DOI:10.1109/apsipa.2013.6694123.

[47] LU L, DONG Y, ZHAO X, LIU J, WANG H. The effect of language factors for robust speaker recognition[C]//2009 IEEE International Conference on

Acoustics, Speech and Signal Processing. IEEE, 2009: 4217－4220.

[48] PATIL H A, KAMBLE M R. A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System[C/OL]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Honolulu, HI, USA: IEEE, 2018: 1047－1053[2019－05－28]. https://ieeexplore.ieee.org/document/8659666/. DOI:10.23919/APSIPA.2018.8659666.

[49] LI L, WANG D, ZHANG C, ZHENG T F. Improving short utterance speaker recognition by modeling speech unit classes[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(6): 1129－1139. DOI:10/gmnnmz.

[50] ZHENG T F. To push on the systematism of trusted identity authentication technologies[J]. Journal of Information Security Research (in Chinese), 2020, 6(7): 574.

[51] CHEN L-W, GUO W, DAI L-R. Speaker verification against synthetic speech[C]//2010 7th International Symposium on Chinese Spoken Language Processing. IEEE, 2010: 309－312. DOI:10.1109/iscslp.2010.5684887.

[52] DE LEON P L, STEWART B, YAMAGISHI J. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. [C]//Interspeech. . DOI:10.21437/Interspeech.2012-135.

[53] SATOH T, MASUKO T, KOBAYASHI T, TOKUDA K. A robust speaker verification system against imposture using an HMM-based speech synthesis system[C]//Seventh European Conference on Speech Communication and Technology. .

[54] KANEKO T, KAMEOKA H, TANAKA K, HOJO N. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion[C/OL]//arXiv:1904.04631 [cs, eess, stat]. [2019－06－07]. http://arxiv.org/abs/1904.04631.

[55] WANG Q, LEE K A, KOSHINAKA T. Using Multi-Resolution Feature Maps with Convolutional Neural Networks for Anti-Spoofing in ASV[C/OL]//Odyssey 2020 The Speaker and Language Recognition Workshop. ISCA, 2020: 138－142[2020－11－21]. http://www.isca-speech.org/archive/Odyssey_2020/abstracts/19.html.

DOI:10.21437/Odyssey.2020-20.

[56] SHANG W, STEVENSON M. A Preliminary Study of Factors Affecting the Performance of a Playback Attack Detector[C/OL]//2008 Canadian Conference on Electrical and Computer Engineering. Niagara Falls, ON, Canada: IEEE, 2008: 000459－000464[2021－03－18]. http://ieeexplore.ieee.org/document/4564576/. DOI:10.1109/CCECE.2008.4564576.

[57] GA\LKA J, GRZYWACZ M, SAMBORSKI R. Playback attack detection for text-dependent speaker verification over telephone channels[J]. Speech Communication, 2015, 67: 143－153. DOI:10.1016/j.specom.2014.12.003.

[58] TODISCO M, DELGADO H, EVANS N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017, 45: 516－535. DOI:10.1016/j.csl.2017.01.001.

[59] CHENG X, XU M, ZHENG T F. Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019[C]//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Lanzhou, China: IEEE, 2019: 540－545. DOI:10.1109/APSIPAASC47483.2019.9023158.

[60] KAMBLE M R, PATIL H A. Novel Variable Length Energy Separation Algorithm Using Instantaneous Amplitude Features for Replay Detection.[C]//INTERSPEECH. . DOI:10.21437/Interspeech.2018-1687.

[61] TAK H, PATIL H A. Novel linear frequency residual cepstral features for replay attack detection.[C]//INTERSPEECH. . DOI:10.21437/Interspeech.2018-1702.

[62] ZHANG L, TAN S, YANG J. Hearing Your Voice Is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication[C/OL]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS ’17. Dallas, Texas, USA: ACM Press, 2017: 57－71[2020－04－12]. http://dl.acm.org/citation.cfm?doid=3133956.3133962. DOI:10.1145/3133956.3133962.

[63] CHEN S, REN K, PIAO S, WANG C, WANG Q, WENG J, SU L, MOHAISEN A. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones[C]//Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on. IEEE, 2017: 183－195.

[64] MOCHIZUKI S, SHIOTA S, KIYA H. Voice liveness detection using phoneme-based pop-noise detector for speaker verification[J]. Threshold, 2018, 5: 0.

[65] Zhou H, Du J, Zhang Y, et al. Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2617-2629.

[66] Hou M, Li J, Lu G. A supervised non-negative matrix factorization model for speech emotion recognition[J]. Speech Communication, 2020, 124: 13-20.

[67] Song P , Zheng W . Feature Selection Based Transfer Subspace Learning for Speech Emotion Recognition[J]. IEEE Transactions on Affective Computing, 1949:1-1.

[68] Wang X, Wang M, Qi W, et al. A Novel end-to-end Speech Emotion Recognition Network with Stacked Transformer Layers[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6289-6293.

[69] Xu M, Zhang F, Cui X, et al. Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6319-6323.

[70] Zhang J, Jiang L, Zong Y, et al. Cross-Corpus Speech Emotion Recognition Using Joint Distribution Adaptive Regression[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 3790-3794.

[71] Gao Y, Liu J X, Wang L, et al. Domain-Adversarial Autoencoder with Attention Based Feature Level Fusion for Speech Emotion Recognition[C]//ICASSP 2021-2021 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6314-6318.

[72] Cai X, Wu Z, Zhong K, et al. Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network[C]//2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2021: 1-5.

[73] Liu J, Wang H. A Speech Emotion Recognition Framework for Better Discrimination of Confusions}}[J]. Proc. Interspeech 2021, 2021: 4483-4487.

[74] Lian Z, Liu B, Tao J. CTNet: Conversational transformer network for emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 985-1000.

[75] Liu J, Chen S, Wang L, et al. Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6339-6343.

[76] Chen J, Wang C, Wang K, et al. HEU Emotion: a large-scale database for multimodal emotion recognition in the wild[J]. Neural Computing and Applications, 2021: 1-17.

[77] Liu J, Song Y, Wang L, et al. Time-Frequency Representation Learning with Graph Convolutional Network for Dialogue-Level Speech Emotion Recognition}}[J]. Proc. Interspeech 2021, 2021: 4523-4527.

[78] Liu J, Liu Z, Wang L, et al. Temporal Attention Convolutional Network for Speech Emotion Recognition with Latent Representation[C]//INTERSPEECH. 2020: 2337-2341.

[79] Zhang J, Liu Z, Liu P, et al. Dual-Waveform Emotion Recognition Model for Conversations[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[80] Luo H, Han J. Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2047-2060.

[81] Wang Y, Shen G, Xu Y, et al. Learning Mutual Correlation in Multimodal

Transformer for Speech Emotion Recognition}}[J]. Proc. Interspeech 2021, 2021: 4518-4522.

[82] Li R, Zhao J, Jin Q. Speech Emotion Recognition via Multi-Level Cross-Modal Distillation}}[J]. Proc. Interspeech 2021, 2021: 4488-4492.

[83] Fu C, Liu C, Ishi C T, et al. MAEC: Multi-Instance Learning with an Adversarial Auto-Encoder-Based Classifier for Speech Emotion Recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6299-6303.

[84] Xu M, Zhang F, Yang J, et al. Exploring the Influence of Noise in Speech Emotion Recognition Devices for Internet of Thing[C]//2020 IEEE International Conference on Energy Internet (ICEI). IEEE, 2020: 128-133.

[85] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021a. MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3682 – 3692.

[86] Jia-Chen Gu, Zhen-Hua Ling, YuWu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021b. Detecting Speaker Personas from Conversational Texts. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1126 – 1136.

[87] Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021a. Emotion Inference in Multi-Turn Conversations with Addressee-Aware Module and Ensemble Strategy. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3935 – 3941.

[88] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021b. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1204 – 1214.

[89] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021c. Conversations Are Not Flat: Modeling the Dynamic Information Flow

across Dialogue Utterances. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 128–138.

[90] Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. DialogueCSE: Dialogue-based Contrastive Learning of Sentence Embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2396–2406.

[91] Han Wu, Kun Xu, and Linqi Song. 2021. CSAGN: Conversational Structure Aware Graph Network for Conversational Semantic Role Labeling. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2312–2317.

[92] Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic aware multi-turn dialogue modeling. Proceedings of the Thirty- Fifth AAAI Conference on Artificial Intelligence (AAAI-21).

[93] Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving Dialog Systems for Negotiation with Personality Modeling. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 681–693.

[94] Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words. Proceedings of the 28th International Conference on Computational Linguistics, pages 437–449.

[95] Zhuosheng Zhang, Hai Zhao. 2021. Structural Pre-training for Dialogue Comprehension. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5134–5145.

[96] Leilan Zhang, Qiang Zhou. 2019. Topic Segmentation for Dialog Stream, Proc. of APASIPA-2019.

[97] Xuejing Zhang, Xueqiang Lv, and Qiang Zhou. 2018. Chinese Dialogue Action Analysis Using Multi-Task Learning Framework. Proc. of IALP 2018.

[98] 张禹尧，蒋玉茹，张仰森. (2021) 基于多尺度自注意力增强的多方对话角色识别

方法，《中文信息学报》，31(6)：101-109.

[99] 周强 (2017) 汉语日常会话的对话行为分析标注研究，《中文信息学报》，35(5)：75-82.

[100] Yoshimura T, Tokuda K, Masuko T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis[C] //Sixth European Conference on Speech Communication and Technology. 1999.

[101] Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis[C] //IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2000, 3: 1315-1318.

[102] Kang S, Meng H. Statistical parametric speech synthesis using weighted multi-distribution deep belief network[C] // Annual Conference of the International Speech Communication Association (INTERSPEECH). 2014.

[103] Ling Z H, Kang S Y, Zen H, et al. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends[J]. IEEE Signal Processing Magazine, 2015, 32(3): 35-52.

[104] Qian Y, Fan Y, Hu W, et al. On the training aspects of deep neural network (DNN) for parametric TTS synthesis[C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 3829-3833.

[105] Fan Y, Qian Y, Xie F L, et al. TTS synthesis with bidirectional LSTM based recurrent neural networks[C] //Annual Conference of the International Speech Communication Association (INTERSPEECH). 2014.

[106] Fernandez R, Rendel A, Ramabhadran B, et al. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks[C] //Annual Conference of the International Speech Communication Association (INTERSPEECH). 2014: 2268-2272.

[107] Zen H, Sak H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C] //IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4470-44745.

[108] Wang Y, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[C] // Annual Conference of the International Speech Communication Association (INTERSPEECH). 2017: 4006-4010.

[109] van den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[C] //9th ISCA Speech Synthesis Workshop. 125-12.

[110] Kalchbrenner N, Elsen E, Simonyan K, et al. Efficient neural audio synthesis[C] //International Conference on Machine Learning (ICML). 2018: 2410-2419.

[111] Ping W, Peng K, Gibiansky A, et al. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning[C] //International Conference on Learning Representations (ICLR). 2018.

[112] Arik S, Chen J, Peng K, et al. Neural voice cloning with a few samples[C] //Advances in Neural Information Processing Systems (NIPS). 2018: 10019-10029.

[113] Jia Y, Zhang Y, Weiss R, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C] //Advances in Neural Information Processing Systems (NIPS). 2018: 4480-4490.

[114] Zhang J X, Ling Z H, Dai L R. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 540-552.

[115] Lee Y, Lee S Y, Rabiee A. Emotional end-to-end neural speech synthesizer[C] //Advances in Neural Information Processing Systems (NIPS). 2017.

[116] Skerry-Ryan R J, Battenberg E, Xiao Y, et al. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron[C] //International Conference on Machine Learning (ICML). 2018.

[117] Wang Y, Stanton D, Zhang Y, et al. Style Tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis[C] //International Conference on Machine Learning (ICML). 2018: 5180-5189.

[118] Wang M, Wu Z Y, Wu X X, Meng H, Kang S Y, Jia J, Cai L H. Emphatic

speech synthesis and control based on characteristic transferring in end-to-end speech synthesis[C] //Proc. ACII Asia. 2018.

[119] Cao Y W, Wu X X, Liu S X, Yu J W, Li X, Wu Z Y, Liu X Y, Meng H. End-to-end code-switched TTS with mix of monolingual recordings[C] //Proc. ICASSP. 2019: 6935-6939.

[120] Zhang Y, Weiss R J, Zen H, Wu Y H, Chen Z F, Skerry-Ryan R J, Jia Y, Rosenberg A, Ramabhadran B. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning[C] //Proc. INTERSPEECH. 2019: 2080-2084.

[121] Zhang Y J, Pan S F, He L, Ling Z H. Learning latent representations for style control and transfer in end-to-end speech synthesis[C] //Proc. ICASSP. 2019: 6945-6949.

[122] Hsu W N, Zhang Y, Weiss R J, Zen H, Wu Y H, Wang Y X, Cao Y, Jia Y, Chen Z F, Shen J, Nguyen P, Pang R M. Hierarchical generative modeling for controllable speech synthesis [C] //Proc. ICLR. 2019.

[123] Wu Z, Virtanen T, Kinnunen T, et al. Exemplar-based voice conversion using non-negative spectrogram deconvolution [C] // The Eighth ISCA Tutorial and Research Workshop on Speech Synthesis, Barcelona, Spain, August 31-September 2, 2013. ISCA, 2013: 201-206.

[124] Stylianou Y, Cappé O, Moulines E. Continuous probabilistic transform for voice conversion [J]. IEEE Trans. Speech and Audio Processing, 1998, 6(2):131-142.

[125] Toda T, Black A W, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory [J]. IEEE Trans. Audio, Speech & Language Processing, 2007, 15(8):2222-2235.

[126] Sun L, Kang S, Li K, et al. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks [C] // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. IEEE, 2015: 4869-4873.

[127] Desai S, Black A W, Yegnanarayana B, et al. Spectral mapping using

artificial neural networks for voice conversion [J]. IEEE Trans. Audio, Speech & Language Processing, 2010, 18(5): 954-964.

[128] Sun L, Li K, Wang H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training [C] // IEEE International Conference on Multimedia and Expo, ICME 2016, Seattle, WA, USA, July 11-15, 2016. IEEE Computer Society, 2016: 1-6.

[129] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning [C] // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. Curran Associates, 2017: 6306-6315.

[130] Kingma D P, Welling M. Auto-encoding variational Bayes [C] // 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014.

[131] Liu S, Zhong J, Sun L, et al. Voice conversion across arbitrary speakers based on a single target speaker utterance [C] // 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. ISCA, 2018: 496-500.

[132] Qian K, Zhang Y, Chang S, et al. AutoVC: Zero-shot voice style transfer with only autoencoder loss [C] // Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. PMLR, 2019: 5210-5219.

[133] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion [J]. J. Mach. Learn. Res., 2010, 11:3371-3408.

[134] Liu L, Ling Z, Jiang Y, et al. Wavenet vocoder with limited training data for voice conversion [C] // 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. ISCA, 2018: 1983-1987.

[135] van den Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio [C] // The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA,

USA, 13-15 September 2016. ISCA, 2016: 125.

第八章　社会媒体处理研究进展、现状及趋势

8.1.研究背景与意义

伴随互联网和移动通信技术的发展，众多社会媒体如微博、微信、知乎等应运而生并迅速普及。相比于传统的大众媒体，从形式观之，社会媒体表现出鲜明的社会化和交互式特点，信息生产、传播和消费的模式均发生了根本性的改变。就功能而言，社会媒体也不再局限于新闻宣传工具，日益成为推动社会发展和治理的重要平台。

社会媒体的普及性和重要性催生了社会媒体处理（Social Media Processing， SMP）技术。它旨在利用计算机和人工智能技术，基于社会媒体中存在的海量异构数据资源挖掘和分析人类的认知、心理和行为模式以及社会的发展变迁规律，并服务于大数据和人工智能时代的政治、经济、教育、宣传等全方位的发展。近年来，在国家战略“互联网+”、“大数据战略”、“新一代人工智能发展规划”的引导和支持下，社会媒体处理进入发展新阶段，并为学术和产业界带来新的机遇和挑战。

在学术领域，社会媒体处理推动计算机科学和人文社会科学深度交叉融合，并取得一系列突出成果。一方面，社会媒体处理所具有的大规模数据自动化处理能力为传统人文社会科学引入了新兴血液，包括先进的研究方法和宽广的研究视角。研究者们可以跨越时间和空间，从多维度（认知、心理和行为）和多主体（人与人、人与物、物与物）考察人类和社会的形态和发展规律。基于此，以 2009 年在《科学》（Science）杂志上发表《Computational Social Science》一文作为计算社会科学的诞生标志，一系列包括计算社会学、计算传播学、计算历史学、计算法学等在内的交叉学科成为新兴热点，引起学界高度关注并蓬勃发展、方兴未艾。另一方面，社会媒体处理为计算机学科带来众多新兴研究问题，包括舆情分析、个性化推荐、用户行为预测、谣言检测等重要的应用问题，推动计算机科学的自然语言处理、数据挖掘、多媒体计算等技术的快速发展和实际应用。

在产业领域，伴随当今时代的媒体形式和功能的不断泛化，社会媒体处理已经渗透到各行各业，助力于各行业的数字化、自动化和智能化转型，如智能教育、智能金融、智能司法等。依托于社会媒体的海量数据优势、万物互联特性、以及高覆盖度和普及度，社会媒体处理改变了诸多行业的传统运行模式，推动了一系列优秀的产业应用诞生，如在线教育平台、信用监管平台、案件检索和分析平台。

正如科学技术的“双刃剑”效应，社会媒体处理全面深入发展、造福社会的同时，也带来了一些问题和隐患，如信息瘟疫、虚假行为主体、算法偏见及隐私泄露等。以社会媒体处理的典型应用社交机器人为例，一方面具备智能的社交机器人可以作为服务人类的工具，应用于工作和生活的辅助管理、情感疗愈等领域，具有重要应用价值，另一方面也是长久以来“图灵测试”中验证人工智能发展程度的重要指标。然而，近年来社交机器人也成为操纵舆论、传播虚假信息的主力，广泛应用在政治传播、科技传播等领 6 域，且随着社交机器人的越发智能化，其检测和识别也会更加棘手。因此，如何应对这把“双刃剑”，也吸引了越来越多跨学科学者的关注。

综上所述，社会媒体处理已经成为一个广受跨学科学者关注的方向，具有重要的研究价值和应用前景。但与此同时，它也仍存在不完善之处，面临很多具体问题和社会隐患。本报告将对社会媒体处理领域的发展现状与关键科学问题、关键技术发展现状及趋势、产业发展现状及趋势做具体介绍。

8.2.发展现状与关键科学问题

近些年，随着互联网以及移动通信技术的发展，社会媒体迅速兴起并得到了广泛的普及，如微博、知乎、抖音、小红书等各类社会媒体平台，都拥有数以亿万计的用户。在这些社会媒体中，用户每天产生海量数据，这些数据包含文本、图片、语音、视频等多种模态，覆盖了社会学、传播学、计算机科学、历史学、认知学、心理学以及语言学等多个文理学科，同时其涉及了政治、金融、军事、外交、法律、教育、文体等人类社会中几乎每一个领域。因此，在社会媒体处理这一方向得到了学界和产业界的广泛关注，他们积极探索在海量社会媒体数据驱动下的数据统一表示、学科交叉研究、以及产学研融合发展等相关问题。

8.2.1. 社会媒体数据的统一特征表示

社会媒体数据具有跨平台、跨模态、跨语言、高噪声等显著特点。社会媒体数据同时存在于大量平行的互联网平台上，不同平台的数据之间具有高度的共生性和互补性。社会媒体数据内容往往跨越多种模态，如文本、图像、声音、视频等，不同的模态内部和之间都具有复杂的数据结构和语义关联。社会媒体数据还会跨越不同的语言，来自全球的用户经常会针对相同的焦点或话题发表各自的内容，形成跨语言的丰富数据。社会

媒体还存在高噪声的特点，其中充斥着大量重复或无意义的低价值数据，甚至存在很多虚假或欺骗性的有害数据。因此，如何构建具有强大表达能力的预训练大模型，能够将跨平台、跨模态、跨语言、高噪声的社会媒体数据在一个统一的特征空间进行准确的特征表示，是社会媒体处理领域的一个基础性的关键科学问题。

8.2.2. 多维度社会网络影响力计算问题

随着移动通信技术的发展，移动互联网应用快速普及，社会网络成了人们传播和分享信息的重要途径。影响力计算问题具有很强的实际应用价值和科学研究意义，是社会网络研究中的重要基础问题之一。现有研究工作大都在经典信息传播模型上展开，而实际应用中则需要从多种不同维度扩展经典的信息传播模型，这方面的研究工作尚不充分。如何从动态影响力、隐私保护、目标导向、网络结构这四个不同维度出发，分析社会网络中用户间影响力形成和变化的不同因素，研究对应的影响力计算问题成为社会媒体处理领域的一个基础性的关键科学问题。

8.2.3. 社会媒体大数据驱动的管理与应用研究

社会媒体环境下的新兴技术快速发展与应用催生了新模式、新业态和新人群，为社会经济生活注入了新活力，进一步丰富和拓展了社会媒体的应用创新领域和应用行业。在社会媒体驱动下，数据管理与行业应用呈现出高频实时、深度定制化、全周期沉浸式交互、跨行业整合、跨学科融合、多主体决策等特性。因此，针对社会媒体驱动的数据特点与行业领域应用特色，在充分发挥多学科（社会学、政治学、传播学、心理学等）交叉研究优势的基础上，如何探索社会媒体大数据驱动管理与应用范式，研究以法律、金融、教育、医疗、政府管理等行业为导向的社会媒体数据的价值发现方法，进而引领相关行业领域的数据管理与应用范式的机理转变，成为社会媒体处理领域的一个基础性的关键科学问题。

8.3.关键技术进展及趋势

社交媒体处理领域涉及多个前沿任务以及一些交叉学科的研究问题，主要包括情感计算，计算传播学，计算社会学，数据挖掘，表示学习，智能教育，智能金融，计算历

史学，智慧司法学，社交机器人，舆情计算，社交多媒体。以下针对各个课题进行任务定义，进展和未来发展趋势的介绍。

8.3.1. 情感计算

文本情感计算的主要任务是研究自然语言中的主观信息（如情感、情绪、态度、评价等）的提取、分析、理解和生成。文本作为人类表达情感情绪的重要载体，文本情感计算是情感计算的一个重要组成部分，也是自然语言处理、文本挖掘等领域的重要分支。文本情感计算可以视为以主观信息为对象的自然语言处理技术。自然语言处理包含自然语言理解、自然语言生成、知识图谱等领域。同样地，文本情感计算也涵盖文本情感分析、情感文本生成、情感图谱构建等方面的研究，它在舆情分析、心理健康监测、评论分析与生成、商业决策等方面有着广泛应用。

8.3.1.1. 近年发展和主流方法

8.3.1.1.1. 文本情感分析

文本情感分析（Liu 2012）是对文本中的主观信息进行分析和理解的技术，具体包括针对情感、情绪、态度、立场等主观信息的分类、抽取、归纳和推理等。文本情感分析按照任务目标可以分为情感极性分类（如正面、负面、中性）、离散情绪分类（如喜、怒、哀、乐等）、立场分类（如支持、反对、中立）、情感信息抽取、情感信息摘要等；按照文本粒度，可以分为词语级、句子级、文档级、属性级情感分析等级别。文本情感分析存在的主要挑战包括：（1）复杂语境下情感分析精度降低（Li et., 2019）：一些复杂语言结构（如否定、转折、隐式情感等）使得情感分析系统的精度显著下降，这一问题广泛存在于各种情感分析任务中；（2）领域适应问题（Gong et. 2019）：在某一领域（即源领域）标注样本上学习得到的情感分析模型通常只在相同领域的测试样本上表现较好，换到其他领域（即目标领域）时，算法性能往往会大打折扣。早期的文本情感分析技术主要针对文本情感的分类，其方法主要分为两类：基于情感字典的规则化方法和基于情感特征的统计机器学习方法。随着深度学习的深入发展，大量的神经网络模型被引入到情感分析任务中，包含卷积神经网络、循环神经网络、递归神经网络、注意力机制网络等。近年来，随着预训练语言模型的兴起，以 BERT 和 GPT 为代表的预训练语言模型

在不同的情感分析任务中均取得了较大的成功。

8.3.1.1.2. 情感文本生成

情感文本生成任务的目标是让模型生成符合指定的情感类别的文本（Zhou et., 2018）。具体而言，生成的文本应当表达出任务指定的情感类别，如开心、难过、愤怒等，这既可以通过情感相关的关键词体现（如开心与“享受”、难过与“哭泣”等），也可以通过隐喻等手法体现（如在难过的情感类别下，“我的心头阴霾不散”）。该任务的挑战有两点：（1）模型生成的文本应该语法正确、通顺连贯。（2）在保证语法性的前提下，生成文本应该蕴含指定的情感类别，并避免产生与指定情感类别矛盾的表述，以防造成歧义。情感文本生成的技术在早期大多基于 RNN 语言模型的方法。近年来，随着预训练模型的发展，情感可控的文本生成逐渐以 GPT 等预训练模型作为基座，并取得了更强大的效果。现有研究主要关注两方面的问题。（1）如何建模情感的表达过程、让文本生成受控于指定情感。（2）如何丰富情感表达的方式和内容，以提高生成的多样性和信息量。针对第 1 点问题，由于情感表达具有显性（如情感关键词）和隐性（如隐喻）的特点，情感表达也是一个动态的过程（有些词语的情感表达强度大，有些强度小），因此现有研究大多采用将拷贝网络与动态记忆单元相结合的方式。一方面，拷贝网络可以显式地在生成文本中插入情感词，另一方面，动态记忆单元可以控制表达情感的过程，在已生成出表达情感的词语后，适时控制生成过程的结束。针对第 2 点问题，由于模型的输入信息十分有限（只有指定的情感类别），因此现有研究大多利用外部知识丰富情感表达的内容（Gao et., 2021）。例如，通过在常识知识图谱检索与情感类别相关的实体（如难过与“分手”、“失业”等）来提升生成文本的信息量。

8.3.1.1.3. 情感图谱构建

传统情感分析方法在特定领域下构建情感词典，依据情感词与文本的映射关系能够实现快速自动情感分析。然而，同一情感词在不同领域和不同方面的情感倾向可能会不同，现有领域情感词典的一个突出问题是缺乏细粒度的、多领域及多方面自适应的情感常识，难以应对多领域的情感分析。当前基于深度学习的情感分析方法依赖于大量高质量标注训练样本，人工标注成本昂贵，同样面临难以实现多领域及多方面自适应的实时在线情感分析的挑战。为了弥补情感计算依赖大规模标注数据、具有强领域特性的特点

中，常常会引入外部的情感知识库提供监督信息，提高模型的泛化性能。然而，当前常采用的外部知识库存存在以下三个问题：（1）缺乏领域适应性：当前常用的情感词典常常只适用于某领域，缺乏领域泛化能力。如情感词“快”在餐厨领域中的“平底锅热得快”表达积极情感，而在电器领域中的“电池消耗快”表达消极情感。（2）缺乏方面适应性：在同一领域中，同一情感词在不同方面的情感极性可能会不同。如 “电池消耗快”以及“系统运行快”，在现存外部知识库缺乏方面泛化能力。（3）缺乏情感推理能力：现存的情感词典以及外部知识库往往只建立词语与情感的一对一的映射关系，无法建模情感词间关系、方面词间关系，以及方面词与情感词的动态多关系。从而导致情感常识成为离散点，无法进行有效关联而失去了情感的推理能力。

8.3.1.1.4. 文本论辩分析

论辩（Argumentation）（Van et. 2002）研究辩论和推理的过程，是一个涉及逻辑、哲学和语言等多学科的研究领域。近年来，在人工智能领域研究论辩催生了一个新的研究课题，即计算论辩学（Computational Argumentation）（Eger et. 2017 ）。学者试图将人类关于逻辑论证的认知模型与计算模型结合起来，以提高人工智能自动推理的能力。论辩挖掘是计算论辩中的重要任务，以文本中包含论辩性内容的部分作为研究对象，旨在自动化识别论辩性文本的结构，论辩语义单元直接的逻辑交互关系等。论辩文本中往往呈现逻辑推理过程，因此语义结构复杂；其文本内容有高度的领域相关性，对于方法的领域迁移性提出了很高的要求；论辩文本体现了人类高级的认知能力，是对人类世界理解的综合运用，依赖于知识融合。对于论辩挖掘的研究主要经过以下几个阶段，（1）理论迁移：对经典论辩理论的迁移和改造使其具备可计算的特点。（2）单体式论辩文本理解：研究论辩基本单元识别和关系分类方法，设计到不同领域的小规模语料标注。（3）交互式论辩文本理解：针对多人参与的论辩场景，研究文本分析框架以及论辩方法（Ji et. 2021）。（4）论辩文本自动生成（Hua and Wang, 2019）：针对某一个特定主题或者其它用户的一段论辩性文本，自动化生成论辩内容。目前的研究热点为交互式论辩文本理解和论辩文本自动生成两个部分。在初期，学者们采用基于特征工程的论辩文本理解方法，近几年基于神经网络的文本编码解码框架开始成为主流。

8.3.1.2. 未来趋势和挑战

8.3.1.2.1. 文本情感分析

虽然现阶段以 BERT 和 GPT 为代表的预训练语言模型在不同的情感分析任务中均取得了成功，但是大部分工作仍是采用预训练加微调的范式。这种范式的缺陷在于语言模型在预训练过程中是脱离于下游情感分析任务的。为了解决此缺陷，最新的基于提示（prompt）的学习范式可能会成为一个比较有发展潜力的研究方向，如何针对下游不同的情感分析任务设计符合预训练语言模型训练目标的 prompt 是值得深入探究的问题。

8.3.1.2.2. 情感文本生成

情感文本生成未来技术发展有两方面的趋势。一是利用大型预训练模型内部的知识。在不引入外部信息的情况下，使得生成文本在情感可控的前提下更加多样、丰富。近期基于提示学习的方法展现出触发大模型内部知识的潜力，未来的情感文本生成的研究或许可以与提示学习方法相结合。二是高效地融合外部知识信息。外部知识信息往往能够提供更好的可控性。然而在基座模型越来越大的趋势下，传统的为小模型所设计融合外部信息的方法可能不再适用（受限于复杂度和效率），此时利用外部知识的方法需要更高的可拓展性。

8.3.1.2.3. 情感图谱构建

针对现有方法难以高效处理多领域及多方面自适应、情感常识离散、缺乏推理机制而难以进行情感推理等问题，其中的一个技术发展趋势是将情感词在多领域、多方面的动态情感倾向知识化。通过构建面向多领域多方面的情感知识图谱，利用知识图谱丰富的表达能力，可以实现领域细粒度情感知识化，通过情感常识关联整合、建模方面词和情感词之间的层级逻辑关系，形成情感知识图谱，有利于领域知识、方面知识及情感知识的动态关联、聚合以及推理，为情感计算的应用，如高效实时的在线情感分析、情感注入的对话系统、情感注入的故事生成等提供具有动态精准的领域自适应情感常识。

8.3.1.2.4. 文本论辩分析

论辩分析的未来研究主要会在三个方向展开：（1）不同场景和粒度的论辩性内容表示方法。从单一论点到论辩性段落再到同一主题下的多立场论点，到整个论辩性文本的知识库构建，这对于论辩性文本挖掘是核心问题但相关的研究还很少。（2）大规模语料集合的构建。目前的论辩性文本研究很大程度上受到数据集合规模小、领域分散的限制，如何构建有标注、无标注的大规模论辩性文本是一个重要课题。（3）论辩性文本生成机制和方法研究。相比叙述性文本，论辩性文本的产生更多的依赖于人类的逻辑推理能力，如何将推理方法融入到文本生成过程中对于论辩内容的自动生成至关重要。

8.4.计算传播学

计算传播学是计算社会科学应用于传播学的研究分支（祝建华等，2014；王成军，2015 & 2017）。它主要关注人类传播行为的可计算性基础，以传播网络分析、传播文本挖掘、数据科学等为主要分析工具，以非介入方式大规模地收集并分析人类传播行为数据，挖掘人类传播行为及过程背后的模式和法则，分析模式背后的生成机制与基本原理，可以被广泛地应用于数据新闻、健康传播、政治传播、计算广告等场景。计算传播学的重要应用领域是计算传播产业（王成军，2016），例如，数据新闻、计算广告、媒体推荐系统、算法新闻等目标（张伦等，2021）。

8.4.1. 近年发展和主流方法

8.4.1.1. 与传统社会科学研究方法的结合

计算传播学的发展根植于传统传播学的研究基础之上，它以计算方法为特色的同时，也继承了传播学研究的问题意识和理论诉求。在这一过程中，近年来的计算传播学研究在方法运用上寻求突破和创新，即体现出本学科内与传统传播学量化研究方法的纵向融合，也有跨学科间对文理交叉研究范式的横向探索。针对不同的研究目的，综合运用传统社会科学研究方法（如深度访谈、实验法、问卷调查）与计算研究方法（如社会网络分析、ABM 仿真模拟、主题建模），将两类方法有机结合，真正实现方法服务于研究目的。例如，在探讨社交网络为社会运动提供的话语机会，一项研究（Mooijman et al., 2018）

利用 Twitter 上公众发帖数据和运动事件信息进行时间序列分析的研究发现，暴力抗争更容易引发道德化讨论；反过来，道德化讨论的热度也预测了暴力抗争事件的发生概率，在此基础上，研究进一步通过问卷调查的方法，发现了人们对抗议道德化程度越高，也就越容易接纳暴力抗争手段。

8.4.1.2. 视觉传播技术的融入

计算视觉技术的融入是近年来计算传播学的重要进展之一。在传统内容分析法难以适应海量在线图像和视频分析的背景下，计算视觉技术开始为传播学所关注。该技术能够对大规模视觉数据进行自动分析，实现物体检测、脸部识别和情感分析等目标。对于传播学而言，计算视觉技术为研究者挖掘视觉数据对人类社会的影响、发现新的理论“绿洲”提供了新的测量工具。目前该技术主要运用于政治传播领域的研究，如分析不同党派媒体的视觉偏见、政客的媒体视觉呈现、社会运动图像的情感分析等。主要涉及三类技术路径：一是借助开源计算视觉库或商业 API 执行标准化任务，如人脸识别；二是借助机器学习方法，通过有监督或无监督的机器学习，实现分类和聚类等图像处理任务；三是与图像亮度、色彩等属性相关的计算美学分析。未来研究中，研究者应该将其与其他计算传播方法（如文本分析、大规模在线实验、社会网络分析等）相结合，以一种多模态的视角探寻人类传播行为特征及其因果机制。

8.4.2. 未来趋势和挑战

8.4.2.1. 计算传播学与主流新闻学的结合

计算传播学在诸多新闻传播学理论与实践研究领域产生了深远的影响。在新闻学实践领域，最重要的影响在于对“数据新闻”这一领域的推动。数据新闻是基于数据的抓取、挖掘、统计、分析和可视化呈现的新型新闻报道方式；其旨在通过数据挖掘与信息可视化技术，将新闻事件以信息图等互动形式展现，从而使得受众更直观、更客观理解新闻事件。数据新闻的出现在一定程度上改变了传统新闻生产方式。

8.4.2.2. 计算传播学与主流传播学的结合

计算传播学研究方法，在融入主流传播学的过程中，一直致力于不断与主流传播学

进行结合。例如，大量的计算传播学研究在探讨经典的传播学理论在新的语境和媒介环境中的实用性——例如，议程设置理论、两级传播理论与意见领袖的作用、级联模型等。这些研究探讨的核心问题是，人类信息传播行为在新的媒介环境中，发生了哪些变化以及哪些方面没发生变化。

8.4.2.3. 计算传播学在信息传播实践中的影响

计算传播学在我国信息传播实践中，具有深远的影响。本文以健康传播、舆论传播和跨文化传播为例，分别简要阐述。在近年来的健康传播领域中涌现出大量计算范式的研究，分别围绕公众对健康疾病或事件的感知、健康促进活动的效果、在线健康社区参与、健康行为的网络传播、疾病监测等多维议题展开（Rains, 2020）。尤其在新冠肺炎疫情的全球范围内肆虐期间，来自公共卫生学、传播学、行为科学、计算机科学等多领域的学者充分利用人类传播行为数据，开展了大量具有实践应用价值的研究，通过对监测疾病爆发、探知信息流行病扩散规律等方面的研究为全球防疫工作提供了宝贵思路。在计算传播涉及的众多研究领域中，我国学者在网络舆情领域进行了广泛深入的探索。研究大多以典型舆情事件或社会热点议题为对象，采用自然语言处理、社交网络分析、时间序列分析、物理仿真建模等综合方法，分别解答了网络舆情发生发展过程中“谁在设置议程”“谁在跟随潮流”“意见如何分布”“情绪如何传播”“舆情演化规律”等多元问题。近年来国内计算舆情研究在选题上也体现出了一定的本土化特色和人文关怀，如对我国反腐议题的网民情绪分析（周莉等，2018）、对新冠疫情期间悼念李文亮医生的网络舆论洞察等（周葆华、钟媛，2021）。在跨文化传播——特别是在中国文化全球传播领域，计算传播学利用数据导向的研究范式，能够量化评估中国文化作品在海外社交媒体等平台的传播效果与信息传播结构。例如，利用计算传播学研究范式， 研究者能够分析海量传播文本语义特征，这为分析受众在不同语境中对文本进行解读提供了基本的技术解决路径。再比如，对于传播结构的量化，对于深入分析文本在异质文化语境中的国家/地区间传播路径，也提供了基本的技术解决思路。

8.5.计算社会学

计算社会科学指称了社会科学在大数据时代所呈现出的新发展、新路径和新范式。计算社会科学既是大数据时代科技进步、数据爆炸和方法创新的产物，又是社会科学长

久以来的计算传统知识积累的成果，致力于应用数据思维、数据资源和数据分析学以研究人类社会行为和社会运行规律等。近年来计算社会科学的发展使新学科成为现实，这种学科创新体现为围绕着数据驱动和算法驱动采取不同融合方式的一系列“问题解决性、应用导向”多元化进路，推动着社会科学范式转换。

8.5.1. 今年进展和主流方法

2009 年 Lazer 等（2009）人发表《计算社会科学》，标志着计算社会科学的诞生。计算社会科学是以大数据及其相关技术的应用为背景的。在这里，“大数据”可以从两个层面加以定义。狭义的“大数据”是指体量异常庞大、结构复杂，以至于传统数据处理方法难以应对的数据集。而广义的“大数据”则不仅指海量数据，还包括获取、传输、存储、挖掘、分析和应用海量数据的一系列方法、技术和模式，后者通常被称为“大数据分析学”（Big Data Analytics）。

2012 年，由来自意大利国家科研委员会的 R. Conte 领衔，来自欧美国家的 14 位学者又在《欧洲物理学刊.专号》（第 1 期）发表了《计算社会科学宣言》（Conte,2012）。这篇《宣言》从时代机遇、技术发展、方法创新、当下挑战和预期影响等五个方面全景式地说明了计算社会科学发展现状及其未来的前景。《宣言》强调，当下时代社会科学将经历一个巨大的范式转变。与实验方法相结合的计算方法，将使社会科学更接近于建立理论、经验事实和研究之间的良好连接。同时，计算社会科学的影响还将更加广泛，其提供的新方法会适用于任何以大数据为资料的研究，增进政策决策与评估的科学性。计算社会科学可以对全球范围内的社会进程进行建模和模拟，进而充分理解与今日世界相关联的、极其复杂的远距离间的相互关系，并用于支持政策制定者的决策，使他们可以有效识别社会发展的最佳路径。最后，计算社会科学的开放性还将极大地提高公民在这一决策过程中的参与程度。这些发展将会开启一个更安全、更可持续和更公平的全球社会。概言之，从 Jim Gary 的《第四范式》，经由 Lazer 等人的《计算社会科学》，再到 Conte 等人《计算社会科学宣言》，这些论断对计算社会科学给出了充分的学科想象空间。

与这些全景式的论证相媲美的是，有关社会科学分支学科的“计算化”进路的探索早在上世纪即可开启。在对于计算机技术有着特殊敏感性的经济学领域，早在 1996 年就出版了《计算经济学手册》，对应用计算机技术开展经济学研究进行了集中评价（Ammans 等，1996）。2006 年和 2014 年，《计算经济学手册》又先后推出了第二卷和第三卷，持

续跟踪计算经济学的前沿发展动态。在政治学领域，尽管尚未有“计算政治学”的统一用语，但政治学量化研究在竞选、议会政治、政治传播等领域较早引入机器学习、非结构化数据挖掘和大规模社会实验等研究方法(吴江、张小劲,2016)。例如，奥地利人工智能研究所所长 Trappl 领衔于 2005 年出版了《为和平编码：国际争端解决与干预的计算机辅助方法》，从统计建模技术介绍了计算社会科学方法在基于案例之上的辅助决策、冲突预测、打击犯罪、危机预警等方面的应用（Trappl，2006）。在社会学领域，长期从事社会模拟的 Michael W. Macy 于 2002 年发表《从要素到行动者：计算社会学与行为者模型化》，首次提出“计算社会学”概念（Macy,2002）。计算社会学是广泛应用计算机技术研究、认知和理解社会现象的社会学分支，包括计算机模拟、人工智能、复杂统计方法、社会网络分析技术等在内的多种手段和工具，通过对多样化社会互动的基础建模方式而提出并检验了关于复杂社会进程的多种理论发现。

此外，计算社会科学还广泛应用于更多的跨学科研究，包括“计算新闻学”、“计算语言学”、“计算犯罪学”、“计量分类学”以及“计算创新”等范畴，均有极其重要的研究进展。同样，在人文学科领域，包括“计算史学”、“计算法学”等分支学科，长期的研究积累加之以现代计算技术的辅助，也产出了大量令人瞩目的研究成果。2016 年，R. Michael Alvarez 编著的《计算社会科学：发现与预测》全面回顾和概括了计算社会科学的发展状况，发现计算社会科学拥有丰富的工具箱，并广泛应用于社交媒体、抗争运动、议会表决、新型政党组建、政府治理以及社会营销学等领域（Alvarez，2016）。

8.5.2. 未来趋势和挑战

计算社会科学的知识发展受到“双重驱动”的关键影响。一是“数据驱动”，即学科在何种程度上利用了本领域产生的数据；二是“算法驱动”，即学科在何种程度上发展了适合自身需要的算法和模型。就此而论，具体的学科领域之于相关社会生活领域的宽狭大小、相关社会生活领域的“数字化生存”程度高低以及数据生成能力的大小和数据密集程度的高低，作为外部主体的体量会严重限定具体学科受到“数据驱动”的压力大小。而具体的学科领域在其前期发展所积累的量化知识总量、计算能力的高低乃至于与其他计算学科的共享融合水平，作为主体的内生变量会严重影响其“算法驱动”的强度。

大数据经济学则既受益于数据驱动又归功于算法驱动，其主要分支包括大数据宏观经济研究、大数据金融学、大数据经济心理学等。刘涛雄等认为，大数据时代极大地拓

宽了信息来源、提高了获取信息的时效性，而新数据的非结构化特征对宏观经济分析的技术和方法提出了新要求（刘涛雄、徐晓飞，2015）。大数据语境下，数据噪声会影响数据质量，因而宏观经济数据挖掘变得十分重要，这就要改进数据挖掘技术，加强对非结构化和半结构化数据的挖掘。实时、快速、海量的数据为更加准确的宏观经济预测提供了可能。此外，机器学习与宏观经济分析方法的结合可以有效解决“维数灾难”，提高宏观经济分析的准确性。大数据影响着政府经济政策制定和评估的变革，进而提升政策的时效性和服务效率。刘志洋、汤珂认为，大数据时代信息产生和传递的速度空前加快，如互联网上的大量信息是实时的，移动互联网和物联网使每个人随时随地都可能制造数据，这导致经济模型可以充分利用数据的实时性，提高分析或预测的时效性，为经济预警和政策制定提供最快速的资料和依据（刘志洋、汤珂，2014）。简言之，大数据带来经济分析的方法论变革，随着信息量的极大拓展和处理信息能力的提高，经济分析从样本统计走向总体普查时代。

大数据政治学主要应用大数据分析学和海量数据资源探究新信息时代的政治现象。孟天广等认为，大数据不仅将政治场域从物理空间扩展到虚拟空间，为政府、公民、企业等行为主体的创造新的互动空间和模式，重塑各主体间的关系模式（孟天广，2018）。大数据在政治领域的影响力主要通过两种机制来体现：一方面，大数据作为一种创新制度推动着更开放、更高效和更智能的治理制度的建设，为政治行为和政策过程创造了新型平台（Meng and Su,2016）；另一方面，大数据所生产并传播的丰富信息突破了传统上政治信息传播和发挥影响的时间和空间限制，信息日益成为政治表达、政治互动、政策决策与制度运行的关键要素，进而推动着治理制度及其现实运作的变化。此外，大数据政治学对数据分析学的依赖取决于政治学研究的方法传统，大数据为政治学研究提供了诸如文本、图片、视频等定性研究素材，而数据分析学作为嫁接定性研究和定量研究的桥梁，为定性资料的定量分析、定量分析的定性阐释提供了可行性。

计算社会学是一门数据驱动的、以数据密集化为特征的交叉学科，其研究和应用的范围十分广泛。其发展大致受到四种相互区别的议题的共同作用：一是传统议题与新兴议题。计算社会学一方面着眼于利用新数据新技术来应对人类世界出现的新问题新挑战；另一方面又致力于以全新方式和全新视角重构和解读社会学研究的经典概念，包括阶层/阶级、社会流动、社会观念等。二是主体性与群体性。计算社会学的灵活性和适应性有助于从点到面地对人与群体的互动展开研究，既着眼于个人行为属性分析的基础性地位，又强调社会关系、社会网络及群体特征分析的宏观架构。三是外部条件性与内部动力性。

计算社会学一方面强调社会生活中信息规律和制度演变等外部条件的探讨，如由信息内容分布的解析到话题发现和言论传播；另一方面又关注此类外部条件与人类个体心灵之间相互作用、相互影响的关系。四是独立性与交互性：计算社会学一方面通过对范式和话语体系的重构来为以往社会学研究中的模糊概念“划界”，另一方面由强调不同概念及不同概念所代表的社会现实之间的交互性，着重探讨个体的倾向性、可信度和影响力等属性对外界条件的响应模式、传导机理和交互适应性，以及关系网络演变。

8.6.数据挖掘

数据挖掘是一个跨学科的计算机科学分支（Clifton 2010; Cortes and Vapnik 1995; Cover and Hart 1967）。它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。它指：“从资料中提取出隐含的过去未知的有价值的潜在信息”（Chakrabarti et. 2006），或“一门从大量资料或数据库中提取有用信息的科学”（Hand et. 2001）。数据挖掘一般设计六类任务：异常检测、关联规则学习、聚类、分类、回归和汇总（Fayyad et. 2001）。当前数据挖掘面临的挑战有：数据安全和隐私问题、噪声和数据的不完整性、数据分散性（不集中）、储存瓶颈、算法的效率与可扩展性等。

8.6.1. 近年发展和主流方法

数据挖掘可以看作信息技术自然进化的结果，对于数据挖掘的研究发展，同样伴随着数据存储、组织及使用的技术发展，主要分为以下几个阶段，（1）20 世纪 60 年代或更早，数据收集和数据库创建阶段，数据使用者直接对原始文件进行处理。（2）20 世纪 70 年代至 80 年代初期，数据库管理系统建立阶段，用户可以通过查询语言、用户界面、查询处理优化和事务管理，方便、灵活地访问数据。（3）20 世纪 80 年代中期至今，高级数据库系统研究阶段。在数据库管理系统建立之后, 数据库技术就转向高级数据库系统、支持高级数据分析的数据仓库和数据挖掘、基于 Web 的数据库。（4）20 世纪 80 年代后期至今，高级数据分析研究阶段。伴随着计算机硬件、数据库技术的不断进步，使得更多的数据易于存储和使用，于是面向大规模数据的深层次分析与决策的算法技术便应运而生（Hastie et. 1996）。

作为应用驱动的领域，数据挖掘吸纳了诸如统计学、机器学习、模式识别、数据库

和数据仓库、信息检索、可视化、算法、高性能计算和许多应用领域的大量技术，主要关注从指定数据挖掘任务中寻找模式类型，包括类 / 概念描述、频繁模式挖掘、分类与回归、聚类分析、离群点分析和演变分析等。当前用于数据挖掘的主流方法主要包括，（1）统计方法：预测统计学、统计假设检验等。（2）机器学习方法：监督学习、半监督学习、无监督学习和强化学习等。涉及的主流算法包括神经网络、支持向量机、最近邻居法（Han et. 2011）、高斯混合模型、朴素贝叶斯方法（Jaseena and Julie 2014）、决策树（Rish 2001）等。（3）数据库系统与数据仓库。（4）信息检索。

8.6.2. 未来趋势和挑战

从 80 年代至今这 40 多年的研究极大地推动了领域的发展，但是面对当前的大数据时代，数据挖掘还有很长的路要走。未来的数据挖掘研究主要会在横向和纵向两个维度上展开：（1）作为一门跨学科的科学分支，数据挖掘与横向的多个领域的应用密切结合是一个重要的方向。（2）理论与应用是相辅相成、共同发展的，数据挖掘在纵向的理论层面也大有可为。

由于数据采集和存储技术的迅速发展，加之数据生成与传播的便捷性，致使各领域数据的爆炸性增长。面对这些数据数量越来越庞大、数据结构逐渐复杂多样的数据集，如何设计更强大的数据挖掘方法针对多样的领域场景进行深入分析是未来的重要方向。长期以来，制药和医疗保健行业备受学者关注。实际上，冠状病毒疫苗的快速发展直接归功于药物试验数据挖掘技术的进步，未来数据挖掘技术还将帮助医药学获得更大的突破。由于当前的数字化进程加速，金融领域的海量数据信息对于金融风控、市场分析等应用也变得至关重要。每天有超过一半的世界人口使用一个或多个社交媒体平台，各行各业的企业都注意到社交媒体数据挖掘的重要性。数据挖掘与几乎所有领域都有交集，在各领域利用蓬勃发展的大数据来解释过去和预测未来不仅是数据挖掘的优点，更是未来研究者的挑战。

深度学习的提出，在数据挖掘领域中是一个重大突破。得益于数据、计算资源、算法这三大法宝，深度学习改变着整个领域。但是，深度学习这个神秘的黑盒也饱受诟病，对于图片分类、围棋等领域尚可；但是当涉及到金融、医疗、无人驾驶等领域时，人们更加需要一个透明的、可信赖的、可解释的数据挖掘模型。

随着数据挖掘的进步，尽管在各领域都取得了不错的进展，但是现有模型缺乏泛化

能力。具体来说，人类在解决某个问题或执行某项任务时，会结合全局知识和以往经验，而非局限于当前的领域知识。然而面对不断涌现的新领域、新问题，现有的大部分数据挖掘研究都是在针对具体的领域任务专门设计。由于不能真正的从数据中学到、积累知识；算法或模型的设计和计算成本过高等原因，数据挖掘通用性、泛化性研究势在必行。

8.7.表示学习

为克服传统符号表示的局限性，表示学习(Bengio et al. 2013)旨在将对象编码成低维连续向量以表征其语义信息。每个对象对应的向量称为其表示(representation)或嵌入(embedding)，而对象间的语义关系通常由对应的表示间的函数（如内积、余弦相似度等）来刻画。和传统的特征工程技术由人工设计对象特征相比，表示学习一般会随机初始化对象的向量表示，并通过优化训练目标来自动学习参数。训练完成后，对象表示即可作为其特征，用于下游任务的预测。除了可以省去特征工程的人工成本之外，表示学习还可以借助深度学习等技术，从数据中捕捉到更加高阶的特征信息。

8.7.1. 近年发展和主流方法

在社会媒体处理领域，表示学习技术常用于两类对象：文本数据（如博客文本等）和结构数据（如社交网络等）。

8.7.1.1. 面向文本数据的表示学习

文本数据一般可看作字词的序列。依表示学习对象的不同，相关技术可大致分为词/字级别与句子/篇章级别的表示学习。

字/词表示学习。谷歌公司于 2013 年提出的 word2vec(Mikolov et al, 2013)模型是最流行的词向量训练工具之一。Word2vec 基于浅层的神经网络构建模型，其优化目标为一定窗口内共现的单词间的相互预测。除作为复杂神经网络模型的底层输入外，词向量本身也广泛用于社会计算研究。(Garg et al, 2018)和(Caliskan et al, 2017)计算了性别相关和职业相关的单词之间的平均向量距离，并验证了职业中的性别偏见的存在。(Sivak and Smirnov, 2019)发现人们在社交媒体中会更多提到和“儿子(son)”词向量相似的单词，而更少提到和“女儿(daughter)”词向量相似的，从而揭示了性别不平等可能在人生早期就出现

了。

句子/篇章表示学习。2017 年之前，基于 word2vec 改进实现的 Paragraph Vector(Le et al, 2014)、基于卷积神经网络(Kalchbrenner et al, 2014)或循环神经网络(Cho et al, 2014)构建的短语/句子/篇章表示学习模型先后被提出。2017 年，(Vaswani et al, 2017)提出了基于多头注意力机制的 Transformer 模型，不仅可以充分地刻画远距离单词间的依赖关系，而且计算方式上易于并行，近年来已被验证优于以往模型，广泛应用于文本编码表示。基于 Transformer 模型扩展的大规模预训练语言模型 BERT(Devlin et al, 2019)、GPT(Brown et al, 2020)等也已成为整个自然语言处理领域的经典编码模型。上述技术也广泛用于社会计算领域的研究中，例如(Mooijman et al, 2018)使用 LSTM 预测 Twitter 中的帖子是否涉及道德话题；(Sheshadri and Singh, 2019)使用 Paragraph Vector 编码新闻，并计算新闻表示间的余弦相似度来研究新闻相似性和公众注意力的关系。

8.7.1.2. 面向结构数据的表示学习

在社会媒体处理领域，大多数结构数据表示学习算法旨在为网络/图中的每个节点学习向量表示。以社交网络为例，表示学习技术可以学习网络中的节点（即用户）的向量表示，使得表示相近的用户其好友关系/偏好兴趣也相似，进而将学得的表示用于好友推荐、用户画像等实际场景。依算法范式的不同，相关技术可大致分为基于浅层神经网络和基于图神经网络的表示学习。

基于浅层神经网络的结构数据表示学习。受 word2vec 的启发，DeepWalk (Perozzi et al. 2014)将词/句子与节点/随机游走进行类比，并直接采用 word2vec 算法进行节点嵌入学习。DeepWalk 通过等概率随机选择下一个节点来生成随机游走序列，2016 年出现的 Node2vec(Grover and Leskovec, 2016)提出了一种邻域采样策略来产生随机游走序列，它够平滑地在宽度优先搜索(BFS)和深度优先搜索(DFS)之间进行插值。LINE(Tang et al, 2015)将节点与其一阶和二阶邻居之间的相似性参数化，然后用其学习网络的节点表示。上述技术被广泛应用于社会计算领域，比如 2017 年微信广告的建模策略中使用了基于 Node2vec 的 Looklike 算法来进行高效的朋友圈广告投放；2018 年阿里巴巴(Wang et al. 2018)提出了 EGES，该算法通过在 DeepWalk 的基础上引入补充信息来解决推荐系统的冷启动问题。

基于图神经网络的结构数据表示学习。图神经网络(GNN)是一种应用于图结构的深

度学习模型。在图神经网络的每一层中，每个节点都会通过消息传递机制，聚合其邻居和自身在上一层的表示来进行更新。作为最具影响力的 GNN 模型之一，图卷积网络(GCN)(Kipf and Welling, 2017)通过节点特征的分层传播对图结构数据进行半监督学习。后来，图注意力网络(GAT)(Veličković et al, 2018)进一步利用注意力机制对邻居特征进行聚合。最近，基于图神经网络的方法被广泛应用于社区发现、谣言检测等社会计算任务：例如 2019 年，(Shchur and Gunnemann, 2019)将伯努利-泊松概率模型整合到 GCN 中，用于重叠社区发现问题；2020 年，(Bian et al, 2020)提出了一种双向图卷积网络来解决社交媒体上的谣言检测问题。

8.7.2. 未来趋势和挑战

结合表示学习技术及其在社会计算领域应用的研究现状，本文概括出以下三点值得关注的研究方向：

图神经网络在社会计算领域扮演重要角色。图神经网络在建模拓扑结构特征的同时，也能将非结构化信息（如文本信息等）作为图中节点/边的特征，在消息传递更新表示的过程中加以利用，从而以其优越的表示学习能力联合表示结构化信息和非结构化信息。目前，图神经网络在学术界和工业界都得到了非常广泛的应用。在未来一段时间，基于图神经网络的方法在面向社会媒体处理的表示学习领域仍会扮演重要角色。

大规模数据的表示学习与预训练。随着社会媒体领域的发展，数据的体量也越来越大，而如何处理并利用大规模的数据逐渐成为了一个备受关注的问题。大规模预训练模型可以充分利用大量无标注数据，并通过微调来适应特定任务。以文本数据预训练为例，GPT-3(Brown et al, 2020)已经达到了 1750 亿参数量的惊人训练规模，并可用于众多社会计算场景中的文本信息处理。结合社会媒体用户行为等大规模数据的表示学习与预训练，也是社会计算领域的重要研究方向。

结合知识、因果与逻辑的表示学习。现有的基于深层神经网络的表示学习方法虽然具有强大的表达能力，但因为缺乏可解释性往往难以用于涉及安全、隐私等问题的社会计算应用中。现实生活中的数据往往之间会存在复杂的关系。例如在知识图谱中，实体与实体之间的关系是复杂的，而如何处理这种多关系数据也是十分重要的。因此，如何结合知识图谱、因果推断、逻辑规则等人类先验知识进行表示学习，也是社会计算领域值得探索的重要研究方向。事实上，目前已有相关工作展开初步探索：例如在推荐系统

领域，NCR(Chen et al. 2021)在表示学习基础上进一步结合逻辑规则，DecRS(Wang et al, 2021)构建因果图来分析推荐模型中的偏差等。

8.8.智能教育

智能教育是基于深度学习、大数据、虚拟现实等新一代信息技术，构建以学习者为中心，贯穿“备课 - 教学 - 联系 - 考试- 评价 - 管理”教育流程各环节的智能化教育环境，实现人才培养更加多元化、更加精准、更加个性化的新型教育模式。智能教育两个主要特征，一是教育流程智能化，具体表现为人工智能技术在教学、考试、评价、管理等环节达到全方位、立体化的智能应用；二是人才培养个性化，即通过人工智能技术建立以学习者为中心、覆盖教学全流程的智能化教育环境，实现更加多元化、更加精准的智能导学与评价，促进人的个性化和可持续发展。人工智能的教育应用是将人工智能技术融入教育核心业务与场景，促进关键业务流程的自动化与关键教育场景的智能化，从而大幅提高教育工作者和学习者的效率，创新教育教学生态。（科技部新一代人工智能发展研究中心 and 罗兰贝格管理咨询公司, 2019）。人工智能赋能教育已成为未来教育变革的重要趋势。通过人工智能技术重塑教育生态，发展智能教育已经成为当前甚至是未来的教育新议题。2019 年 3 月，联合国教科文组织发布《教育中的人工智能：可持续发展的挑战和机遇》的报告，提出了人工智能教育发展的愿景、目标、途径、挑战等。人工智能教育发展的愿景是促进人工智能教育的可持续发展；人工智能赋能教育的目标是改善学习和促进教育公平；为人工智能教育时代做好准备的两个途径是：构建面向数字化和人工智能赋能世界的课程，通过后期教育和培训增强人工智能能力（任友群 and 万昆, 2019）。报告提出人工智能教育发展的六个挑战：提升制定全面的人工智能公关政策的能力；确保教育中人工智能的全纳和公平；教师与人工智能驱动的教育的双向准备；开放、高质量和包容性强的教育数据系统构建；人工智能教育相关研究的重要价值发挥；数据收集、使用和传播伦理的关注。报告对我国发展人工智能教育有如下启示：坚持立德树人，培养具备人工智能思维的中国公民；消除数字鸿沟，警惕人工智能的马太效应；构建公平而有质量的人工智能教育生态系统；以跨学科融合促进高校知识生产模式转变；新技术赋能人工智能教育；重视人工智能教育发展的伦理问题。

8.8.1. 近年发展和主流方法

当前，多项人工智能技术正逐步在教育领域开展应用，主要包括计算机视觉与机器学习、知识图谱、自然语言处理、机器人与智能控制等。

8.8.1.1. 计算机视觉与机器学习

计算机视觉是一门研究如何使机器“看”的科学，用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉，并进一步做图形处理，使电脑处理成为更适合人眼观察或传送给仪器检测的图像。机器学习技术是指机器通过对客观世界的观察获得经验，再利用经验改善自身性能的过程。如果模型是基于多层人工神经网络构建的，这一类监督式学习通常被称为深度学习。计算机视觉与机器学习在教育中已有较为广泛的应用。基于所采集的学生多维度数据，学校和教师可以对学生的学业成绩做出预测，对其可能的学习障碍和困难进行分析，对其退学（尤其在慕课学习环境中）的风险进行预警等（卢宇 and 马安瑶, 2021）。

8.8.1.2. 知识图谱

知识图谱是基于图的一种结构化的知识表示方式，本质上是一种大规模语义网络，包含较大数量的实体以及实体之间的多种语义关系。近年来，教育知识图谱的构建逐渐活跃，尤其是相继建立了针对慕课平台上的课程类知识图谱以及针对中小学学科类的知识图谱。基于所构建的教育知识图谱，智能化教育系统可以自动解答学生所提出的学科知识类的问题。另外，基于教育知识图谱，系统还可以进行相关教学资源与课程的个性化、精准化推荐（郑庆华 and 董博, 2019）。

8.8.1.3. 自然语言处理

自然语言处理技术主要用来实现人与智能机器之间通过自然语言进行有效交互。当前，自然语言处理技术在教育中也有诸多应用。例如，短文自动评分系统已经在考试中使用多年，并被不断改进以接近人类的评分水平。口语自动测评系统也已经开始广泛应用于中考等关键性考试，并已被嵌入各类语言学习软件中（严晓梅 and 高博俊, 2019）。

8.8.1.4. 机器人与智能控制

机器人作为人机共生的主要载体之一，涵盖了智能感知与推理、规划与决策、控制与交互等（张学军 and 董晓辉, 2020）。教育领域的机器人可以简单分为教育服务类机器人与教学用途类机器人。教育服务类机器人通常作为不可拆分的软硬件整体，直接服务于教学过程，完成特定的教学任务，如通过与学生的互动完成知识传授或情感陪伴。教学用途类机器人则通常由可拆分组合的硬件以及可编程的软件组成，作为机器人教育的载体或 STEM、创客课程的教学辅助工具。教学全流程可分为“备课 - 教学 - 联系-考试 - 评价 - 管理”六大场景，其中“备课”场景为开端，“管理”场景为末端，“教学”场景为中心，“练习”“考试”“评价”三大场景为支撑。随着人工智能技术的发展以及全域数据的积累

8.8.1.5. 大数据分析分析与挖掘

大数据、人工智能等技术的发展使得基于数据的多元教学评价，尤其是过程性评价成为可能（杨宗凯 and 吴砥, 2019）。新型的教学评价更注重学习者的学习过程与学习行为。利用智能技术，开发智能教育助理，建立智能、快速、全面的教育分析系统，实现专业教学活动全过程的评估监测与管理，覆盖教师和学生两个主体，以及课前、课中与课后三个环节；对教师的课前备课、课中教学、课后反思和学生的课前预习、课中学习、课后复习进行实时监测；利用基于数据驱动的过程化评价，对学生的线上、线下学习过程监督；通过智能分析与诊断，实时干预学生学习行为从而提升学习效果；提供在线评教功能让学生对教师的教学进行评价。

8.8.2. 未来趋势和挑战

随着人工智能理论和技术的不断发展和完善，人工智能在教育领域应用的广度和深度将大幅拓展。基于人工智能技术适配性和成熟度两个维度的表现，可以将智能教育场景分为三类（科技部新一代人工智能发展研究中心 and 罗兰贝格管理咨询公司, 2019）。

（1）当前 AI 主要用武之地场景

这类场景具备较高的人工智能技术适配性和成熟度，以基于智能识别技术的教学辅助类场景为主。

（2）AI 应用空间有限场景

这类场景在人工智能技术适配性和成熟度上均较低，主要集中在核心教育流程的“评”和“管”环节。

（3）AI 潜在提升发力场景

这类场景为适合通过人工智能革新、但受制于技术成熟度的场景，具有较大的潜在价值。这类场景主要集中在判断和推荐领域，普遍为服务属性。从技术层面来看，该类场景当前挑战来源于知识图谱构建、多元大数据的结构化处理等方面，一旦实现技术突破该类场景将很快实现落地。

教育人工智能是一个跨学科、跨部门、跨体系的新兴探索领域。教育人工智能的推进策略，需采取双向赋能策略，兼顾至上而下与至下而上相结合，兼顾教育主体与企业政府相辅相成，兼顾前瞻研发与实际应用顾全平衡（杨晓哲 and 任友群, 2021）。在中短期内，预计将基于大数据的智能综合分析与诊断能力，覆盖“备课- 教学 - 联系 - 考试 - 评价- 管理”核心流程下更多场景、接入教学核心环节。智能教育随着认知技术的成熟，掌握学生的能力、实现人性化交互，促进大规模“因材施教”的达成。

8.9. 智能金融

当前的金融界正面临“手工业”升级“大工业”的拐点，处于从大数据化转向自动化的升级的重要进程中，产生了一系列新思路、新工具、新模式和新业态，继而形成了“智能金融”。智能金融涵盖量化投资、智能风控、金融知识计算等方面的研究，部分技术已获得了广泛的应用。

8.9.1. 近三年进展和主流方法

（1）量化计算

量化计算旨在运用人工智能展开量化投资，主要包括公司基本面分析、资产定价和资产配置三个领域(陈梦根, 2013)(赵大伟 and 李文华, 2020)。其目标是结合金融投资领域知识，设计智能算法，更好更快地分析公司基本面，更加准确地对资产进行定价，获得更优的资金分配，辅助投资者进行资产的管理(薛亮 et al., 2018)。目前该方向的进展包括：1）设计开发上市公司财务欺诈预警算法与系统，设计财务欺诈评价体系，同时在中国和美国市场进行验证。2）研究基于机器学习与公司基本面和技术面特征的资产收

益预测方法(李斌 et al, 2019)。3）基于传统经济学模型和深度学习方法，研究结合宏观市场及经济信息的序列化资产价格预测模型(凌立文 and 张大斌, 2019)(Thakkar A and Chaudhari K, 2021)。4）研究和实现基于在线机器学习技术的投资组合选择系统、完整的量化投资回溯测试系统，能够实现较高的投资收益(Pan W et al, 2020)。

8.9.2. 智能风控

随着人工智能、大数据等技术在银行业的应用，拥抱智能风控已成为行业共识，传统风控正在向智能风控转型。智能风控可以通过大数据等获取到更多维度的外部数据，从更多层面刻画客户风险视图，与业务数据形成联动，降低定价风险、违约风险等。智能风控可以覆盖包括贷前、贷中、贷后三个阶段的信贷业务全流程，依托智能风控技术与传统风控模型互补，可以对客户风险进行更为及时有效的识别、预警和标识，同时实现全链条自动化、智能化(肖馨 et al, 2019)(杜浩云 and 张红波, 2021)。主要方法分为三类：1）贷前风控，它是整个信贷流程的基础，技术创新主要集中在反欺诈和征信两大环节。目前已经很多金融机构选择和上游数据供应商或第三方智能反欺诈机构合作，通过金融机构内部数据和第三方数据整合，基于高纬度变量和丰富应用场景，构建反欺诈模型，同时利用大数据、机器学习等技术动态化反欺诈规则，提高欺诈案件识别率，实现数据和技术的互补。以央行征信为代表，传统征信机构主要采集、加工和使用线下渠道数据为主进行信息共享，以便授信机构掌握贷款申请人的历史贷款申请、批准、使用和归还情况；随着大数据的发展，征信数据所包含的领域和来源领域越来越广，大量个人征信数可被采集，与传统个人征信数据互补，有效提升了数据的多元性和可获性，满足了网络借贷的个人征信需求。2）贷中风险，能够实现对在线交易进行仿冒和欺诈识别，对借款人进行实时管控，有效防范和控制欺诈交易等贷中风险威胁。创新主要集中在信用评分和交易反欺诈两大环节。主要通过机器学习技术构建针对业务信息中的欺诈特征与风险的自动化识别与评估，为不同场景提供反欺诈模型。还可通过关联学习、图谱学习等，实现生成式模型自动检测异常风险，开展风险评估。3）贷后风控，确保贷款安全，案件防控和业务管理质量等。利用机器学习处理多维弱变量数据，可以精准估计违约风险，制定风险管理策略、风险偏好、风险限额和风险管理政策和程序，通过自动监控策略执行情况及时优化调整，提升业务端风险管理体系的有效性，打造信贷风控闭环。

（3）金融知识计算

金融知识计算是利用大数据与人工智能技术，构建智能产业链知识图谱，深度洞悉新兴产业发展趋势，挖掘适合本区域主导产业招商的目标企业与人才，寻找招商路径精准触达，同时，实现招商引资项目从招前评估、项目落地以及招后监管的全流程管理，创新招商引资模式，提升招商引资效率(陶睿 et al., 2019)(吕华揆 et al., 2020)。当前研究热点主要包括：1）利用宏观经济数据、产业经济数据、微观经济数据与周边区域或相似区域对比，了解本区域经济发展态势。定制产业地图，展示本区域的营商环境、交通网络、产业布局、载体信息等。2）开展产业智能分析、招商雷达定位、企业尽调评估、招商协同管理、智能营销推广、产业智库顾问、风险预警监控、招商可视化等。3）基于大数据的企业投资意愿分析系统，代替传统人工分析，从企业扩张能力、投资关系、产业匹配等多个强因素，综合评价企业投资意愿，并对企业群进行打分排序，精准推荐优质潜在招商企业。

8.9.3. 未来发展趋势和挑战

8.9.3.1. 量化计算

当前，量化计算采用的评价及优化目标方法虽然能够保证模型在单个资产上的预测性能，却无法评价预测的多个资产间的相关关系，导致后续产品表现有限；对资产特征与资产价格间的关系建模还无法捕捉深层次的特征关系和时序上的结构信息；资产间、市场以及宏观经济因素对资产价格的影响和预测还未真正取得很好的成果。如何结合宏观市场、经济信息、资产价格特征，来构建预测模型，实现优秀的量化投资模型，并应用于实际基金产品，是个挑战(Pan W et al, 2020)。

8.9.3.2. 智能风控

在数字经济时代，智能风控能力逐渐成为影响金融机构可持续发展的重要因素。有效应对金融服务新形态，传统金融机构发力智能风控建设势在必行。未来，各家金融机构积极转型，引入人工智能和大数据等新技术，整合内外部资源，搭建全新的智能风控系统，打通前、中、后等各环节，不断实现全风险智能管控，全面助力风险管理水准提升，将势在必行(季成 and 叶军, 2020)。

8.9.3.3. 金融知识计算

随着数字经济的繁荣兴起，企业数据成倍递增，其中也包含了企业全息画像、企业知识图谱、产业发展指数等数据库，基于这些产业数据支撑，可以帮助决策部门统筹、洞察整个地区各个主导产业在产业规模、企业数量、投融资规模、技术专利情况、人才团队规模等多个维度的数据变化和趋势演进，也帮助决策部门实时、动态监测本地区产业发展现状和问题，掌握产业未来发展大势，进而制定更加科学、有效的产业发展扶持政策(胡慧芳 and 郑芬芳, 2020)。挖掘和应用企业和产业数据，构建图谱和产业大脑，已逐渐成为产业结构调整和企业发展规划的重要参考，这为我们科学化地制定区域产业政策提供了全新的视角。

8.10. 计算历史学

计算历史学是基于自然语言处理、知识图谱、事理图谱、网络分析、图像标记、时空地理分析、多语文本对读分析以及基础设施平台建设等方面技术，对历史乃至人文进行深度研究的新兴交叉领域。

计算历史学有两个主要特征：历史数字化与数字历史化。前者希望将过去难以计算的历史能够通过各种处理结构性数据与非结构性数据的数字计算进行量化转译，完成历史数字化工作。后者希望在引入各种数字计算工具之际，也能从历史人文角度去再赋义数字工具的人文温度，亦即从历史人文角度对数字计算方法附加上历史人文性的思考与量化计算权重，而非仅是拿来主义式的直接使用各种数字计算工具，历史人文学者在使用过程中也会对数字计算方法提出人文向度的算法迭代建议。

计算历史学的应用场景，除能辅助历史人文研究工作者进行历史研究外，还希望更进一步，通过历史数字化与数字历史化的计算历史学交叉发展下，培养出一批能量化视读历史的人文工作者，为建设数字中国贡献一份力量。

2016 年 5 月 17 日，习近平总书记在哲学社会科学工作座谈会上指出，“一个国家的发展水平，既取决于自然科学发展水平，也取决于哲学社会科学发展水平。一个没有发达的自然科学的国家不可能走在世界前列，一个没有繁荣的哲学社会科学的国家也不可能走在世界前列”。我国高等教育要培养时代所需的高素质创新人才，不仅需要建设“新工科”，也需要建设“新文科”。教育部在 2019 年 4 月 4 日发出的《关于实施一流本科专

业建设“双万计划”的通知》中也提出，“推动新工科、新医科、新农科、新文科建设，做强一流本科、建设一流专业、培养一流人才，全面振兴本科教育，提高高校人才培养能力，实现高等教育内涵式发展”。

计算历史学应对国家大力提倡新文科建设这一大背景，此一交叉领域的目标，是想通过具有处理长时段、复杂性材料优势的数字计算方法，推动对历史人文进行连续性以及宏观结构式的探索与研究。在计算历史学研究领域趋势部分，目前可见计算历史学研究向度既深且广，包含古今文学研究、古今思想研究、近代报刊语言研究、古今图像学研究、古今人物网络研究、古代时空地理研究、东亚区域史研究等，可见计算历史学发展包含一切古今人事时地物的领域研究趋势。

基于上述研究趋势，计算历史学的未来目标是，希望将计算历史学的研究产出连结到大众历史的历史公众教育、图博档等机构的文化创意工程等。通过计算可视化数据，让大众更容易接触历史与了解历史，且更进一步可响应十四五规划中的建设数字中国、打造具有国际竞争力的数字产业集群、加强国家重大文化设施与文化项目建设、协同文化创意旅游业打造数字文旅，以及有效加强党史、新中国史、改革开放史、社会主义发展史教育工作的号召。以上这些重要规划方向若能使用计算历史学研究成果作为基础，当能获得巨大有效的推动力量。

8.10.1. 近年发展和主流发展

目前，计算历史学交叉学科使用到的数字计算技术主要包括：可视觉技术、机器学习、知识图谱、自然语言处理；而所处理的历史人文议题从时间向度上包含从古代到现当代，从议题上包含从文学、历史、思想到艺术。以下依据计算历史学论坛过去四届邀集的全球计算历史学知名学者在「历史数字化」以及「数字历史化」两方面作为目前进展的说明。

历史数字化方面，主要是通过计算技术对长时段或复杂历史进行数字深描的过程。其进展主要包含三大方面：

其一，通过自然语言处理技术，对中国古代乃至近代巨量历史文本进行词汇、概念、话语等多向度的挖掘，借以探讨中国古今文学历史等思想转型的宏观轨迹，如南京大学学衡研究院暨历史学院的邱伟云老师报告的“词汇、概念、话语：基于文本挖掘技术的中国近代思想史研究” 第一届论坛（金观涛等，2016），以及南京大学艺术学院陈静老师报

告“基于《中国思想家评传》的思想家地图及思想谱系研究探索”（第二届论坛）；清华大学人文学院严程老师报告“基于数字人文方法的古典文献再发现”（第三届论坛）。

其二，通过网络分析技术，对中国古今人物进行人物网络的撷取，藉以发现网络对于古今历史事件发展中的重要推力，如北京大学中国古代史研究中心胡斌老师报告“中国历代人物传记资料库（CBDB）的建设与使用”（第一届论坛）（包弼德，2017）。

其三，通过图像语义识别技术，对古今图像如中国近代商业报纸广告图像以及中国传统山水画等进行人机互动的语义识别工作，借以从大量的图像材料中自动过滤出具有意义的图像范式，使图像史研究者得以在此基础上直接进行图像学的历史研究。如南京大学艺术学院陈静老师报告“Advertising Chinese Modern Society: graphesis, concept modeling, historical method”（第一届论坛）（陈静，2017）、中国美术学院中国思想史与书画研究中心王平老师报告“从山水到风景——中国山水画“主题”的数字人文研究“（第二届论坛）（王平等，2018）。

其四，建立综合式的计算历史学研究大型平台，将自然语言处理、网络分析、知识图谱以及时空地理分析等技术整并，有效推进大型的计算历史学研究工作，如德国马克斯普朗克科学史研究所陈诗沛老师报告“中国古方志门类分析”（第三届论坛）（Shih-Pei Chen, 2016）、山东大学东北亚学院苗威老师报告“东亚历史研究的数字人文空间”（第三届论坛）、法鼓文理学院洪振洲老师报告“中国佛教寺庙志数位典藏的建置与数位分析应用”（第四届论坛）。

而在数字历史化方面，主要是各种前沿数字计算技术的报告，而计算历史学中的数字技术报告与众不同的是，特别强调历史人文学者在运用了前沿的各种数字计算技术进行历史人文研究后，通过比对过去历史人文研究经典成果，回过头再对各种数字计算技术提出历史人文向度的计算权重建议，以调整原有的数字计算技术算法，使得算法能更加具备历史性与人文性。在此方面进展主要包括：清华大学计算机科学与技术系刘知远老师报告“自然语言处理在计算社会科学中的应用”（第一届论坛）、中科院自动化所模式识别国家重点实验室何世柱老师报告“从知识表示发展历史理解知识图谱”（第二届论坛）、南京大学历史学院王涛老师报告“NLP 技术之于文本对读的实践与发现”（第二届论坛）、哈尔滨工业大学丁效老师报告“基于事理图谱的历史事件演绎与反绎”（第三届论坛）、北京大学信息管理系王军老师报告“宋明理学研究的传承与展望—一项基于社会网络及文献计量的实验” （第四届论坛）（杨海慈、王军, 2019）、华东师范大学信息管理系许鑫老师报告“老子思想研究—来自华东师范大学的实践”（第四届论坛）、北京语言大学

汉语国际教育研究院饶高琦老师报告“近现代报纸里的语言演变和历史信息挖掘”（第四届论坛）（饶高琦, 2016; 饶高琦、李宇明, 2019, 孙琦鑫、饶高琦、荀恩东, 2020）。

8.10.2. 未来趋势和挑战

从上述计算历史学近四年来的进展图景中可见，计算历史学在未来领域关键技术发展部分，主要围绕在自然语言处理、知识图谱、事理图谱、网络分析、图像标记、时空地理分析、多语文本对读分析以及基础设施平台建设等方面进行深度研究。在领域趋势发展部分，则可见计算历史学研究向度既深且广，包含古今文学研究、古今思想研究、近代报刊语言研究、古今图像学研究、古今人物网络研究、古代时空地理研究、东亚区域史研究等，可见计算历史学发展包含一切古今人事时地物的领域研究趋势。除上述发展趋势外，近年来历史学界所兴起的全球史视野下的情感史研究、记忆史研究、多种语言文本语料概念对比分析研究，则有赖能处理中国古代到近代语言的情感分析技术与多语语义分析技术的长足发展。

历史人文学界目前正全力进行从古至今各种文字与图像史料的数字化工作，例如传统古代方面的中国基本古籍库，近当代方面则有香港中文大学建置具有一亿两千万字的中国近代思想史专业数据库（1830-1930），以及收录了期刊 520 余种，文章 53 万余篇的晚清期刊全文数据库（1833-1911)，收录了期刊 25,000 余种，文章 1000 余万篇的民国时期期刊数据库（1911-1949)，乃至于由中国社会科学院近代史研究所主持的“抗日战争与近代中日关系文献数据平台”，于 2020 年 9 月 1 日平台已上线报纸 1046 种、期刊 2343 种、图书 71071 册，以上数据平台若能配合前沿的 OCR 技术，有望在未来取得巨 s 量的文字文本；另如方正当代报纸库（1949-)收录报纸超过 1.4 亿篇，同步出版 500 种报纸，收录已停刊报纸 300 种，业已是可全文检索的数据库。

未来计算历史学在上述巨量的中国从传统到近当代图文史料基础上，将有赖于计算机学界各种前沿技术对文字与图像材料进行数字化、数据化、智能化等数字计算工作。而历史人文学界也将从人文性角度对各种前沿数字计算算法提供具有人文温度的调参建议。期待之后在计算机学界与历史人文学界的通力合作下，能基于数字化、数据化、智能化的数字发展以及有效加强党史、新中国史、改革开放史、社会主义发展史教育工作等十四五国家战略规划下，一同为建立数字中国而努力。

8.11. 智慧司法学

智慧司法旨在通过人工智能、大数据和云计算等技术实现司法业务的网络规范化、信息公开化和智能自动化等需求，推进司法体制综合配套改革。通过大数据和人工智能技术手段，为司法机关提供安全可靠的技术支持，提升司法效率和促进司法公正。 2016年中共中央、国务院印发《国家信息化发展战略纲要》，明确全方位实施“科技强检”，进一步推进检察工作现代化，明确建设各级“智慧法院”，提高案件办理各环节的信息化水平。随后，最高法、最高检和司法部等相继宣布了各自领域的智慧司法规划，科技部部署了国家重点研发计划智慧司法部分专题，极大地促进了智慧司法在我国的研究和应用，各个省市也在各自的区域内开展智慧司法的应用，扩充了智慧司法的普及程度和范围。

8.11.1. 近年发展和主流方法

8.11.1.1.判决预测

早期的基于深度学习方法的法律判决预测研究通常对几个子任务独立建模或者联合建模独立预测。随后有一些研究者提出，在法律判决预测的几个子任务之间存在依赖关系，前导任务的结果有助于后续的预测任务，对子任务进行联合预测。Zhong 等人（Zhong et. 2020）提出，法律判决预测的三个子任务之间都存在着依赖关系，他们提出了一个拓扑多任务学习模型 TOP-JUDGE，按照有向无环图的拓扑逻辑顺序预测每个子任务的输出，从而实现了对三个子任务的联合预测。Yang 等人（Yang et. 2019）对这一工作进行了改进，他们提出了一个多任务学习模型 MPBFN-WCA,将三个子任务之间的影响从单向改为了双向，最后把所有的前向信息和反馈信息结合起来，获得最终的预测结果。Chen 等人（Chen et. 2020）指出先前的研究仅仅预测总的刑期，但实际上被告经常被指控犯有多项罪行。针对这个不足之处，他们提出了更加切合实际需要的基于罪名的刑期预测任务(CPTP)，并构建了相关的数据集，使用深度门控网络(DGN)捕捉案情文本的细粒度特征，达到了目前在 CPTP 任务上的最佳表现。

8.11.1.2.案例检索

法律案例检索技术是司法文书检索系统的技术基础，其任务目标是对于给定的案例，

从案例库中找出与其相关的案例。在判例法国家（如英美等西方国家），法院在审理案件时依照“遵循先例”的原则，将先前法院的相似判例作为审理和裁决的法律依据，律师在开庭前需要对以往相关案例有充分的了解。在成文法国家（如中国），相似案例也是司法从业人员的重要参考。Shao 等人（Shao et. 2020）认为相比于通用文本检索任务，法律案例之间的相关性不仅仅指语义层面，还应该包括法律要素、事实描述、事件因果等多方面内容，在此基础上 Shao 等人（Shao et. 2020）提出融合法律文书结构的法律案例检索模型，该模型利用查询案例与候选案例各个段落间的交互信息计算法律案例之间的相关性。Tran 等人（Tran et. 2019）采用了一种基于摘要的类案检索模型，首先获取法律案例的摘要，然后计算查询案例与候选案例文本（含摘要）之间的词汇重叠度，同时融合了词汇特征和潜在特征，进一步提高了检索性能。

8.11.1.3.案件要素识别

案件事实认定通常是法院审理案件的基础性工作，案件事实认定直接决定了争议焦点归纳和法律法规适用。而案件要素的识别，能够对对案件事实的认定起到辅助作用。在刑法中，法律要素通常指犯罪构成要件，构成要件是成立犯罪所必须的条件，Li 和 Zhao 等（Li et. 2021）通过从法律案件的事实描述中获取主客观要素来实现法律指控的多粒度推理。Chen 等针对毒品类刑事案件提出了基于实体特征和多任务框架的三元组提取系统，有助于挖掘非结构化裁判文书中的案情要素。民法中没有对法律要素的明确定义，法律要素一般指司法文书中事实描述部分等关键信息，Li 和 Zhang 等 （Li et. 2019）将婚姻法案例中的法律要素定义为婚姻状态、子女个数等关键信息。案情要素抽取的结果可以用于案情摘要、可解释性的类案推送以及相关知识推荐等司法领域的实际业务需求中。

8.11.1.4.司法问答系统

Quaresma 等人借助计算语言学理论进行语义分析，采用本体论和逻辑推理进行语义和语用解释回答问题，开发了一个葡萄牙语司法问答系统。Monroy 等人利用图结构搭建了西班牙语法律问答系统，系统根据问题使用 TF-IDF 抽取一组法律条款作为答案。为了达到更好的性能，Taniguchi 和 Kano 利用概念解释，Tran 等人将相关文件形式化为图形，以帮助推理。近年来，深度学习被广泛应用在法律问答系统中，如 Morimoto 等

人将注意力机制（Attention）、Gain 等人将 BERT 应用在法律问答系统中。Huang 等人将法律领域知识图谱作为知识库辅助法律，取得了很好的问答效果。Zhong 等人提出了从中国国家司法考试中收集的法律领域问答数据集 JEC-QA。该数据集需要强大的逻辑推理能力检索相关材料和回答问题。Duan 等人提供了一个中文司法阅读理解 (CJRC) 数据集，文件来自判决书，问题由法律专家注释。CJRC 数据集可以帮助研究人员通过阅读理解技术提取元素。

8.11.1.5.评测比赛

CAIL：中国法律智能技术评测比赛（Challenge of AI in Law，CAIL）由国家最高人民法院和中文信息学会联合举办。它面向全球学术界和工业界的研究者和开发者，旨在促进中国法律智能技术的创新发展。随着领域交叉的不断深入和司法需求的不断增加，该比赛从 2018 年设立的三个任务（罪名预测、法律条款推荐、刑期预测）发展到了 2021 年的七个任务：阅读理解、类案检索、司法考试、司法摘要、论辩理解、案情标签预测、信息抽取。

COLIEE：法律信息提取/蕴含竞赛（The Competition on Legal Information Extraction/Entailment）是由人工智能与法律国际会议（ International Conference on Artificial Intelligence and Law，ICAIL）在 2014 年举办的法律人工智能评测竞赛，目的是促进法律从业人员和研究人员对法律信息处理方法的研究和讨论。在 2018 年前，COLIEE 由两项任务组成：相似案例检索和日本律师考试文本蕴含（Recognizing Textual Entailment，RTE）。从 2018 年开始，加入了加拿大判例法的两个任务：法律案件抽取和法律案件蕴含。截至目前，COLIEE2021 已经从两个任务增加到了五个任务，分别是法律案件抽取、法律案件蕴含、民法抽取、法律文本蕴含和法律问答。

8.11.2. 未来趋势和挑战

智慧司法作为国家智慧司法战略目标推进的基础，为推动国家司法信息化建设提供了方法支持。自然语言处理技术在智慧司法信息处理研究中的实践也为其他交叉领域研究带来了新的思路。但同时伴随智慧司法在图像、文本、大数据等多领域研究的不断深入，通用领域中的技术也在司法领域中暴露出类似的弱点，如数据依赖、可解释性等问题仍亟待解决。

从技术与业务角度出发，智慧司法信息处理研究未来的研究方向包括：

（1）法律文书中的语言与日常生活中的语言文本在语言风格上的差异研究。这种差异主要表现为词汇差异和篇章结构差异。因此如何利用成熟方法（如语言模型）避免差异化带来的表征不匹配，需要从司法领域角度开展进一步的研究。

（2）判决的因果推理。在裁判文书中，法官会记录完整的审理过程和结果，包含法官做出判断的推理过程，所使用的法律依据，以及最后的裁判结果。加强对法律文书文本的推理和论辩结构的研究能够为法律人工智能模型提供良好的可解释性，因而是法律人工智能走向实际应用所必须要面对的挑战。

（3）法律知识引入。法律知识对于单纯依靠数据驱动的自然语言理解模型有着至关重要的作用，不但可以提高模型性能，还可以指导模型避免数据自身带来的偏差。这里所指的法律知识包括但不限于法条、推理模式、指导案例。

总之，智慧司法现阶段已经取得了一定成果，但仍然存在不少应用需求亟待满足。对相关技术的进一步研究，需要司法从业人员与相关技术工作者的共同努力。

8.12. 社交机器人

社交机器人是一种在社交网络中自主运行社交账号并且有能力自动发送信息和链接请求的智能程序(Boshmaf et al, 2011)， 是在社交网络中扮演人的身份、拥有不同程度人格属性、且与人进行互动的虚拟 AI 形象(张洪忠等，2019)。

8.12.1. 近年发展和主流方法

8.12.1.1.传播力最大化

首先来看传播力最大化。在社交媒体上，我们希望机器人可以生成高质量文本或选取高质量文本进行传播，那么评价内容在社交媒体上是否是高质量主要在于评价其传播力的大小。那么就需要一个评价传播力的指标并使其最大化，可以结合社交媒体的特点，例如综合转发量，评论量，点赞量给出传播力的评价指标。此项技术的关键在于对社交媒体上的文本传播链进行分析和利用推荐系统相关方法。此项技术可以应用于社交媒体传播方式的建模，分析与预测，更好的理解社交媒体的传播与影响力问题，还有可以根据文本本身特征和传播特征相结合进行谣言检测。

8.12.1.2.可控交互内容生成

再来看可控交互内容生成。社交机器人的一个重要的社交属性就是与用户交互。我们希望社交机器人可以根据用户的发布内容进行高效的交互，也就是说基于用户发布的内容进行交互内容生成，同时我们希望机器人生成的交互内容是可控的。在这一点上可以更细化的分为两大技术，一个是可控评论生成，一个是辩论生成。对于可控评论生成技术，核心在对文本(如新闻文本，微博文本等)生成评论的基础上，进一步考虑嵌入外部因素(如情感等)的可控交互内容生成，将机器阅读理解和外部因素(如情感等)嵌入到 encoder-decoder 生成结构中相结合来做文本生成。(Qin et al, 2018)在 2018 年给出了腾讯新闻筛选出来的的新闻-评论对数据集，并给出了详细的评测指标。进一步可控评论生成可以从可控对话生成相关技术迁移，ECM(Zhou et al, 2018)是第一个在对话中考虑情感因素的工作。在传统的 encoder-decoder 基础上引入了三个机制:情感类别嵌入，内部记忆和外部记忆。近年来，随着预训练模型的发展，情感可控的文本生成逐渐以 GPT 等预训练模型作为基座，并取得了更强大的效果。现有研究主要关注两方面的问题。1. 如何建模情感的表达过程、让文本生成受控于指定情感。2. 如何丰富情感表达的方式和内容，以提高生成的多样性和信息量。对于辩论生成，可以看作是论辩分析与文本生成相结合的任务，同时可以引入外部知识，针对用户发布的内容，生成相应的反驳内容。其中包含通过论辩分析方法抽取出用户发布内容的论据与论点，并考虑将论据论点嵌入到传统的 encoder-decoder 文本生成结构中。(Alshomary et al, 2021)提出并验证了一个假设，即辩论句中存在可以攻击的论点，针对这些论点可以生成反驳句，同时给出了一个分类方法用来筛选可攻击论点。可控交互内容生成可以应用于在社交媒体上对舆论进行积极正向引导，以及舆论攻防中。

8.12.1.3.角色化对话生成

最后再来看角色化对话生成。在社交媒体上，我们希望这个社交机器人具有一定的设定人物背景，也就是人设，同时社交机器人可以根据自己的人设和用户进行更加丰富的互动。目前的核心技术在于将人设信息嵌入到大规模预训练对话模型上，如 GPT-2，BERT 等。(Song et al, 2021)提出了 BoB 模型，由一个基于 BERT 的编码器和两个基于 BERT 解码器构成。一个解码器用于人设一致性理解，另一个用于响应生成，在角色化生成上取得了不错的效果。角色化对话技术可以应用于社交媒体上的私信功能，基于人

设背景和用户进行更好的私信互动。

8.12.2. 未来发展和趋势

社交机器人已经可以在社交媒体上产生较广泛的影响力，这份影响力可能被别有用心者用作虚假信息传播的媒介。为了鉴别此类社交机器人，社交机器人检测任务应运而生。研究者使用多种方式鉴别真实社交媒体中可能存在的社交机器人，(Dorri et al, 2018)提出了 SocialBotHunter，它将社交图的结构信息与用户的社交行为信息统一起来，以便在类似 Twitter 的社交媒体中检测恶意的社交机器人，(Fazil et al, 2021)提出了 DeepSBD，通过双向长短期记忆（BiLSTM）和卷积神经网络（CNN）架构建模用户的行为模型，以辨别社交机器人。尽管上述研究者认为社交机器人是对社交网络的破坏，一些调查显示社交机器人的影响力也可以被应用于积极的方面，(Suárez-Serrato et al, 2018)的分析表明，机器人实际上帮助了优质信息在人类用户之间的扩散，因此一些研究者也在尝试绕过社交机器人检测。例如，使用上一节中所述的角色化对话生成技术生成的推文与人类更类似，将有可能规避社交机器人检测。在未来，围绕社交机器人检测的攻防将是一个热点和趋势。

社交机器人目前已经得到了广泛的应用，类似 Athena 2.0(Walker et al, 2021)的对话式社交机器人不仅可以应用于语音助手等场景，也在推特等社交媒体中被成功的应用(Savvopoulos et al, 2018)，一些新的社交机器人设计框架帮助社交机器人与用户就各种话题进行友好的交流(Bowden et al, 2018)，从而使得社交机器人可以更好的完成信息宣传、引导等任务。

8.13. 舆情计算

舆情计算是指研究使用自然语言处理技术，多语言文本语义理解技术、图像技术、跨平台信息追踪等技术，对来自所有互联网公开信息，如常见的资讯网站和社交媒体：新华网、腾讯新闻、百度贴吧、论坛、新浪微博、微信、博客等数据进行计算分析的一项任务。

现代社会是一个信息驱动的社会，每天都有大量的信息产生，舆情计算的第一任务是利用强大的大数据计算能力实现了互联网信息的实时收集、挖掘和智能检索，保障数据的及时性、完整性，易用性和准确性。拥有足够多的舆情数据支持，为语言、图像处

理等技术打好数据基础。舆情的计算不仅需要具备强大的数据采集和处理能力，还需要具备强大的价值挖掘能力，普通的关键词检索、敏感信息过滤等手段对舆情的分析过于片面，不能很好的提升模型、系统的泛化能力，图像处理、自然语言处理等人工智能技术则使舆情分析系统在分析方式、分析对象、分析能力等方面更加“智能”，能够适应更为复杂的互联网信息内容和传播方式。通过舆情计算不仅可以洞察观点、情绪、口碑、社情民意，为企业提供商业情报，辅助商业决策，还能为政府机构挖掘社情舆论，提升社会治理水平。

8.13.1. 近年进展和主流方法

舆情计算技术主要包括网络数据采集、舆情数据清洗、数据分析和预测等关键步骤。以下分别针对这些关键步骤进行描述和介绍。然后我们分析近三年舆情计算的专利情况。

8.13.1.1.网络数据采集

舆情计算技术中所使用的数据主要是通过网络爬虫进行搜集。目前常用的网络爬虫技术有 Nutch、Crawler4J、Heritrix 等。其中，Nutch 技术是最常用的一种工具，该方法可以定制化的完成页面检索和拓展抓取；此外该方法还支持大规模并行处理，可以将算法部署到不同的计算机集群中，进行协同处理。（马梅等，2016）基于 Nutch 开发了一种面向新浪微博的网络爬虫技术，该方法克服了官方 API 接口中无法下载大量信息的缺点，为使用微博数据进行社会舆论分析提供了极大的便利。（杨青等，2021）通过八爪鱼数据采集系统,从百度,雅虎两种搜索引擎上获得样本网站并分别获取相关数据，利用数据分析法和内容分析法总结产品结构的现状及其存在的问题。此外，基于 PageRank、Fish Search 算法的爬虫策略也常常被使用在数据收集过程中。

8.13.1.2.舆情数据清洗

舆情数据清洗是指对使用网络爬虫技术所获取的数据进行处理。经过爬虫获取的数据中含有大量的无关链接、广告等无关内容，数据清洗的目的也就是去除这些噪声数据。（王少鹏等，2014）提出了一种聚焦式的网络舆情数据清洗方法，这种方法根据需求对获取的数据聚焦，在海量数据中快速寻找所需要的目标。（冯力等，2021）采用改进 Levenberg-Marquardt 算法(L-M BP 神经网络算法)来建立数据清洗模型,首先对样本进行

预处理,对建立的异常数据进行训练和得到的结果反复验证,得到的误差控制在 3.0%以内,且模拟的网络值能真实反应的变化趋势.该模型适用对异常数据清洗数据和预测.（Suvorov R，2014）提出了一种基于排序的智能数据爬取和处理技术，该技术采用主动学习的手段对爬虫技术进行了改进，使改进后的技术在数据获取阶段对大量网络数据进行分类和预处理，该方法在音频、视频的获取和处理方面应用广泛。

8.13.1.3.数据分析

舆情数据的分析和预测，是舆情计算技术中最关键的步骤。通过设计恰当的算法对获取的数据进行分析，发掘其中的热点话题，并对其传播影响、舆情等级进行评估，采用合理的手段对舆论进行引导和管控。在舆情分析方面长常用的技术手段有贝叶斯分类器、支持向量机（SVM）、随机森林、AdaBoost、贝叶斯网络、神经网络等技术。（马梅等，2016）使用改进型孪生支持向量机对从新浪微博获取的大量数据进行分析和训练，经过实验验证，该方法适用于对中文语料进行分析；（田俊静等，2021）借鉴决策树和回归模型的思想，结合政策、互联网以及市场经济情况，构建了一种多模态数据联合分析模型，该方法在大量训练数据的支撑下，可以完成多维度社会热点信息的挖掘和提取；（Feng 等，2014）在 Single-Pass 模型的基础之上，提出了一种周期性的 Single-Pass 聚类算法模型,这种方法在话题的聚类指标上远优于原始的 Single-Pass 算法。（张乾浩等）通过 Lingo 聚类算法来分析网页中的热点话题，并设计改进了排序算法发掘热点话题之间的关联性。（聂方彦等，2017）提出了一种基于 FCM 的组建聚类算法，这种算法可以根据所分析数据自身的特点和分布状况，自适应的完成数据的分析工作；此外该算法还可以完成在线式增量学习，也就是说对于新建入的数据，算法可以不经过重新训练，只进行迁移学习即可，这一特点极大地提升了该聚类算法的泛化能力和迁移部署性能。

8.13.1.4.舆情预测

舆论预测是数据分析后的一项重要步骤，该过程主要是为舆情监控、舆情预警提供重要的参考，并且帮助制定各级政府部门、企业单位制定有关的应对措施。目前舆情预测方面的研究相对广泛，有多种方法可供使用。（吴谦等，2019）使用改进的粒子群算法和 BP 神经网络,设计了一套基于百度搜索指数的网络舆情预测模型。（张和平等,2021）深入的研究马尔科夫模型，构建了一种基于 HMM 模型的舆论预测模型，该模型会结合

网民之间的个体差异、行为特征等个性化因素，对社会舆论话题在人与人之间的传播性进行预测，此外，该算法还可以分析不同事件的传播规律。（何炎祥等，2016）使用概率关联模型，建立动态贝叶斯分析网络，这种方法可以预测网络舆论之间关联性的走势，是一种非常高效的网络舆论动态预测模型。（韩玉鑫等，2019）通过对时序的 RNN 模型进行分析，针对新浪微博数据之间的关联性找了一种最优时序模型，该模型对于早期热点事件的传播和演化具有高度的准确性。（王超等，2018）结合我国互联网自身的特点，提出了使用 ARMIR 动态时序模型同时结合人工神经网络的方法动态的预测网络舆情的发展趋势，该方法根据从历史数据角度出发，对于某些特定的问题且具有一定周期性的任务可以产生较好的预测结果，但是也存在模型训练时间长、运行效率较差、训练困难等缺点。

8.13.1.5.舆情计算相关专利

近两年关于舆情计算的专利共有 321 项，上表列举了近来部分舆情计算相关专利，可以看出，专利主要集中在上述四项技术以及完整的舆情计算平台，其中数据挖掘与数据分析占重较大。且舆情计算具有明显的领域跨度，在众多领域均有应用，有法院舆情、微博舆情乃至医院智慧服务舆情等等。另外，目前舆情计算方面的研究相对广泛，各类深度学习方法可供使用。

8.13.2. 未来趋势和挑战

舆情分析的需求几乎涵盖所有行业，无论是公司级、企业级、还是各级政府机构，即时的舆情信息对于维护、提升公司、政府形象意义非凡。舆情分析服务能为目标客户提供多维度的信息挖掘和高附加值的洞察分析，具有巨大的企业和社会价值。高精准的舆情分析对舆情计算技术提出了新的要求。舆情计算方法也面临着诸多挑战。

8.13.2.1.数据实时采集

舆情计算的目的是对决策提供支持，具有时效性、真实性的舆情资源更有利于迭代舆情计算模型。互联网舆情，本质上是对互联网公开信息的采集、分析、研判，并产生业务价值，是一个价值数据挖掘的过程，鉴于舆情计算的时效性、动态性、特殊性、真实性，与其数据获取与传统的数据挖掘又有很大差别。未来舆情计算流程更加倾向于动

态采集数据，不仅通过网络的方式进行舆情采集，保证信息来源的广度，还可通过众包等方式进行舆情信息的获取保证舆情的可靠性、真实性，再通过动态计算得出结论，整个分析过程不间断进行。

8.13.2.2.多模态舆情计算

时下舆情计算的方法主要使用 NLP 技术对文本内容进行情感分析、主题抽取、关键词提取来获取舆情信息，鉴于互联网时代信息资源的多样性、复杂性，单纯的自然语言处理技术已经不能满足舆情系统的需求。后续的舆情计算可以考虑融入多模态技术，使用图片、表情符号、视频、声音等多维度的数据信息进行建模，对于舆情系统准确性提升大有裨益。

8.13.2.3.多语种舆情计算

放眼国内外、多语种分析技术如火如荼。但是国内舆情研究机构并不算少，但关注的都是国内舆情和网络舆情，多语种、信息化和对接整个社会和国家发展战略的舆情全球化才是未来国内舆情计算的核心。多语种舆情分析能及时进行海外舆情监测、海外安全形势分析、海外品牌口碑分析，为国家综合竞争力提升助力。同时多语种舆情计算也关注例如维吾尔语、藏语、蒙古语等小语种内容，及时获取少数民族民意、民声，对于国家稳定、民族团结意义非凡。

8.13.2.4.舆情计算模型提速

舆情集计算过程主要涉及情感分析、文本分类、知识图谱、自然语言理解等技术，其中词性标注和命名实体识别能够有效的描绘出网络关键词与热词。长短句形的句法分析不仅能够在海量文本数据当中提炼出话题与意图信息，还可以计算出信息当中所表达的情绪，以实现舆情分析。此外，文本聚类与分类的准确性直接映射出互联网指数所提供信息内容与趋势分析的可靠性。伴随着文本内容复杂性、多样性的提升，开始使用深度模型对数据建模分析，但是深度模型的速度、效率在一定程度上又限制了舆情计算的发展，未来运用 NLP 技术进行舆情计算更关注模型的时效性、运算效率。

8.14. 产业发展现状及趋势

近年来，社会媒体平台已经成为互联网中最主要的一种信息媒介，衍生出多种以信息共享、信息服务为核心的网络应用，所产生的大规模社交用户生成数据蕴含了海量的数据信息。作为重要科技发展纲领，习近平总书记曾指出：“要运用大数据促进保障和改善民生。大数据在保障和改善民生方面大有作为。 要坚持以人民为中心的发展思想，推进“互联网+教育”、“互联网+医疗”、“互联网+文化”等。”

对于海量社会媒体数据的有效处理与利用具有重要意义，能够有助于解决多种业务场景以及相关社会需求。在学术界，研究人员在 ACL、EMNLP、The Web Conference、ICWSM、SIGKDD 等代表性国际会议上持续推动社会媒体处理相关的研究工作，针对社交关系预测、用户画像、兴趣建模、谣言检测等科研问题，涌现出一批以深度学习为主要解决途径的相关工作。同时，产业界更强调社会需求，紧密围绕具有相关实际需求的业务或者社会问题进行相关工作的开展，在抗击疫情、突发事件救援等方面都起到了重要作用。

根据社会媒体处理领域的现阶段情况，下面分别从技术和应用两个方面对于领域产业发展现状以及趋势进行介绍。

8.14.1. 社会媒体处理技术快速发展

随着社会媒体平台用户量和活跃度的提升，相关业务数据规模日益增长。为更好地支撑大规模、跨模态的业务应用场景，社会媒体处理技术快速发展，呈现出以下发展趋势：

8.14.1.1.模型由浅层到深层

近年来，深度学习技术在推荐系统、用户行为预测等社会计算业务场景中得到了非常广泛的应用。2016 年谷歌提出了将线性模型和深度模型相结合生成嵌入的 Wide & Deep models (Cheng et al. 2016)模型，为用户提供应用推荐；2017 年雅虎公司提出基于循环神经网络进行新闻推荐(Okura et al., 2017)；2018 年阿里巴巴提出了结合注意力机制的深度嵌入模型用于用户点击率预测(Zhou et al., 2018)。伴随着 Transformer、图神经网络技术的出现，最新的社会媒体处理技术逐渐向着层数更深、参数更多的方向发展。2019

年，阿里巴巴提出基于 Transformer 的用户行为序列建模模型用于电商推荐(Chen et al., 2019)；2021 年，华为将图神经网络应用于用户点击行为预测(Guo et al., 2021)；百度结合异质图神经网络提出多语言 POI 检索模型 HGAMN (Huang et al., 2021)，并服务于百度地图应用。

8.14.1.2.信息处理由单模态到多模态

社会计算领域中的真实业务场景通常具有多种类型的数据：如非结构化的文本、图片、视频以及结构化的社交网络、知识图谱等。因此，社会媒体处理技术也逐渐从处理单模态信息向处理多模态信息方向发展，涌现出众多和深度学习技术相结合的工业界应用。例如 2020 年美团点评提出了 MKGAT(Sun et al., 2020)，将多模态知识图谱应用到推荐系统中; 2021 年 Facebook 提出 VisRel(Borisyuk et al., 2021)，通过多模态技术结合媒体理解和文本理解技术进行多媒体搜索；阿里巴巴提出的 SEMI(Lei et al., 2021)应用多模态技术和预训练技术进行短视频推荐，并服务于淘宝平台。

8.14.1.3.新兴技术探索业务落地

学术界新兴的元学习、大模型预训练等技术也在社会媒体处理场景中得到了应用。这一类新兴技术可充分利用社会媒体用户产生的大量无监督数据或适用于小样本场景，从而在大规模人工标注成本昂贵的情况下更加适合业界应用。2021 年百度地图团队提出 MST-PAC(Fan et al., 2021)，一种基于元学习的时空个性化 POI 即时检索模型；阿里巴巴提出基于元学习和语言模型预训练的 MeLL（Wang et al., 2021a）用于用户意图检测。微软结合强化学习和预训练模型用于文本广告生成(Wang et al., 2021b)，并将提出模型应用到微软必应动态搜索广告（DSA）中。

8.14.2. 社会媒体大数据及相关处理技术赋能多领域应用

作为互联网大数据的重要组成部分，社会媒体数据近年来增量显著、用户规模不断扩大，相应的智能分析处理技术也随之不断发展，赋能了一批新兴应用。尽管社会媒体应用场景与商业化程度各不相同，但基于社会媒体数据的应用探索总体上呈现加速发展趋势，其商业化模式日趋成熟。

在智能教育领域，社交媒体借助其数据互联、传播广泛等媒介特点，对于教育领域

具有重要的促进作用。各种互联网以及社会媒体平台的线上教育工具以及资源在疫情期间发挥了重要作用，线上教育已经成为日常教育的重要途径之一。作为典型代表应用，雨课堂为教学过程提供在线的数据化、智能化信息支持，形成了新型智慧教学解决方案；社交媒体教育用户（例如微博教育用户或者教育账号）数量也呈逐年增多趋势，依靠社交媒体平台进行教育信息的传播与普及，吸引了较多的社会关注。

在智能金融领域，考虑到未来银行智能化、平台化与生态化的新趋势，多家互联网公司均构建了自己的金融服务平台，如蚂蚁集团的综合金融服务平台、科大讯飞的智慧金融讯飞开放平台，百度的金融智能获客平台等。这些平台利用人工智能、大数据与云计算等技术，融合各地金融实践建立了监管与服务系统。

在舆情计算领域，由于对于获取资源和处理技术能力的限制，中小企业难以独立进行专业的舆情应用开发。针对这一需求，多家机构开始通过构建共享开放的舆情分析平台对外提供服务，如新华智云、百度智能云的舆情服务于监控系统、北鲲舆情监控系统等。与此同时，情感分析商业化平台也在不断涌现与完善，各大机器学习平台也均提供了情感分析的 API 接口，如讯飞、腾讯云、阿里云、百度云的情感分析 API 等语义智能分析平台等，有效支撑了中小企业用户的相关技术服务。

在重大突发事件中，社交媒体平台以及相关处理技术发挥了重要的应对作用。疫情期间，健康码以及行程码通过综合分析多来源的网络平台大数据，实现了疫情的精准溯源与追踪，有效推动了抗疫工作的顺利开展。在郑州暴雨救援期间中，一份命名为《待救援人员信息》共享文档自发由民间救援组织设置和传播，为救援队的行动提供了准确资讯。社交媒体能够凝聚较为分散的关键信息，从而形成更具有针对性、时效性的信息聚合。

此外，社会媒体处理技术也同步在多个领域持续发力，包括智能司法、智能交通等，相信会有更多成熟完善的应用呈现在相关领域。经过多年的探索与研究，社会媒体处理技术已经成为相关企业的核心技术，所形成的技术处理平台有效提升了研究者对于社会媒体数据的挖掘与分析能力。这种技术与应用相交互的影响力正在进一步深化与拓展，对于我国居民生活质量的提升、国民经济的平稳发展起到了重要作用。

8.15. 总结及展望

伴随着互联网和移动通信技术的进步，社会媒体处理在为学术和产业界带来新的机

遇的同时，也带来了隐私泄露、算法偏见等隐患，呈现出“双刃剑”效应。本文首先从学术领域和产业领域介绍社会媒体处理的起源和发展，然后阐述了社会媒体处理的发展现状和关键科学问题。涉及广泛的学科、模态和领域是社会媒体处理的重要特点。社会媒体处理在不同领域的发展进一步推动了数据统一表示、学科交叉研究、产学研融合等相关问题的探索。本文接着介绍了十余个社会媒体处理领域的重要研究方向，并详细分析了近五年来的研究进展与发展趋势。最后，本文从技术和应用两方面介绍了社会媒体处理领域产业发展现状和趋势。

随着近期图神经网络、Transformer、预训练等技术的出现，社会媒体处理技术快速发展。基于图神经网络的结构化数据处理技术已经逐渐成熟并用于社会媒体处理的实际应用中，如舆情分析和推荐系统，并取得了巨大的成功；结合预训练等新兴技术利用大规模无标签数据的模型范式也将在情感计算、谣言检测等众多社会媒体处理问题上大有作为。

虽然现在的社会媒体处理技术在学术领域和产业领域都取得了成功，但由于大部分技术是基于深度模型的，虽然表达能力强大，但具有缺乏可解释性、鲁棒性的缺陷，从而导致模型安全性难以保障的问题。随着因果推断、知识图谱等对人类先验知识进行学习的技术的发展、以及数据驱动+知识驱动的人工智能新范式的提出，如何在社会媒体处理技术中有效运用人类知识，将是涉及算法可靠性的重要研究方向。

由于社会媒体处理通常具有跨学科、跨模态、跨领域以及数据多样化的特点，预计未来一段时间的社会媒体处理技术会继续向着“多模态、大规模”的趋势发展。比如在跨模态数据方面，目前无论是在学术领域还是产业领域都有很多通过结合图片、视频、文本等各种模态下数据的社会媒体处理方法，并在短视频推荐、多媒体搜索等真实场景中得到了应用。未来的社会媒体处理技术发展将有望横跨各种模态、领域及学科高速前进。

社会媒体处理专委会在中国中文信息学会的领导下，致力于研究最前沿的社会计算技术，提升我国信息技术创新能力，力争使我国信息技术朝数字化、数据化、智能化发展，运用大数据技术促进保障和改善民生，全面助力我国社会媒体处理技术提升，推进“互联网+教育”、“互联网+医疗”、“互联网+文化”等跨领域发展，将势在必行。

8.16. 参考文献

- (Eger et al., 2017) Eger S, Daxenberger J, Gurevych I. Neural end-to-end learning for computational argumentation mining. ACL 2017.
- (Gao et al., 2021) Silin Gao, Ryuichi Takanobu, Wei Peng, Qun Liu, Minlie Huang HyKnow: End-to-End Task-Oriented Dialog Modeling with Hybrid Knowledge Management. Findings of ACL 2021.
- (Gong et al., 2019) C Gong,J Yu, R Xia,Unified Feature and Instance Based Domain Adaptation for End-to-End Aspect-based Sentiment Analysis. EMNLP 2019.
- (Hua and Wang., 2019) Xinyu Hua and Lu Wang, Sentence-Level Content Planning and Style Specification for Neural Text Generation. EMNLP 2019.
- (Ji et al., 2021) Lu Ji, Zhongyu Wei, Jing Li, Qi Zhang and Xuanjing Huang, Discrete Argument Representation Learning for Interactive Argument Pair Identification. NAACL 2021.
- (Li et al., 2019) Xin Li, Lidong Bing, Piji Li, and Wai Lam.2019. A unified model for opinion target extraction and target sentiment prediction. AAAI 2019.
- (Liu et al., 2012) Bing Liu,Sentiment analysis and opinion mining Synthesis Lectures on Human Language Technologies, Morgan&Claypool Publishers, 2016.
- (Liu et al., 2021) Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, Minlie Huang, Towards Emotional Support Dialogue Systems. ACL 2021.
- (Van et al., 2002) Van Eemeren FH, Grootendorst R, Henkemans AF. Argumentation: Analysis, evaluation, presentation. Routledge; 2002.
- (Zhou et al., 2018) Hao Zhou,Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, Bing Liu, Emotional Chatting Machine:Emotional Conversation Generation with Internal and External Memory. AAAI 2018.
- 王成军，2016） 计算传播学的起源、概念与应用. 编辑学刊,3:59-64.
- （王成军，2017).计算社会科学视野下的新闻学研究：挑战与机遇. 新闻大学, 4:26-32.
- （张伦，彭泰权，王成军，梁海，祝建华，2021）.<从边陲到主流的一条自然路径：华人计算传播学者的参与和体验>。李立峰、黄煜（编），《传承与创

新：中华传播研究 40 年》（页 399-419）。香港：香港中文大学出版社.

- （周葆华，钟媛，2021）.“春天的花开秋天的风”：社交媒体、集体悼念与延展性情感空间——以李文亮微博评论（2020-2021）为例的计算传播分析[J].国际新闻界, 43(03):79-106.DOI:10.13495/j.cnki.cjjc.2021.03.005.
- （周莉,王子宇,胡珀，2018）.反腐议题中的网络情绪归因及其影响因素——基于 32 个案例微博评论的细粒度情感分析[J].新闻与传播研究, 25(12):42-56+127.
- （祝建华，彭泰权，梁海，王成军，秦洁，陈鹤鑫，2014).计算社会科学在新闻传播研究中的应用. 科研信息化技术与应用. 5 (2), 3-13.
- （Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. 2018). Moralization in social networks and the emergence of violence during protests. Nature human behaviour, 2(6), 389-396.
- （ Rains, S. A. 2020). Big data, computational social science, and health communication: A review and agenda for advancing theory. Health communication, 35(1), 26-34.
- (刘涛雄 and 徐晓飞, 2015) 刘涛雄、徐晓飞.互联网搜索行为能帮助我们预测宏观经济吗？[J].经济研究，2015（12）.
- 刘志洋、汤珂.互联网金融的风险本质与风险管理[J].探索与争鸣，2014（11）.
- 孟天广、郭凤林.大数据政治学：新信息时代的政治现象及其路径探析[J].国外理论动态，2015（1）.
- 吴江、张小劲.大数据国际政治研究的回顾与展望[J].华中师范大学学报（人文社会科学版），2016（7）.
- 喻丰、彭凯平、郑先隽.大数据背景下的心理学：中国心理学的学科体系重构及特征[J].科学通报，2015（5）.
- Ammans H M, Kendrick D A, John R. Handbook of computational economics, v1[M]. Elsevier, 1996.
- Conte R, Gilbert N, Bonelli G, et al. Manifesto of computational social science[J]. European Physical Journal Special Topics, 2012(1).
- Gary Jim. eScience-The Revolution is Starting, in Hey Tony (ed.), The Fourth Paradigm: Data-Intensive Scientific Discovery[M]. Microsoft Research, 2009.
- King, et al. Designing Social Inquiry: Scientific Inference in Qualitative

Research[M]. Princeton University, 1994.

- Lazer D, Pentland A, Adamic L, et al. Computational social science[J]. Science, 2009(5915).
- Macy, Michael W., Willer Robert. From Factors to Actors: Computational Sociology and Agent-Based Modeling[J]. Annual Review of Sociology, 2002(28).
- Marchi S D, Page S E. Computational Social Science: Discovery and Prediction [M]. Cambridge University Press, 2016.
- Trappl Robert. Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention (Advances in Group Decision and Negotiation) [M]. Netherlands: Springer, 2006.
- (Clifton et al, 2010)Clifton C., Encyclopædia Britannica: Definition of Data Mining. 2010.
- (Cortes et al, 1995)Cortes C., Vapnik V., 1995, September. Support-vector networks[J]. Machine learning (pp. 273-297).
- (Cover et al, 1967)Cover T., Hart P., 1967, January. Nearest neighbor pattern classification[J]. IEEE transactions on information theory (pp. 21-27).
- ((Soumen et al, 2006)Soumen C., Martin E., Usama F., Johannes G., Jiawei H., Shinichi M., Gregory P.S., Wei W., 2006, April. Data Mining Curriculum. ACM SIGKDD (pp. 1-10).
- (Hand et al, 2001)Hand D., Mannila H., Smyth P., 2001, August. Principles of Data Mining. MIT Press, Cambridge, MA. ISBN 0-262-08290-X (pp. 1-425).
- (Fayyad et al, 1996)Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996, March. From Data Mining to Knowledge Discovery in Databases. AI Magazine (pp. 37-54).
- (Hastie et al, 2008)Hastie T., Tibshirani R., Friedman Jerome., 2008, August. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 1-745).
- (Han et al, 2001)Han J., Pei J., Kamber M., 2011, November. Data mining: concepts and techniques[M]. Elsevier (pp. 1-26).
- (Jaseena et al, 2014)Jaseena K.U. and Julie M. David, 2014, April. Issues, challenges, and solutions: big data mining. CS & IT-CSCP (pp. 131-140).
- (Rish et al, 2001)Rish I., 2001, January. An empirical study of the naive Bayes classifier[C]//IJCAI 2001 workshop on empirical methods in artificial intelligence (pp. 41-46).
- (Bengio et al, 2013)Y. Bengio, A. Courville, and P. Vincent, Representation

learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798– 1828, 2013.

- (Bian et al, 2020) T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, & J Huang. (2020, April). Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 549-556).
- (Brown et al, 2020) T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, et al., Language models are few-shot learners, in Proc. of 34th Annual Conference on Neural Information Processing Systems, 2020.
- (Caliskan et al, 2017) A. Caliskan, J. J. Bryson, and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science , vol. 356, no. 6334, pp. 183–186, 2017.
- (Chen et al. 2021) Chen, H., Shi, S., Li, Y., & Zhang, Y. (2021). Neural Collaborative Reasoning. In Proceedings of the Web Conference 2021 (pp. 1516-1527).
- (Cho et al, 2014) K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation, in Proc. 2014 Conf. on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014, pp. 1724−1734.
- (Devlin et al, 2019)J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, inProc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 4171−4186.
- (Garg et al, 2018)N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proc. Natl. Acad. Sci. USA , vol. 115, no. 16, pp. E3635–E3644, 2018.
- (Grover and Leskovec, 2016) A. Grover and J. Leskovec, Node2Vec: Scalable feature learning for networks, in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 855-864.
- (Kalchbrenner et al, 2014) N. Kalchbrenner, E. Grefenstette, and P. Blunsom, A

convolutional neural network for modelling sentences, in Proc. 52ndAnn. Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 2014, pp. 655−665.

- (Kipf and Welling, 2017) T. N. Kipf, and M. Welling, 2017. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations.
- (Le et al, 2014) Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.
- (Mikolov et al, 2010) T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, Recurrent neural network based language model, in Proc. 11th Ann. Conf. of the Int. Speech Communication Association, Makuhari, Japan, 2010, pp. 1045−1048.
- (Mikolov et al, 2013) T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in Proc. 27th Ann. Conf. on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 2013, pp. 3111−3119.
- (Mooijman et al, 2018) M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani, Moralization in social networks and the emergence of violence during protests, Nat. Hum. Behav., vol. 2, no. 6, pp. 389–396, 2018.
- (Perozzi et al, 2014) B. Perozzi, R. Al-Rfou, and S. Skiena, DeepWalk: Online learning of social representations, in Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data.
- (Shchur and Gunnemann, 2019) O. Shchur and S. Gunnemann. Overlapping community detection with graph neural networks. In KDD Workshop DLG'19, 2019.
- (Sheshadri and Singh, 2019) K. Sheshadri and M. P. Singh, The public and legislative impact of hyperconcentrated topic news, Sci. Adv., vol. 5, no. 8, p. eaat8296, 2019.
- (Sivak and Smirnov, 2019) E. Sivak and I. Smirnov, Parents mention sons more often than daughters on social media, Proc. Natl. Acad. Sci. USA, vol. 116, no. 6, pp. 2039–2041, 2019.
- (Tang et al, 2015) J. Tang, M. Qu, M. Z. Wang, M. Zhang, J. Yan, and Q. Z. Mei, LINE: Large-scale information network embedding, in Proc. 24th Int. Conf. on

World Wide Web, Florence, Italy, 2015, pp. 1067-1077

- (Vaswani et al, 2017) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in Proc. Ann. Conf. on Neural Information Processing Systems 2017, Long Beach, CA, USA, 2017, pp. 5998−6008.
- (Veličković et al, 2018) P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, 2018. Graph Attention Networks. In International Conference on Learning Representations.
- (Wang et al, 2018)J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, & D. L. Lee, (2018, July). Billion-scale commodity embedding for e-commerce recommendation in alibaba. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 839-848).
- (Wang et al, 2021) Wang, W., Feng, F., He, X., Wang, X., & Chua, T. S. (2021). Deconfounded Recommendation for Alleviating Bias Amplification. arXiv preprint arXiv:2105.10648.
- (科技部新一代人工智能发展研究中心， 罗兰贝格管理咨询公司，2019)科技部新一代人工智能发展研究中心， 罗兰贝格管理咨询公司[M]. 智能教育创新应用发展报告. 2019.08
- （刘邦奇等，2021）刘邦奇，张金霞，许佳慧，胡婷婷，朱广袤. 智能技术赋能因材施教：技术框架、行业特点及趋势——基于智能教育行业发展实证数据的分析[J]. 电化教育研究，2021，42(02):70-77
- (卢宇等，2021)卢宇，马安瑶，陈鹏鹤. 人工智能+教育：关键技术及典型应用场景[J]. 中小学数字化教学，2021，10:5-9
- (任友群等，2019)任友群，万昆， 冯仰存. 促进人工智能教育的可持续发展——联合国《教育中的人工智能：可持续发展的挑战和机遇》解读与启示[J]. 现代远程教育研究. 2019，31(05):3-10
- (严晓梅等，2019)严晓梅，高博俊，万青青，尹霞雨. 智能技术变革教育的发展趋势——第四届中美智慧教育大会综述[J]. 中国电化教育，2019(07):31-37
- (杨晓哲和任友群，2021)杨晓哲，任友群. 教育人工智能的下一步——应用场景与推进策略[J]. 中国电化教育. 2021(01):89-95

- (杨宗凯等，2019)杨宗凯，吴砥，陈敏. 新兴技术助力教育生态重构[J].中国电化教育，2019，385(02):1-5.
- (张学军和董晓辉，2020)张学军，董晓辉. 人机共生:人工智能时代及其教育的发展趋势[J]. 电化教育研究，2020，41(04):35-41
- (郑庆华等，2019)郑庆华，董博，钱步月，田锋，魏笔凡，张未展，刘均. 智慧教育研究现状与发展趋势[J]. 计算机研究与发展，2019，56(01):209-224
- (陈梦根，2013)陈梦根.算法交易的兴起及最新研究进展[J].证券市场导报，2013(9):11-17.
- (肖馨等，2019)肖馨，马远，陈璐.商业银行智能风控探索[J].中国金融，2019(11):44-46.
- (陶睿等，2019)陶睿，吴继春，谢胜强，郑海涛，毛子舒.深度学习和知识图谱在智能监管中的应用研究[J].金融纵横，2019(8):56-66.
- (赵大伟等，2020)赵大伟，李文华.人工智能技术在债券行业应用问题研究[J].金融与经济，2020(12):86-90.
- (季成和叶军，2021)季成，叶军.智能银行:关键要素、重点场景和完善路径[J].南方金融，2020(3):74-82.
- (杜浩云和张红波，2021)杜浩云，张红波.智能风控技术护航数字化转型[J].中国金融，2021(21):71-72.
- (薛亮等，2018)薛亮，刘丽颖，虞文杰.股票市场预测的小波神经网络模型[J].经济研究导刊，2018，0(3):95-95.
- (凌立文和张大斌，2019)凌立文，张大斌.组合预测模型构建方法及其应用研究综述[J].统计与决策，2019，0(1):18-23.
- (李斌等，2019)李斌，邵新月，李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济，2019(8):61-79.
- (胡慧芳和郑芬芳，2020)胡慧芳，郑芬芳.基于知识图谱的战略性新兴产业十年研究回顾[J].科研管理，2020，41(11):240-256.
- (吕华揆等，2020)吕华揆，洪亮，马费成.金融股权知识图谱构建与应用[J].数据分析与知识发现，2020，4(5):27-37.
- (Pan et al., 2020) Pan W, Li J, Li X. Portfolio Learning Based on Deep Learning[J].

Future Internet, 2020, 12(11):202.

- (Thakkar and Chaudhari, 2021) Thakkar A, Chaudhari K. A Comprehensive Survey on Deep Neural Networks for Stock Market: The Need, Challenges, and Future Directions[J]. Expert Systems with Applications, 2021, 177(2):114800.
- （金观涛等，2016）金观涛、刘青峰、邱伟云：〈《新青年》的数位人文研究〉，收于思想史编委会编著：《思想史 5》（台北：联经，2016 年 9 月），页 283-309.
- （包弼德，2017）包弼德：〈群体、地理与中国历史:基于 CBDB 和 CHGIS〉，《量化历史研究》2017 年第 Z1 期，页 213-246.
- （陈静，2017）陈静：〈数字档案化广告蜉蝣：以中国商业广告档案库（1880-1940）为例〉，《江海学刊》 2017 年第 2 期，页 165-171.
- （王平等，2018）王平、钮亮、金观涛、刘青峰：〈五代北宋山水画的数位人文研究（二）－以“渔隐”主题为例〉，《数位典藏与数位人文》第 1 期（2018 年 4 月），页 127 - 147.
- （Shih-Pei Chen, 2016）Shih-Pei Chen, “Remapping Locust Temples of Historical China and the Use of GIS,” Review of Religion and Chinese Society, Vol.3, No.2, 2016, pp.149–163.
- （杨海慈、王军, 2019）杨海慈,王军：〈宋代学术师承知识图谱的构建与可视化〉，《数据分析与知识发现》2019 年第 6 期，页 109-116
- （饶高琦, 2016; 饶高琦、李宇明, 2019, 孙琦鑫、饶高琦、荀恩东, 2020）饶高琦：《时代精神：基于 1946 年到 2015 年报刊语料和隐含主题模型的历史热词提取》，《语言 s 规划学研究》2016 年第 2 期，页 40-58；饶高琦、李宇明：《基于词频逆文档频统计的词汇时间分布层次》，《中文信息学报》2019 年第 11 期，页 31-38；孙琦鑫、饶高琦、荀恩东：《基于长时间跨度语料的词义演变计算研究》，《中文信息学报》2020 年第 8 期，页 10-22.
- CAIL: http://cail.cipsc.org.cn/
- COLIEE: https://sites.ualberta.ca/~rabelo/COLIEE2021/
- (Chen et al., 2020) Chen Y, Sun Y, Yang Z, et al. Joint Entity and Relation Extraction for Legal Documents with Legal Feature Enhancement[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020:1561-1571.
- (Gain et al., 2021) Gain B, Bandyopadhyay D, Saikh T, et al. IITP@COLIEE 2019: Legal Information Retrieval using BM25 and BERT[J]. 2021.

- (Huang et al., 2020) Huang W, Jiang J, Qu Q, et al. AILA: A Question Answering System in the Legal Domain[C]//Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20. 2020.
- (Li et al., 2018) Li J, Zhang G, Yan H, et al. A Markov Logic Networks Based Method to Predict Judicial Decisions of Divorce Cases[C]//Proceedings of in IEEE International Conference on Smart Cloud (Smart Cloud). 2018:129-132.
- (Li et al., 2021) Li L, Zhao L, Nai P, et al. Charge Prediction Modeling with Interpretation Enhancement Driven by Double-layer Criminal System[C]//Proceedings of World Wide Web. 2021:1-20.
- (Monroy et al., 2009) Monroy A, Calvo H, Gelbukh A. NLP for shallow question answering of legal documents using graphs[M]. GELBUKH A, ed.//Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 498–508[2021–06–04].
- (Quaresma and Rodrigues, 2005) Quaresma P, Rodrigues I P. A Question Answer System for Legal Information Retrieval[C]// The Eighteenth Annual Conference on Legal Knowledge and Information Systems, Brussels, Belgium, 8-10 December 2005.
- (Shao et al., 2020) Shao Y, Mao J, Liu Y, et al. BERT-PLI: Modeling Para-graph-Level Interactions for Legal Case Retrieval[C]//Proceedings of International Joint Conference on Artificial Intelligence. 2020:3501-3507.
- (Taniguchi and Kano, 2016) Taniguchi R, Kano Y. Legal Yes/No Question Answering System Using Case-Role Analysis[J]. Springer, Cham, 2016.
- (Tran et al., 2019) Tran V, Nguyen M, Satoh K. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model[C]//Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. 2019:275-282.
- (Yang et al., 2019) Yang W, Jia W, Zhou X I, et al. Legal judgment prediction via multi-perspective bi-feedback network[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.2019: 4085-4091.
- (Zhong et al., 2020) Zhong H, Wang Y, Tu C, et al. Iteratively questioning and answering for interpretable legal judgment prediction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:1250-1257.
- (Zhong et al., 2020) Zhong H, Xiao C, Tu C, et al. JEC-QA: A Legal-Domain Question Answering Dataset[C]//. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 9701-9708, 2020.
- (Duan et al., 2019) Duan X, Wang B, Wang Z, et al. CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension[J]. In Proceedings of

CCL. Springer. 2019.

- (Boshmaf et al, 2011)Boshmaf, Y., Muslukhov, I., Beznosov, K. and Ripeanu, M., 2011, December. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th annual computer security applications conference (pp. 93-102).
- (张洪忠等，2019)张洪忠,段泽宁,韩秀.异类还是共生:社交媒体中的社交机器人研究路径探讨[J].新闻界,2019(02)
- (Qin et al, 2018)Qin, L., Liu, L., Bi, W., Wang, Y., Liu, X., Hu, Z., Zhao, H. and Shi, S., 2018, July. Automatic Article Commenting: the Task and Dataset. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 151-156).
- (Zhou et al, 2018)Zhou, H., Huang, M., Zhang, T., Zhu, X. and Liu, B., 2018, April. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
- (Alshomary et al, 2021)Alshomary, M., Syed, S., Potthast, M. and Wachsmuth, H., 2021. Argument Undermining: Counter-Argument Generation by Attacking Weak Premises. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021 : 1816–1827.
- (Song et al, 2021)Song, H., Wang, Y., Zhang, K., Zhang, W.N. and Liu, T., 2021. BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021 : 167–177.
- (Dorri et al, 2018)A. Dorri, M. Abadi and M. Dadfarnia, 2018. SocialBotHunter: Botnet Detection in Twitter-Like Social Networking Services Using Semi-Supervised Collective Classification. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2018, pp. 496-503.
- (Fazil et al, 2021)M. Fazil, A. K. Sah and M. Abulaish, 2021. DeepSBD: A Deep

Neural Network Model With Attention Mechanism for SocialBot Detection. In IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4211-4223, 2021.

- (Suárez-Serrato et al, 2018)Suárez-Serrato, P., Velázquez Richards, E.I. and Yazdani, M., 2018. Socialbots supporting human rights. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 290-296).
- (Walker et al, 2021)Walker, M., Harrison, V., Juraska, J., Reed, L., Bowden, K., Cui, W., Patil, O. and Ratnaparkhi, A., 2021. Athena 2.0: Contextualized Dialogue Management for an Alexa Prize SocialBot. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 124-133).
- (Savvopoulos et al, 2018)Savvopoulos, A., Vikatos, P. and Benevenuto, F., 2018. Socialbots' first words: can automatic chatting improve influence in Twitter?. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 190-193).
- (Bowden et al, 2018)Bowden, K. K., Wu, J., Oraby, S., Misra, A., & Walker, M. (2018). Slugbot: An application of a novel and scalable open domain socialbot framework. arXiv preprint arXiv:1801.01531.
- (韩玉鑫，2019)韩玉鑫. 基于神经网络的微博观点检测方法研究.
- (何炎祥等，2016)何炎祥，刘健博，孙松涛. 基于神经网络的微博舆情预测方法[J]. 华南理工大学学报：自然科学版，2016，44(9):6.
- (马梅等，2016)马梅，刘东苏，李慧. 基于大数据的网络舆情分析系统模型研究[J]. 情报科学，2016，34(3):5.
- (聂方彦，2017)聂方彦. 基于模糊 C 均值的舆情等级分类模型研究[J]. 软件导刊，2017，16(6):3.
- (田俊静等，2021)田俊静，兰月新，夏一雪，等. 基于决策树方法的网络舆情反转识别与实证研究[J]. 2021(2019-8):121-125.
- (王超等，2018)王超，彭湃，李波. 舆情短文本挖掘的数学模型及其实现[J]. 数学建模及其应用，2018，7(3):9.
- (王少鹏，2014)王少鹏. 基于 LDA 的文本聚类在高校网络舆情分析中的应用研究[D]. 首都师范大学，2014.
- (吴谦，2019)吴谦. 基于机器学习的微博舆情预测模型研究[D]. 中国人民公

安大学，2019.

- (杨青和邱扶东，2021)杨青，邱扶东. 虚拟旅游产品结构优化研究——基于网络数据采集[J]. 2021(2018-4):97-101.
- (张和平和陈齐海，2021)张和平，陈齐海. 基于灰色马尔可夫模型的网络舆情预测研究[J]. 2021(2018-1):75-79.
- (Cao et al., 2014) Cao F, Zhang Z, Jing Y, et al. A model of ecological monitoring and response system for Internet public opinion[J]. International Journal of Multimedia and Ubiquitous Engineering, 2014, 9(5): 373-390.
- (Suvorov et al., 2014) Suvorov R, Sochenkov I, Tikhomirov I. Training Datasets Collection and Evaluation of Feature Selection Methods for Web Content Filtering[C]// International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Springer International Publishing, 2014.
- (Borisyuk et al., 2021)Borisyuk, F., Malreddy, S., Mei, J., Liu, Y., Liu, X., Maheshwari, P., & Rangadurai, K. (2021, August). VisRel: Media Search at Scale. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 2584-2592).
- (Chen et al., 2019) Chen, Q., Zhao, H., Li, W., Huang, P., & Ou, W. (2019, August). Behavior sequence transformer for e-commerce recommendation in alibaba. In Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data (pp. 1-4).
- (Cheng et al. 2016) Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems (pp. 7-10).
- (Fan et al., 2021)Fan, M., Sun, Y., Huang, J., Wang, H., & Li, Y. (2021, August). Meta-Learned Spatial-Temporal POI Auto-Completion for the Search Engine at Baidu Maps. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 2822-2830).
- (Guo et al., 2021)Guo, W., Su, R., Tan, R., Guo, H., Zhang, Y., Liu, Z., ... & He, X. (2021). Dual Graph enhanced Embedding Neural Network for CTRPrediction. arXiv preprint arXiv:2106.00314.
- (Huang et al., 2021)Huang, J., Wang, H., Sun, Y., Fan, M., Huang, Z., Yuan, C., & Li, Y. (2021, August). HGAMN: Heterogeneous Graph Attention Matching

Network for Multilingual POI Retrieval at Baidu Maps. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3032-3040).

- (Lei et al., 2021)Lei, C., Liu, Y., Zhang, L., Wang, G., Tang, H., Li, H., & Miao, C. (2021, August). SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3161-3171).
- (Okura et al., 2017)Okura, S., Tagami, Y., Ono, S., & Tajima, A. (2017, August). Embedding-based news recommendation for millions of users. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1933-1942).
- (Sun et al., 2020)Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., ... & Zheng, K. (2020, October). Multi-modal knowledge graphs for recommender systems. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 1405-1414).
- (Wang et al., 2021a)Wang, C., Pan, H., Liu, Y., Chen, K., Qiu, M., Zhou, W., ... & Cai, D. (2021, August). Mell: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3649-3659).
- (Wang et al., 2021b)Wang, X., Gu, X., Cao, J., Zhao, Z., Yan, Y., Middha, B., & Xie, X. (2021, August). Reinforcing Pretrained Models for Generating Attractive Text Advertisements. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3697-3707).
- (Zhou et al., 2018)Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., ... & Gai, K. (2018, July). Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1059-1068).

第九章　知识图谱领域研究发展、现状及趋势

9.1.引言

知识图谱（Knowledge Graph， KG）旨在描述客观世界的概念、实体、事件及其之间的关系，知识图谱把信息表达成更接近人类认知世界的形式，把互联网内容转化成计算机可理解和深度关联的语义。大数据时代需要把数据转化成知识，为数据增添语义信息，获得对大数据的洞察，使数据产生智慧，以提供决策支持等智能服务。

知识图谱的发展历史源远流长，从经典人工智能子领域——知识工程，到互联网时代的语义 Web，再到当前很多领域构建的数千亿级别的现代知识图谱，以及大规模知识在搜索推荐、智能问答、特定领域知识图谱和大数据分析的广泛应用。知识图谱已经成为以知识为代表的知识获取和应用服务的人工智能技术，是人工智能的知识驱动方法的核心。

知识图谱兼具人工智能、大数据和互联网的多重技术基因，是来自于知识表示与推理、自然语言处理、机器学习和数据库技术等等多个领域的交叉融合。知识图谱同时也是不断发展的新领域，并在不断与深度学习、联邦学习、区块链、物联网（IoT）、视觉计算等众多领域的新发展进一步融合，不断更新和进步。

当前人工智能正在从感知智能向认知智能发展。以大数据深度学习为代表的数据驱动的人工智能方法在感知智能上取得了很大成功，但是实践证明当前方法还不具有人类认知水平的能力，如推理和对客观世界深度理解的能力。以知识图谱为代表的知识驱动方法，虽然可以对客观世界的深层语义结构进行表示和推理，但是存在知识不足带来的知识计算的脆弱性。因此，结合数据驱动的深度学习和知识驱动的知识图谱技术已经成为人工智能从感知智能向认知智能发展的可行途径。

9.2.关键任务与科学问题

知识图谱关键任务主要包含两个方面：知识图谱构建和知识图谱应用。具体来说，知识图谱构建是研究如何在计算机内部表示、组织知识内容，并从外部多类型数据中用何种方法获取海量知识内容，最终形成知识图谱系统。其相关技术包括：知识表示与建模、知识获取、知识融合等。知识图谱应用主要研究如何利用知识图谱更好地解决实际应用问题，所涉及的技术主要包括：知识存储与管理、

知识推理、语义搜索与知识问答等。具体来说，其中关键任务如下：

知识表示与建模：知识图谱以结构化的形式描述客观世界中概念、实体间的语义关系，将信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解海量信息的能力。

知识获取：知识图谱获取是根据确定知识表示模型，从分布异构的海量数据资源中获取知识的过程。目前，知识图谱的构建一般多依赖已有的结构化数据，通过映射到预先定义的 Schema 或本体来快速冷启动，然后利用自动化抽取技术，从半结构化数据和文本中提取结构化信息来补全知识图谱。其中基于机器学习的信息抽取技术是其中最为关键的技术，这两年也成为研究界关注的热点。

知识融合：在知识图谱的构建过程中，获取的知识往往是碎片的、冗余的，很多时候都需要将多个来源数据中的知识（实体或概念）映射到统一的命名空间中。主要包含本体概念层面和实体实例两个层面的融合任务。

知识图谱存储与查询：这一任务的主要目标是搭建图数据库并建立知识图谱查询引擎，这也是很多知识图谱项目在实际场景中进行应用的基础工作。知识图谱的存储需要综合考虑知识的结构、图的特点、索引和查询优化等问题。

知识推理：知识图谱推理的目标是利用图谱中已经存在的关联关系或事实来推断未知的关系或事实，在知识图谱的各项应用任务中发挥着重要作用，这包括链接预测、补全属性、检测错误和识别语义冲突，拓展问句语义，提升推荐精准和可解释性等多种任务。

语义搜索与知识问答：知识图谱可以将用户搜索输入的关键词，映射为知识图谱中客观世界的概念和实体，搜索结果直接显示的满足用户需求的结构化内容；而知识问答系统将知识图谱看成一个大规模的知识库，通过理解将用户的问题转化为对知识图谱的结构化查询。

图数据挖掘与知识分析：知识图谱作为一种基于图结构的数据，可以充分利用各种图挖掘算法对知识图谱进行深度分析，这包括基于图论的基础图算法，也包括图嵌入、图神经网络等图表示学习新方法。

这些任务中所涉及的关键科学问题包括：

- 在知识表示方面，知识表示是知识图谱的基础，是对于客观世界知识中所蕴涵的语义内容以及关联进行的刻画和描述。知识表示既要考虑知识的表示与存储，又要考虑知识的使用和计算，其中需要解决的关键问题是：1)建立什么样的知识表示形式能够准确地反映客观世界中不同粒度、层次的知识？2)知识表示如何支持高效知识推理和计算，从而使知识表示具有得到新知识的推理能力。
- 在知识图谱构建方面，主要任务包含知识获取和知识融合两项核心技术。

其中需要解决的关键问题包括：1）从多种异构数据资源中如何获取知识？主要包括结构化(如数据库数据)、半结构化(如互联网上的表格数据等)、非结构化资源(如文本数据等)对象和多模态数据（图像、语音、视频、文字）中如何获取知识。2）针对不同类型、不同结构的知识（实体、事件、常识等），面对复杂场景（小样本数据场景、领域迁移场景等）如何精准的获取知识。3）针对异构知识资源中获取的片面、冗余的知识，如何学习不同知识之间的映射、关联关系，实现不同知识的相互补充，构建统一、规范的知识图谱？

- 在知识图谱应用方面，其目标是面对具体下游应用（问答、推荐等），如何基于知识图谱建立智能知识服务，提升应用的智能化水平。其中需要解决的关键问题包括：1）面对所构建的图结构海量知识内容，如何实现知识的高效存储和快速查询？2）面向多类型下游应用，如何实现知识的精准匹配、查询和计算？3）面向未知知识，如何在已有知识图谱的基础上，实现大规模知识推理？

知识图谱不是单一技术，构建及应用好知识图谱需要建立系统工程思维。知识图谱本质是质量更高的数据，沉淀知识是每个领域都需要持续投入的系统知识工程。同时知识图谱技术涉及数据、算法、工具和系统等多个维度的任务和目标，其价值也需要通过多个技术点的叠加才能最大限度发挥。

9.3.关键技术进展及研究趋势

知识图谱的关键技术涉及自然语言处理、语义网、数据库、信息检索、数据挖掘、机器学习等多个领域，相关研究工作在近年来越来越多地受到国内外学者的关注。已有研究工作主要围绕知识表示与建模、知识获取、知识融合、知识推理、知识资源建设等方面展开，下面将分别进行介绍。

9.3.1. 知识表示与建模

概念是凝聚人们认知世界的粘合剂，是人们理解客观世界的线索，是人们对客观世界中的事物在不同层次上的抽象化描述。很多知识图谱更关注实体及实体间关系，而从基于本体的知识表示方法来看，概念模型可以看作知识图谱的“骨骼”，是不可或缺的一部分。知识根据本身特性可细分为实体类、事件类和规则类等，表示方法不尽相同。知识建模就是从大规模数据中获取这些不同性质的概念知识，建立知识图谱顶层知识模型的过程，其主要任务包括概念抽取、概念层次学习和概念属性挖掘，具体地：

概念抽取旨在从语料中识别领域相关的术语，通常为代表该领域的某一个具体概念的词组。概念层次学习旨在确定概念间的上下位关系，判断两个概念之间是否存在 SubClassOf 的关系，将识别得到的概念和关系组织成一个合理的分类体系，通常为树或者有向无环图。属性是概念的内涵表达，描述概念的特征或性质，具有描述和鉴别概念的功能。概念属性挖掘旨在对给定概念从不同类型的数据源中自动获取其属性集合，并对属性重要程度进行量化计算。

弱资源问题是知识建模任务的核心挑战。由于知识增速爆炸，大量概念缺乏充分的描述性信息，也缺乏高质量标注的数据集。以维基百科为例，约 14%的概念的描述信息不超过一段文本，同时概念间上下位关系也存在大量缺失或错误。

9.3.1.1. 技术进展

9.3.1.1.1. 概念抽取

概念抽取任务于 2007 年首次被形式化定义[Wang, 2007]，即给定特定类别下的若干概念示例，目标为扩展改类别下的概念集合，而初期的主要研究多基于模板匹配和特征工程，2011 年起逐渐形成目前主流的研究方法，即基于文本的自举迭代模型[He, 2011]。然而，此类方法的共性问题为语义漂移，即在扩展过程中，偶然引入的错误会使噪声不断累积传播，使得后续扩展轮次的效果持续变差。因此，近年来的研究主要是围绕该问题的缓解，核心思路包括多集合协同扩展和迭代过程检验两种，具体地，

多集合协同扩展主要思想是假设待扩展概念可以划分为不同细粒度的类别，某一类别下的噪声概念对于其他类别可能为正确概念，如果能在扩展过程中逐渐明确类别边界，同时将可能产生噪声的概念集合同时扩展，就可以互相提供监督，减少噪声的引入。代表性方法包括多样本集成 SetExpan[Shen, 2017]、MOOC 课程概念扩展算法[Yu, 2019;2020]以及协同扩展法 CoExpan[Huang, 2020]。

迭代过程检验的核心思想是在每一次迭代计算过程中使用一定的外部信息进行扩展纠正，以直接减少后续的错误传播。Yan 等使用蒙特卡罗树搜索算法对于模版的质量进行更精确的评估[Yan, 2019]，Zhang 等提出的 CGExpan 在扩展过程中引入上位概念的指导[Zhang, 2020]，Shen 等进一步将上位概念发现与概念扩展联合学习[Shen, 2020a]。

9.3.1.1.2. 概念层次学习

概念层次学习的关键是判断给定的概念对的上下位从属关系，该任务最早的

方法论可以追溯至 1992 年总结出来的 Hearst Pattern [Hearst, 1992]，即基于模式匹配的方法，尽管其所依赖的词法、句法匹配模式语言相关性强，仅在部分西方语言中明显存在，但至今仍是主流方法之一，且催生了基于维基分类体系的上下位识别算法，即主要利用维基开放分类的和词条标签的词法模式完成概念体系模型的建立[Ponzetto, 2007; Gupta, 2016;2018]。

另一种思路主要建模上下位关系的不对称性，基于对给定概念对概率分布的估计来判断其上下位关系。随着文本嵌入技术的发展，此类方法自 2014 年起逐渐繁荣[Julie, 2014]，其本质是将上下位关系发现形式化为多标签分类问题，核心则是合理地对概念节点进行表示学习，融合节点的局部特征、结构特征[Mao, 2020]、文本特征[Shen, 2021]，从而实现精准分类。

此外，还有一类研究扩展了该任务的外延，即不需逐对进行关系识别，而是对整个概念体系树进行建模，以得到更准确的上下位关系[Zhang, 2018; Shen, 2018;Shen, 2020b]。

9.3.1.1.3. 概念属性识别

与概念层次学习类似，概念属性识别最普遍的方法也是基于词汇句法模式，通常利用利用少量种子属性在半结构化数据 HTML 页面及非机构化文本中的模式学习和属性识别[Pasca, 2007; Ravi, 2008; Bellare, 2007]，或者将属性识别与其他相关任务进行联合学习，如实例扩展[Zhang, 2016]。另一类研究是评估属性对给定概念的重要程度，如 Lee 等人在 Probase 上提出基于排序学习对属性典型性概率化学习[Lee, 2013]，Lajus 等人提出一种必要属性过滤方法，其基本假设是必要属性不会随着分类树对概念的划分而发生明显变化[Lajus, 2018]。

9.3.1.2. 技术发展趋势

随着知识图谱的普及以及数据与知识双轮驱动第三代人工智能方法的发展，越来越多场景对知识概念模型的需求在逐渐提高，而为了应对弱资源问题进行的技术探索将是知识建模技术发展的主线，具体可以体现在如下几个方面：

符号与数字融合：大规模预训练模型在很多任务上显示出统治性效果，也证明了其中存储了大量源于无标注数据的知识。近来快速发展的 Prompt Tuning 技术提供了一种与超大规模模型交互的方式，且 Prompt 的设计与知识建模中常用模式具有一定程度上的相似性，因此，探索将符号化的知识模型与数字化的预训练模型结合，利用预训练赋能高质量知识模型的构建，是未来的主要方向之一。

通用与领域融合：通用域概念模型的资源丰富，特定领域内高质量语料则相

对匮乏但对知识模型的需求更迫切，Yu 等尝试将维基分类体系应用于课程概念扩展中[Yu, 2019]，取得了一定效果，因此，高效利用已有通用域模型提升领域建模质量，也是未来需要探索的重要方向。

多语言知识融合：结构化知识在语言上表现出严重的不平衡性，概念模型更是如此，如 DBpedia 本体、Schema.org 和维基开放分类中英文规模和质量均高于中文，国内 OpenKG 也在致力于通用的中文概念模型，但整体上滞后，因此构建多语言融合概念模型，高效利用资源丰富语言的知识积累，也是重要发展趋势。

9.3.2. 知识获取

知识获取（Knowledge Acquisition）的目标是由无结构化的数据中抽取结构化数据并存储起来，其基本任务可分为命名实体识别、关系抽取、事件抽取等任务。命名实体识别在于从输入数据中自动获取人名、地名、机构名等命名实体，关系抽取承接于命名实体识别在于从输入数据中获取实体之间的相互关系，事件抽取则在于识别输入数据中包含的事件类型以及事件对应的角色。以上任务是在给定抽取目标的前提下获取满足要求的文本片段，其挑战可归结为如何建模文本片段和抽取目标的对应关系。

然而，随着网络数据的爆炸性增长，人们很难事先确定抽取目标，故而开放域信息抽取逐步成为知识获取中的一个重要分支，开放域信息抽取无需事先定义抽取目标，而获取一系列的“主语、谓语、宾语”三元组结构。开放域信息抽取可分为无指导和有指导两类。无指导方案受限于事先定义的模板，而有指导方案则对训练数据异常敏感。故此，开放域信息抽取面临的最大挑战即是如何减少对模板和数据的依赖性。

其次，传统的知识获取仅针对文本，而互联网上存在大量多模态数据，且多个模态的知识相互补充、增强。基于此，多模态信息抽取成为热门的研究领域。模态不同意味着数据分布不同，也意味着建模方式的不同。故此，多模态信息抽取的挑战即在于如何对齐和融合多模态数据，通过多个模态蕴含知识的互补提升知识获取的准确性和完整性。

9.3.2.1. 技术进展

9.3.2.1.1. 命名实体识别

早期多是基于规则和字典的方法，观察命名实体的构成，总结出规则集并构建模板，其在特定语料上有较高效果，但可移植性差。上世纪九十年代出现基于

统计的机器学习方法，本质是序列标注，有较好效果，但依赖特征的人工选取。基于深度学习的方法主要通过深度神经网络获取词和文本的向量表示，减少特征工程的工作量，同时结合上下文语义信息，但需要大量标注语料。尤其预训练模型的提出大大提升识别效果，自此成为主流方法，其核心是利用预训练模型从通用领域中学习含上下文语义信息的词向量表示，再在特定任务中进行微调，同时结合各种深度学习模型进一步提升效果。资源丰富的命名实体识别任务已取得较好进展，目前研究者的关注点集中于有限样本条件下的命名实体识别，包括不连续命名实体（如嵌套、间断等）、命名实体识别模型的领域迁移、细粒度命名实体识别、命名实体识别任务中时间和可解释性等[Huang, 2021; Lison, 2020]。

9.3.2.1.2. 关系抽取

早期的关系抽取主要依赖人工构建的规则在文本中进行模板匹配，这种方式很难处理长距离依赖关系。随着深度学习的发展，采用神经网络的方法已成为该领域的主流方法，但是精确标注样本往往要求昂贵的成本。远程监督是一种通过对齐知识图谱中的实体与语料中的实体提及，自动标注大量样本的方法，然而这样会制造大量噪声，如何降低噪声影响是需要解决的主要问题。目前，复杂语境下的实体关系联合抽取和文档级关系抽取逐渐成为该领域的研究热点。联合学习 NER 与关系抽取任务，有利于增强任务之间的特征共享、缓解级联错误问题，可分为基于参数共享和基于联合解码的两种思路。而如何提升模型依据跨文档内长距离依赖特征的推理能力是文档级关系抽取任务的主要挑战，目前主流工作多采用图结构构建文档图，通过 GNN 学习节点间的信息传递，取得不错效果[Fu, 2019]。

9.3.2.1.3. 事件抽取

早期事件抽取以模板方法为主，在特定领域可以取得较好效果，但可移植性差。随着机器学习方法的流行，研究者借鉴文本分类思想，将事件抽取转换为分类问题，通过人工选取特征提高泛化能力。当前基于深度学习的方法逐渐成为主流，其利用模型自动从连续向量中学习更加抽象的特征，将事件抽取视为分类和序列标注任务。深度学习的方法按照模型学习流程可以分为基于流水线和基于联合学习。基于流水线的方法先学习事件识别模型，再学习分类模型。基于联合学习方法则不单独识别触发词而是直接进行端到端的学习。传统事件抽取是基于句子级的，然而很多事件的主客体存在跨句子联系，故此提升模型跨句抽取事件的能力成为当前研究热点。近几年随着事理图谱的研究深入，人们逐渐关注事件推理、事件泛化方面的工作，这些工作从已抽取到的事件元素泛化事件类型，进而

挖掘事件在时间维度和空间维度的联系来支撑推理[Du, 2021]。

9.3.2.1.4. 开放域信息抽取

开放域信息抽取对于知识的构建至关重要，可以减少人工标注的成本和时间。早期的方法首先从文本中抽取得到候选知识三元组，然后利用预先训练的分类器对这些候选目标进行判别，最后利用全局信息对知识元组进行重排序以获得高质量知识。但是这种方法抽取知识的准确率和召回率并不理想。为此，提出使用浅层句法特征进行开放式信息抽取，利用浅层句法约束来消除错误知识元组以及无意义知识元组的抽取。人们发现这些基于词性和句法的开放式信息抽取技术对于复杂句子往往无法有效抽取准确且全面的知识元组，于是提出将复杂语句转换为多个句法结构简单的分句，从而简化目标文本。知识并不局限于陈述事实，作为对事实限定的条件也是一种知识。引入条件能够更加准确的描述事实，当前有研究者专注于条件知识的抽取，并将其应用于下游检索或推理任务中[Jiang, 2019]。

9.3.2.1.5. 多模态知识获取

早期多模态知识获取研究使用多模态信息辅助纯文本知识获取，这一阶段主要利用多模态信息进行文本知识的消歧和完善，处理范式是在纯文本处理模型的框架下，加入编码其他模态信息的分支，分别得到纯文本和其他模态的表示，最后使用模态融合得到多模态表示，该研究的核心在于如何进行模态融合，传统方法有简单的向量操作，例如向量相加，乘积，双线性池化等，实现比较简单，在融合性能上具有不错的效果，但是建模能力有限。目前常用的方法是基于神经网络的融合，例如结合门控机制的注意力机制，基于堆叠多层注意力机制的多模态Transformer。目前主要关注的问题是如何从不同模态中共同抽取知识，即分别从不同模态中抽取知识，然后采用跨模态对齐工具进行知识的链接与消歧[Li, 2020]。

9.3.2.2. 未来技术发展趋势

当前深度学习模型和预训练语言模型的出现能够很好的应对训练数据充足、知识获取目标简单的情况，故知识获取的未来技术发展趋势可归结为：

1、小样本知识获取：深度学习模型都需要大量的标注数据，甚至模型的性能与标注数据量成正比。元学习和 prompt 方法的出现从性能上来看还不足以突破数据规模的限制。弱监督方法虽然能够缓解人工标注数据的难度，但是噪声过大，因此如何在基于少量标注样本和大量无标注样本，获得高性能的抽取模型具

有极大的研究价值。

2、跨语言、跨领域知识获取：目前知识获取的目标是针对某一特定语言或某一具体领域，然而不同语言的语法特征不同，不同领域也有各自的专有名词和术语，这就导致了在某一语言或某一领域性能优越的模型，迁移到别的语言或领域下效果会不尽人意。因此，如何利用迁移学习实现跨语言、跨领域知识获取方法即成为热门的研究方向。

3、篇章级知识获取：目前主流的知识获取任务大多局限在对单个句子的处理，文档具有复杂的结构信息，很多实体或关系的抽取需要依赖于对文档全局的理解，受限于预训练模型的输入长度限制以及篇章级抽取技术的效果不尽如人意，篇章级知识抽取必然成为新的研究热点。

4、多模态时序知识获取：当前的多模态知识获取方法并没有明显的引入时间信号。现实世界中的知识尤其是事件是动态变化的，文本会描述连续的事件，而视频更会描述事件随时间的变化情况。故而，在知识获取中考虑时序信息，从多模态数据中获取动态变化的知识即成为当前的研究趋势。

9.3.3. 知识融合

知识融合（Knowledge Fusion）的目标就是将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序间的交互建立互操作性。知识融合的常见处理流程主要包括：输入、预处理、本体匹配、实体对齐、真值发现和输出 6 个环节。关键技术包括本体匹配（也称为本体映射）、实体对齐（也称为实例匹配、实体消解）以及真值发现（也称为真值推断）等。面临的核心挑战主要包括大规模、异构性、低资源等问题。

9.3.3.1. 技术发展脉络和进展

9.3.3.1.1. 实体对齐

实体对齐旨在发现指称真实世界相同对象的不同实例。例如，DBpedia 中的 *Mount_Everest* 和 Wikidata 中的 *Q513* 均指称珠穆朗玛峰。如何从语义上消解实体间的异构性是实体对齐待解决的关键科学问题。

传统的实体对齐方法可以分为基于等价关系推理的方法[Glaser, 2009]和基于特征相似度计算的方法[Volz, 2009]。近年来，伴随着表示学习（Representation Learning）技术在图像、视频、语音、自然语言处理等领域的成功，如何将表示学习技术用于实体对齐成为一个新的研究热点。下面着重介

绍基于表示学习的实体对齐技术进展。

如图 1 所示，基于表示学习的实体对齐方法典型框架以两个不同知识图谱作为输入，并收集已知实体对齐，然后在嵌入模块和对齐模块中输入这两个知识图谱和已知实体对齐，通过学习到的嵌入表示来度量实体相似性。

现有工作主要研究如何学习高质量的实体嵌入表示。根据利用的实体特征，可以分为两大类：基于关系的方法和属性增强的方法。前者利用知识图谱的关系结构进行表示学习，主流技术包括平移模型（如 TransE [Bordes, 2013]）、循环神经网络[Guo, 2019]、以及图神经网络（GNN）[Kipf, 2017; Wang, 2018; Wu, 2019; Cao, 2019; Li, 2019; Mao, 2020; Sun, 2020; Yu, 2021]。属性增强的方法除了基于关系结构进行表示学习，还额外引入属性信息，如实体描述[Chen, 2018]、实体名称[Zhang, 2019; Zeng, 2020]、图像信息[Chen, 2020; Liu, 2021]以及其他普通属性等[Trisedya, 2019; Liu, 2020]。近期一些工作考虑将表示学习技术与传统实体匹配技术相结合[Qi, 2021]、优化表示学习效率[Mao, 2021]，以及引入主动学习技术[Berrendorf, 2021; Zeng, 2021; Liu, 2021]等。

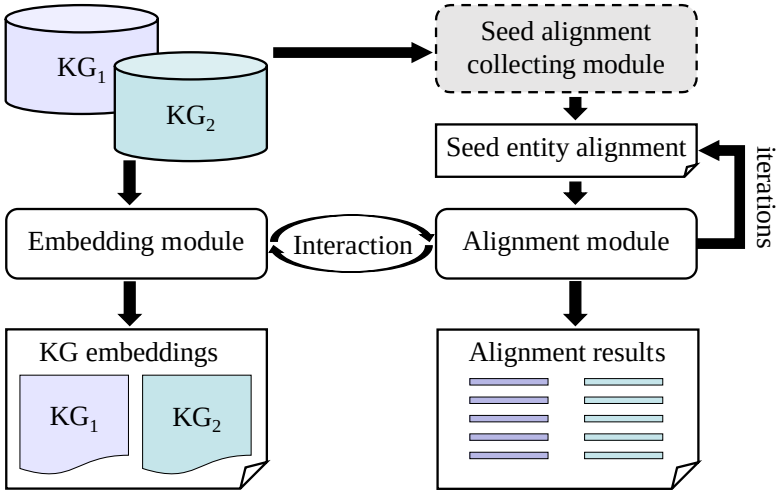


图 1：基于表示学习的实体对齐方法典型框架[Sun, 2020]

9.3.3.1.2. 真值发现

真值发现一般通过冲突检测、真值推断等技术消解知识融合过程中的冲突，再对知识进行关联与合并，最终形成一个一致的结果。如何处理冲突是真值发现的主要研究问题。

常见的真值发现方法可以分为 3 类：迭代模型、优化模型、概率图模型。迭代模型[Yin, 2008; Pasternack, 2010]首先给每个数据源设置初始置信度，交替式地先基于数据源置信度估计真值，再基于真值更新数据源置信度，直到算法收敛或

达到终止条件。优化模型[Li, 2014]通常以最小总体误差作为优化目标，通过连续优化找到最优的真值。而概率图模型[Wang, 2015; Li, 2019; Cao, 2020]假设数据源以特定的概率分布产生噪音数据，并利用最大期望算法来最大化观测值似然函数，从而找到概率分布的参数和真值。

最新研究也开始运用表示学习技术。Knowledge Vault [Dong, 2014]基于张量分解和路径排序技术使用知识图谱中的知识验证新的事实，并渐进式地将正确的事实添加到图谱中。OKELE [Cao, 2020]使用图神经网络推断实体可能具备的属性，同时使用概率图模型和最大期望算法推断出正确的属性值。CASE [Lyu, 2021]先将数据源、问题和观测值之间的关联建模成异构信息网络，再学习网络中节点的嵌入表示，并估计出真值的嵌入表示，最后选择最相似的观测值作为结果。

9.3.3.1.3. 本体匹配

本体匹配侧重发现本体模式层的等价或相似的类、属性或关系。近年来，关于本体匹配的研究进展不多，不过值得一提的是，2021 年 LogMap [Jiménez-Ruiz, 2011]获得了语义网科学联盟（SWSA）十年最有影响力论文奖。

9.3.3.1.4. 未来技术发展趋势

知识融合在过去几年里得到了广泛的关注，未来可能的研究方向包括：

首先，表示学习是近年来的研究热点，未来一段时间将依然延续这一趋势。主要研究热点可能包括表达能力更强的图神经网络、自监督表示学习、可以融合多种特征的自适应表示学习，以及基于预训练模型的实体对齐技术等。此外，知识融合在现实应用中仍面临很多挑战，如大规模（千万级实体的）知识图谱实体对齐、基于表示学习的知识图谱划分、时序或动态知识图谱实体对齐，以及带有悬挂实体的知识图谱对齐等。实体对齐的鲁棒性与可解释性等也有待深入研究。

其次，持续融合有可能成为知识融合的新发展方向，基于增量式学习，不断融合现有知识图谱的新增实体或其它开放数据，打造终身学习的知识融合系统。同时，真值之间通常存在关联性，利用图谱内部知识自动化挖掘这种关联性推断出一致的真值也有待深入研究。此外，也可以设计众包系统结合人工干预提高真值推断的准确性。

最后，知识融合与其他知识图谱研究领域的互动也值得关注。例如，当前知识库问答主要针对单个知识图谱，而利用多源知识融合可以弥补单一知识图谱不完备的缺点，提高回答（例如多跳问题）的效果。此外，知识融合与知识抽取、知识推理等也有很多可以结合的方向。

9.3.4. 知识推理

知识图谱的推理任务而言，主要是利用图谱中已存在的关联关系与事实来推断未知的关系或事实。知识图谱上主要的推理方式可大致分为基于符号的推理与基于统计的推理[Li, 2020]。基于符号的推理通过利用语义框架来形式化上述问题，并利用预先定义的规则来推断出隐含的知识。基于统计的推理试图找到合适的统计模型来拟合样本，并利用模型来预测出图谱中实体之间推断关系预期的概率。随着大数据和深度学习的兴起，基于神经网络和表示学习的知识图谱推理方法逐渐流行并被广泛应用，包括基于知识图谱嵌入与预训练的推理、基于神经符号集成的知识图谱推理、基于神经网络的知识图谱推理以及基于本体表示的知识图谱推理等。不同的表示方法和推理模式在近年来的知识图谱推理发展中呈逐渐融合的趋势，提升了知识图谱推理方法的鲁棒性、可迁移性、可解释性、可应用性等。

9.3.4.1. 技术发展脉络和进展

9.3.4.1.1. 基于神经网络和表示学习的知识图谱推理

基于知识图谱嵌入与预训练的推理 与词向量的思想类似，知识图谱嵌入推理将实体和关系映射到向量空间，称为实体或关系的嵌入表示，嵌入表示支持通过计算获得实体或关系的语义信息，例如，实体的相似度、关系的性质以及实体和实体之间的关系等。作为最早的知识图谱嵌入表示学习方法之一TransE[Bordes, 2013]将头实体到尾实体的映射看作向量的平移翻译，模型简单有效，但对复杂的关系表达能力不足，例如，关系的自反性、可逆性、传递性以及组合性等，随后众多可编码关系多样语义的模型被提出[Wang, 2014][Théo, 2016][Sun, 2019]，逐渐提升了知识图谱嵌入推理方法的表达能力。随着大规模预训练语言模型在自然语言处理领域取得卓越的进步，基于“预训练+服务”理念的知识图谱预训练模型被提出[Zhang, 2021]，将知识图谱嵌入推理方法实用化，基于其补全和推理的能力为下游知识图谱应用任务提供更好的知识服务。

基于神经符号集成的知识图谱推理 神经推理和符号推理各具优缺点，符号推理依赖于规则和本体这类难获取的知识，神经推理这种数据驱动的方法无法得到精确的预测同时无法提供良好的解释。因此基于神经符号集成的知识图谱推理致力于让两种方法优势互补。其中一种集成方法是用符号知识（如规则、路径等）约束神经模型的训练，作为目标函数的一部分以优化预测结果[Niu, 2020]，另一种集成方式是将神经模型应用到符号推理过程中，为符号推理加入软逻辑，以避免知识图谱不全导致的推理链中断问题[Ho, 2018]，还有一类工作同时进行符号

推理和神经模型的双向集成，使两者形成动态互补[Zhang, 2019]。这些基于神经符号集成的知识图谱推理方法在鲁棒性、可迁移性、可解释性方面相较非集成模型均有显著提升。

基于图神经网络的知识图谱推理 受到图神经网络在同构网络研究上的启发，应用于知识图谱推理的图神经网络主要基于知识图谱的图结构进行学习。与之不同的是，知识图谱推理还需要考虑节点和边的语义类型信息以支持更复杂的逻辑推理。对比可以隐含地捕获图结构的基于图谱嵌入的推理，基于图神经网络的推理会显式地对图结构以及节点特征进行编码，因而可以有效地利用实体的邻居实体信息和连接关系进行推理。此类典型的算法有 R-GCN[Schlichtkrull, 2018]、CompGCN[Vashishth,2020]以及 ExpressGNN[Zhang, 2020]等，并已广泛应用于长尾关系抽取、实体对齐、零样本图像识别、对话生成以及推荐系统等应用中。随着知识图谱规模不断扩大，大图数据的处理，也就是基于大数据的计算引擎的图计算也是需要深入研究和考虑的技术问题。

基于本体表示学习的知识图谱推理 三元组和图结构信息能较好地支持简单直接的知识图谱推理，而复杂的推理往往依赖于知识框架，又称为本体，其中的实体以概念与属性为主。本体嵌入表示主要侧重于将概念层次体系、概念之间的逻辑组合关系、属性的层次体系、概念和属性之间的逻辑组合以及属性自身的性质（如：传递性、对称性、自反性）等这类抽象的知识编码到稠密、连续的语义空间中，即如何将本体语义和逻辑表达进行向量化表示。典型的本体嵌入模型是 EL Embedding[Kulmanov 2019]，该模型将轻量级的 EL 本体利用高维的球形空间来表示，用球心之间的位置来编码概念之间的关系。现有的本体表示学习研究较多的是概念之间的层次关系以及连接关系，对于更富在的逻辑表达包括组合语义（或/且/非）、存在量词和全称量词涉及较少。受益于近些年来复杂知识库问句查询的研究[Ren, 2020]，本体表示学习[Chen,2021]在向量空间的逻辑表达能力得到了进一步的探索。

9.3.4.1.2. 推理应用

推理已广泛应用于知识图谱、自然语言处理以及图像等领域的学术研究[Wang, 2017] [Ji, 2020]，例如面向知识图谱的查询问答研究中，嵌入表示方法成为了单步推理的标准模块，组合推理过程则大量采用了神经符号集成的推理方法；在自然语言处理领域以及图像处理领域，知识图谱作为背景知识配合上述推理方法，常常用于克服标注数据的稀疏性、包含噪音以及缺乏领域知识的问题。

除了在学术界，基于神经网络和表示学习的知识图谱推理在工业界也有一定

的应用价值。借助上述方法中具有的不确定特性，可以进一步优化产业模型中排序的精度。例如，阿里巴巴借助知识图谱推理技术增强了商品分类、对齐、推荐以及假货识别等任务以提升商品数据的管理和维护，华为将其应用于故障诊断、异常检测等任务以快速响应客户需求，字节跳动将该推理技术来辅助“用户意图”的识别上，从而进一步提高知识库问答、知识搜索的效果；腾讯则利用补全的信息来提升实体链接和实体推荐的精度；小米公司同样利用这些推理模型的结果提升了实体链接、实体推荐的效果，从而更好的为小爱同学等一系列产品进行赋能。

9.3.4.1.3. 未来技术发展趋势

从推理方法来说，知识图谱嵌入与预训练模型可以学习实体和关系的表示，本体嵌入表示可以学习类的层次关系以及各种公理和规则，图神经网络方法可以充分捕获图的邻接信息以及子图结构，这些仍将是当前知识图谱推理的研究热点。同时为了完成更高层次的推理，多种方法需要同时运用，进一步深挖更复杂的逻辑规则结构，因此神经符号集成的推理方法也将是未来的研究热点。

从应用角度分析，目前的知识图谱推理方法在中等规模的标准数据集上取得了不错的效果和明显的进步，但面对超大规模、低资源以及人机协作的应用依然面临众多挑战，例如在超大规模应用上提升推理方法的易用性和推理效率，在低资源应用中提升推理方法的鲁棒性和可迁移性，在人机协作应用中提升可解释性和人可介入性都是在把知识图谱推理实用化的道路上需要重点关注的问题。而在工业应用中，如何综合利用不同的推理将知识图谱中缺失的知识进行补全、将错误的信息进行矫正，从而提升意图识别、实体链接、实体推荐等任务的实际效果，仍是诸多工业界关注的核心问题。

9.3.5. 知识图谱资源

在知识图谱领域的研发中，知识资源既是一种主要的输出物，在一些任务中也作为重要的输入之一：知识资源是知识建模、获取与融合的目标产物，是知识推理与应用的数据基础，也是诸多自然语言处理任务的辅助资源。大规模知识资源的构建通常面临精确性（即知识资源的质量）和完备性（即知识资源的规模）如何平衡的问题，难免不精确、不完备。因此，在初次构建之后，知识资源仍需不断扩充和完善，持续提高质量和规模。

根据知识覆盖的范围不同，传统上将知识资源分为三类：世界知识、常识知识、语言知识，当然，这种分类并非严格正交。其中，世界知识描述的对象是现实世界；常识知识描述的对象是大部分人都知道的日常世界；语言知识描述的对

象是自然语言本身。这三类知识资源的构建、演化和利用存在一定差异。例如，万维网文本中存在大量的世界知识可供抽取，而常识知识则很少直接被提及。因此，相关研究面临的挑战和采用的技术也有所区别。以下分别简要回顾上述三方面知识资源的发展历史和前沿进展。

9.3.5.1. 技术发展脉络和进展

9.3.5.1.1. 世界知识

世界知识是关于现实世界的事实和具体事物的知识。世界知识的表示可追溯到 1956 年 Richens 提出的语义网络 [Richens, 1956]以及随后兴起的专家系统。1999 年初步形成的 RDF 和 OWL 等规范则作为语义网的推荐标准直至今日。2012 年谷歌提出的知识图谱引发了世界知识研究和构建的新热潮。目前，一些被广泛使用的大型世界知识图谱包括 DBpedia、YAGO、Freebase、Wikidata 等。

DBpedia 是一个由社区发起的从维基百科的信息框中自动提取结构化知识形成的 RDF 知识库 [Auer et al., 2007, Bizer et al., 2009, Lehmann et al., 2015]。至 2021 年，DBpedia 核心库包含约 9 亿条三元组结构的世界知识。YAGO [Suchanek et al., 2007, Tanon et al., 2020]与 DBpedia 在同一时期发起，是一个自动构建的世界知识库，其来源与 DBpedia 类似，并在此基础上与 WordNet 语言知识库相结合，具有一定的逻辑推理能力。Freebase [Bollacker et al., 2007]与 Wikidata [Vrandecic and Krötzsch, 2014]则都是由相应社区成员协同编辑形成的众包知识库。中文的世界知识在近十年同样有较多进展，如百科型的 Zhishi.me [Niu et al., 2011]、XLore [Wang et al., 2013]、CN-DBpedia [Xu et al., 2017]等，以及一些面向特定领域的知识库，例如地理领域的 CKGG [Shen et al., 2021]等。

对于世界知识库的构建，一些方法采用知识挖掘的思路，将结构化或半结构化的已有数据转化为知识图谱格式，并集成多个来源的数据，如上述提到的 CKGG。另一些方法采用信息抽取思路，以非结构化的文本作为输入，主要步骤包括实体识别、关系抽取、实体对齐等，其中关系抽取的研究尤其百花齐放。例如，近期的一些工作包括基于表填充和全局关联的方法 [Ren et al., 2021]、基于抽象语义表示的方法 [Zhang et al., 2021]、面向开放域的抽取 [Liu et al., 2021]、文档级抽取 [Ru et al., 2021]、跨文档抽取 [Yao et al., 2021]、零样本和少样本抽取 [Sainz et al., 2021]等。

9.3.5.1.2. 常识知识

常识知识是关于日常世界的知识，被大部分人所了解[Liu and Singh, 2004]。常识知识库的构建可以追溯到 1984 年由 Douglas Lenat 创建的 Cyc 项目 [Lenat and Guha, 1989]，其使用逻辑框架描述常识知识。为了在自然语言处理等任务中更有效地处理常识，2004 年构建的常识知识图谱 ConceptNet [Liu and Singh, 2004]已被广泛使用。近年新构建的常识知识库或直接从万维网文本抽取三元组结构的常识知识 [Tandon et al., 2014]，或以事件为中心 [Mostafazadeh et al., 2020]，侧重于常识事件之间的因果关系抽取。

在近期的一些工作中，Hwang 等人 [2021]将大规模预训练语言模型中的隐式常识知识转化为显式的知识库，构建了常识知识库 ATOMIC-2020；Fang 等人 [2021]将语言知识库 ASER [Zhang et al., 2020]自动转化为与 ATOMIC 形式类似的常识知识库。

9.3.5.1.3. 语言知识

语言知识是有关于自然语言本身的知识。WordNet [Miller, 1995]是普林斯顿大学在 1980 年代开始构建的语言知识库，其以词义为基本单元。受 WordNet 启发，后续的 FrameNet [Baker et al., 1998]以框架为基本单元描述语言知识。BabelNet [Navigli and Ponzetto, 2010]将维基百科链接到 WordNet 中，并通过机器翻译补充低资源语言知识。知网 [Dong and Dong, 2003]是有较大影响力的中文语言知识库，以义原作为最基本的语义描述单元。

近期的一些工作中，Wang 等人 [2021]采用一种基于框架的统一表示架构，集成了语言知识库 FrameNet、世界知识库 YAGO 和常识知识库 ConceptNet。WordNet 也常作为知识图谱补全的实验数据集，如 Bai 等人 [2021]利用双曲嵌入对 WordNet 进行补全。

9.3.5.2. 未来技术发展趋势

知识资源的构建已受到越来越多研究人员的重视，CIKM 和 ISWC 等国际会议常年设有资源类论文投稿，鼓励研究者构建和发布知识资源。由于知识资源对于若干下游任务的重要性，在可以预见的将来，知识资源研究热潮还将持续。

知识资源的研究重心有向“两高”迁移的趋势：高阶知识和高质量知识。一方面，除了表达实体关系的知识图谱，表达事件关系的事理图谱的相关工作越来越多。另一方面，如何验证和提高知识资源的质量，对知识资源的利用至关重要。

这些有望成为知识资源研究的新热点。

此外，如何定义知识资源将变得非常有趣。近年来，已经有大量研究表明：预训练语言模型中包含丰富的世界知识和常识知识，这些模型是否属于知识资源？知识资源一定是图谱等离散形式的吗？作为不同形态的知识资源，离散的知识图谱和连续的预训练语言模型如何有效结合形成互补效应？预训练语言模型能完全取代语言知识吗？随着大规模预训练语言模型的不断涌现和研究的持续深入，这些问题将进一步推动知识资源的发展。

9.4.知识图谱产业应用

知识图谱应用落地是系统性的工程问题，在经过大量的知识图谱研究与产业化落地实践后，业界逐步形成了行业知识图谱应用落地的全流程，称为行业知识图谱的生命周期。目前，在知识图谱生命周期中，也存在着知识图谱建设到应用周期过长，图谱构建过程难度较高，需要专业技能，跨项目，跨领域迁移成本高，数据、知识、模型、算法等可复用性程度低，应用构建复杂，需要技术人员深度开发等工程性的挑战。

为解决上述问题，在现阶段的工业级的应用场景中，国内外越来越多的企业和研究机构开始引入当前热门的平台化方案等相关技术，即围绕生命周期构建相应的行业知识图谱服务平台，在平台的基础上进行应用的构建，实现一个功能完整的知识图谱信息系统来支撑知识图谱的应用落地。

9.4.1. 知识图谱平台

知识图谱服务平台主要负责构建知识图谱和提供具体场景应用服务，将来自上游数据提供方的初步结构化数据进行信息抽取、知识融合、知识加工，逐步构建起知识图谱，再为下游最终用户提供具体场景下基于知识图谱的数据智能化应用服务，可显著提高各行业中知识图谱的落地效率和效果，应用领域包括金融、客服、工业、科研、医疗等。目前，国外主流的知识图谱平台有：Palantir 可拓展大数据分析平台、IBM Watson Discovery 服务及其相关产品所使用的知识图谱框架 Knowledge Studio、Oracle 知识图谱平台、Metaphactory 知识图谱信息系统解决方案平台，以及开源知识图谱项目 LOD2。

公司	平台名称	平台简介	主要特点（服务）
Palantir	Palantir 平台	Palantir 是用于知识图谱创建、管理、搜索、发现、挖掘和积累的可扩展的大数据分析平台	数据集成、搜索发现、知识管理、算法引擎、算法引擎
IBM	IBM Watson Discovery 知识图谱框架	Watson Discovery Services 使用该框架并提供相关服务，这些服务已经部署在 IBM 以外的许多行业配置中。	该框架直接支持 Watson Discovery，它关注于使用结构化和非结构化的知识来发现新的、不明显的信息，以及发现之上的相关垂直产品；该框架允许其他人以预先构建的知识图谱为核心构建自己的知识图谱。
Oracle	Oracle 知识图谱平台	Oracle知识图谱平台基于其自身多年的存储经验,在具有明显优势的存储层上进行构建,上层通过W3C标准的RDF和OWL组织和表示图谱,使用SPARQL对数据统一查询服务。	对数据存储与访问的支持性比较好,可以实现基于内存的并行图计算,提供许多工具完成从各种大数据平台、关系数据库到知识图谱的映射与转换。
Metaphacts	Metaphactory 平台	Metaphactory 提供了一套从知识存储、知识管理到知识查询与应用开发的端到端的知识图谱平台解决方案。	Metaphactory 主要针对结构化数据进行查询和管理，且兼容常见的知识图谱存储形式，实现不同数据源、不同格式的知识图谱混合查询，提供了搜索、可视化和知识编辑管理的接口,可用于知识图谱资产管理，快速应用程序构建和面向最终用户的交互。
Stardog	Stardog 平台	Stardog是一个企业级知识图谱平台,通过将数据转换成知识,使用知识图谱进行组织,对外提供查询、检索和分析等服务。	Stardog能够把关系数据库映射成虚拟图,并且支持OW2的推理和 Gremlin,但其仅对结构化数据(RDBMS Excel等)的处理,没有针对非结构化数据的知识抽取,也不具有知识融合功能。
/	LOD2 开源知识图谱项目	LOD2 是构建结构化链接数据的企业级管理工具和方法	提供一个搜索、浏览和生成链接数据的平台,其侧重于链接数据的生命周期管理,而对于其他类型的数据需要首先转换成链接数据。

表 1 国外知识图谱平台

近年来，国内知识图谱平台也发展迅速：华为知识图谱云提供了集本体设计、信息抽取、知识映射、多源融合以及增量更新等功能为一体的一站式知识图谱构建平台；腾讯知识图谱实现集成了图数据库、图计算引擎和图可视化分析的一站式解决方案；百度的 AI 开放平台面向公安、法律、金融、医疗等领域为客户提供从数据到知识应用的全方位服务；阿里巴巴基于其建立的多领域知识图谱“藏经阁”，研发知识引起产品，在淘宝、天猫、盒马鲜生等几十种产品上成功支撑商品、旅行、智能制造等多个领域的知识服务。互联网巨头正积极构建知识图谱，针对垂直领域构建知识图谱平台，促进知识图谱的发展与落地。

同时，传统解决方案商旗下知识平台和初创型知识服务平台以其在具体领域中的垂直深耕，并整合了知识图谱的设计、构建、编辑、管理、应用等全生命周期实现，在市场上也具有一定的竞争力。这类典型的知识平台有：明略知识图谱信息系统 SCOPA，其提供了基于知识图谱技术的知识管理和洞察分析平台，实现从客观数据汇聚到抽象知识沉淀的认知跃迁，为组织提供知识驱动的辅助决策；柯基数据的认知智能引擎提供全周期的知识图谱构建和运维管理平台，平台通过动态本体实现多源异构数据的知识获取与融合存储，可构建复杂的多模态知识图谱，提供从基础数据到知识管理、知识应用的全方位智能服务，赋能医药、军工、能源、金融等行业的数智化转型。PlantData 知识图谱管理系统（Knowledge Graph Management System，KGMS），以行业知识图谱全生命周期为理论指导，结合多行业、数十个项目实战经验，打造全流程一体化的管理平台。星环科技的知识图谱全场景解决方案，内置全套数据组件，使用 3D 空间图实现知识图谱的可视化，并提供了成熟的行业模板；渊亭 DataExa-Sati 认知智能平台，可帮助客户打造行业知识图谱，帮助企业快速生成成熟的解决方案；此外还有包括达观数据、

东软、北大医信、鼎富科技等等一批知识图谱平台提供商。企业级的知识图谱信息系统、知识工作自动化平台、知识图谱平台软件服务等方案相继被各厂商提出，正快速成为以知识图谱为核心的新一代信息系统发展的有力支撑。

公司	平台名称	平台简介	主要特点（服务）
百度	知识图谱开放平台	基于知识图谱、自然语言、搜索与推荐等核心技术，依托高效生产、灵活组织、便捷获取的智能应用知识的全链条能力，提供企业知识应用全生命周期一站式解决方案，助力企业提升效率、提高决策智能水平	数据引入、服务接入、知识生产与组织、平台化综合管理、知识搜索
腾讯	腾讯知识图谱（Tencent Knowledge Graph，TKG） 腾讯知识图谱一站式平台	腾讯知识图谱是一个集成图数据库、图计算引擎和图可视化分析的一站式平台。腾讯知识图谱用于构建和分析包含千亿级节点关系的知识图谱，并支持在图谱上搭建企业级应用服务	知识图谱自动构建、图谱在线查询、提供多种图计算模型、图数据可视化展现、图查询语言、独立部署
阿里巴巴	藏经阁 阿里巴巴知识图谱服务平台	以多源大规模数据为对象，研究从大数据向通用、领域知识转化的共性关键技术，研发并推出知识建模、知识获取、知识融合、知识推理计算和知识赋能的平台服务	通过实现知识建模、知识获取、知识融合、知识推理计算和知识赋能五个模块，提供从数据、信息、知识到知识服务一整套技术平台化服务，同时，特定领域知识图谱可插拔，特定领域知识图谱加载后，可以提供特定领域的知识服务
华为	华为云 知识图谱 KG	华为知识图谱是一款知识图谱构建工具，提供一站式知识图谱构建平台，提供本体设计、信息抽取、知识映射、多源融合以及增量更新等功能	本体设计，信息抽取，知识映射，知识融合，知识服务（知识图谱问答、智能文案系统、行业知识图谱解决方案、智能知识推荐）
明略科技	知识图谱信息系统 SCOPA	可视化数据分析平台，构建在明略自研知识图谱数据库 NEST 之上，实现知识图谱行业解决方案快速落。目前已应用到公共安全、金融、税务、工业等多个行业几百个项目中	关系网络分析、时空轨迹碰撞、实时多维检索、信息比对碰撞、智能协作系统、实时数据接入
柯基数据	KGDATA 知识图谱平台	柯基数据知识图谱平台通过动态本体实现多源异构数据的知识获取与融合存储，可构建复杂的多模态知识图谱，提供从基础数据到知识管理、知识应用全方位智能服务，已赋能医药、军工、能源、金融等行业多客户多业务部门的数智化转型	多模态知识图谱、动态本体构建、非结构化数据标注与训练、结构化数据增量更新、事件抽取、语义检索、智能问答、智能推荐
海义知科技	PlantData 知识图谱认知智能中台	KGMS：企业级知识图谱管理平台； KGBuilder：配置式自动化图谱构建工具； KGAssist：插件式知识服务助手； KGRobot：会话式图谱机器人开放平台； KGPro：统一知识图谱分析引擎；	关联分析、路径分析、图数据探索、图谱可视化、推理、自然语言检索、智能BI、语义标注
达观数据	达观智能知识图谱平台	基于客户的多源异构数据整合构建知识中台，为客户量身打造基于知识图谱的数据智能化应用，为制造、政务等行业客户提供业务场景智能升级服务	文本挖掘、智能推荐、垂直搜索、文档智能审阅、企业级搜索引擎、客户意见洞察、光学字符识别、机器人流程自动化、数据挖掘分析、文本审核
渊亭科技	DataExa-Sati 认知智能平台	渊亭Dataexa-Sai认知智能平台能够帮助客户打造行业知识图谱,采用分布式服务架构和自研分布式图计算引擎,实现行业级知识图谱构建和分析,从可视化知识建模、多源异构知识提取和知识融合、万亿级别高性能图存储计算引擎、复杂知识推理等角度,快速、精准地从知识图谱中提取出有价值的信息,帮助企业快速生成成熟的解决方案	聚焦金融、政务、国防、工业互联网四大行业，为客户提供认知中台、AI中台、数据中台三大中台产品及AI+行业解决方案，打通“数据-AI-认知”的闭环服务。
海致星图	海致星图金融知识图谱平台	海致星图金融知识图谱平台从零散数据中发现知识，帮助组织机构实现业务智能化	银行智能营运分析：自动化分析财务报表、外源文档、行内文档，提高银行运营决策、产品设计、营销推广、风险管理效率。
星环科技	知识图谱平台 Sophon KG	星环知识图谱平台（Sophon KG）是一款集知识的获取、融合、存储、计算以及应用为一体的自 研知识图谱产品。支持拖拽式图谱构建、知识抽取、知识存储、分布式图谱计算、知识推理以及图谱查询分析	零代码图谱构建、交互式图谱分析、文本标注、图算法数据挖掘、智能语义检索

表 2 国内知识图谱平台

9.4.2. 知识图谱行业应用

知识图谱于 2012 年被谷歌正式提出的初衷是为了改善搜索，基于谷歌知识图谱的搜索不是简单地返回网页的超链接，而是真正理解用户请求并将其链接到

现实世界认知概念的指代，然后返回指代的相关结果，可大幅度提升用户的搜索体验。截至目前，谷歌的知识图谱涵盖了广泛的主题，包括超过 10 亿个实体和 700 亿条事实。与之同时期的，微软必应（Bing）知识图谱也针对搜索场景，它包含了物理世界的知识，如人物、地点、事物、组织、位置等类型的实体，以及用户可能采用的行为。当用户输入搜索文本时，如果知识图谱中存在相关的知识时，必应搜索引擎将显示来自必应知识图谱的知识面板，可充分展示用户感兴趣的内容。领英图谱（LinkedIn graph）也是微软公司旗下的知识图谱应用，其中包括人员、工作、技能、公司、位置等实体，可实现更加有效的职场社交。脸书（Facebook）公司拥有全球最大的社交知识图谱，该图谱以用户为中心，同时包括用户关心的其他信息如兴趣爱好、从事行业等信息，基于图谱的知识资源可增加用户对脸书产品的体验，包括内容搜索和兴趣推荐等。在搜索及社交应用场景中，国内与国外相同，有相应的大型互联网厂商提出的知识图谱，例如百度、搜狗的面向搜索的知识图谱，以及面向社交场景的微博图谱。

经过近 10 年的发展，当前知识图谱的应用俨然远超其最初的搜索场景，由相对通用的搜索、问答、推荐等场景向核心业务决策过程转变。在行业应用方面，随着面向不同行业的知识图谱落地应用，以信息系统为载体的知识图谱典型应用（包括智能问答、推荐系统、个人助手等）也逐渐走进各个行业领域。

知识图谱在国外有着较为成熟的行业应用积累，如 IBM Watson 最早被研发应用于医疗领域，随着产品的不断延伸也逐步应用于金融等其他领域中。Palantir 相关产品已经分别应用于国防安全与金融领域，形成包括反欺诈、网络安全、国防安全、危机应对，保险分析、疾病控制、智能化决策等解决方案。国内人工智能及知识图谱在产业中地落地也呈现井喷得态势，知识图谱在国内的行业应用落地已经处于世界领先水平，在金融、情报分析、能源电力、医疗、工业、教育、政务、公安、营销和客服等场景均得到了广泛应用。

		知识图谱平台	知识图谱应用										
			通用	领域									
				公安领域	金融领域	能源领域	客服领域	医疗领域	教育领域	司法领域	营销领域	舆情领域	政务领域	工业领域
大数据智能公司	明略科技	√		√	√	√					√	√	√	√
	国双	√		√		√				√	√	√	√	√
	海致	√		√	√						√			
	百分点	√		√	√							√	√	√
	一览群智	√		√	√							√		
	海义知科技	√				√								√
	柯基数据	√			√	√	√	√	√	√			√	√
	蓝凌	√												
	文因互联	√			√									
互联网公司	阿里巴巴	√	√	√	√	√	√				√	√		√
	腾讯	√	√	√	√	√	√	√			√	√		
	百度	√	√		√	√	√	√						√
	京东数科	√					√				√		√	
	Google	√	√			√	√		√		√	√		
	amazon	√	√				√			√		√		
	美团	√	√		√		√				√	√		
	今日头条	√	√											
	搜狗	√	√											
AI 公司	科大讯飞	√						√	√	√				
	第四范式	√												
	松鼠AI	√							√					
	追一科技	√												

表 3 知识图谱行业应用

本节将以金融领域、工业领域、能源领域、医疗领域和电商领域的行业应用为主，介绍知识图谱的应用情况。

知识图谱在金融行业应用广泛。金融数据增速迅猛且包含各行业的数据信息，这些信息又以文字、表格、图形等形式存储在大量文档中，格式非标准统一且以碎片化存在，传统风控方法难以满足应用需要。凭借知识图谱强大的统一知识表示及数据存储的能力，通过构建如企业图谱、专利图谱、产业链图谱等金融领域知识图谱，实现多源异构数据的知识化整合。文因互联用深度语义分析技术，将非结构化金融文档转为知识图谱，并基于推理机和知识库管理系统技术，实现大规模金融知识建模和流程机器人，在上交所、北交所、投行、评级、资管多个场景成功落地。

知识图谱应用可拓展至传统制造业领域。在汽车行业设备故障维修流程中，存在高度依赖人工、缺乏辅助工具、排查周期长、故障重复发生等问题，通过解析汽车故障维修手册、产品手册、维修记录、FMEA 等资料，构建百万级设备故障图谱，利用自然语言处理、语义检索、事理推理等技术，形成一套汽车生产制造过程中多个环节的故障辅助诊断系统，系统支持根据故障描述，结合往期案例分析，自动定位故障位置、故障原因、推理排查措施，可有效提升故障维修效率，沉淀经验形成企业财富和核心竞争力。

随着电网信息化、智能化水平的不断提升，电力设备的功能较以往更加复杂，其日常的运行维护，包括故障诊断，也更加依赖于专门的电力知识。由于缺乏有效的电力知识提取、组织、管理、展示等技术，运维人员不得不依靠自身经验去诊断电力设备故障，不仅效率低，准确率也难以保障。PlantData 知识资源服务平台将知识图谱技术应用于电力能源领域，从已有电力技术文献中提取知识并建

立知识库，辅助运维人员开展电力设备故障诊断，最终大幅提高其工作效率，保障电网安全。

在医药大健康行业，需时刻了解国内外行业最新动态、情报和技术等，但行业存在大量的医学文献、医学论文、医学指南和企业内部的幻灯资料等都没有结构化，无法实现高效的检索和关联这些非结构化文档，也无法智能的发现一些关键知识来辅助医学人员做出决策。在产品销售阶段，学术营销、提供学术咨询目前也都是通过电话、邮件、会议等方式，效率低下，无法满足为患者服务的时效性。基于柯基数据的行业知识图谱认知智能引擎技术，可快速构建复杂的全科或者领域内知识图谱，实现医药大健康智能专家虚拟助理，辅助医学人员在医药情报发现、专家智能推荐、医学信息推广、患者健康管理和疾病用药专业知识咨询等领域，大幅度提升工作效率和智能化水平，最终实现企业的数字化转型。

美团点评 NLP 中心构建了大规模的餐饮娱乐知识图谱 —— 美团大脑。美团点评作为中国最大的在线本地生活服务平台，覆盖了餐饮娱乐领域的众多生活场景，连接了数亿用户和数千万商户，积累了宝贵的业务数据，蕴含着丰富的日常生活的相关知识。在建的美团大脑知识图谱已有数十类概念、数十亿实体和数百亿三元组。目前，美团大脑已经在搜索、金融等场景中初步验证了知识图谱的有效性，包括智能搜索、商户赋能以及金融风险管理和反欺诈。

9.4.3. 知识图谱应用发展趋势

知识图谱对于大数据智能具有重要意义，知识图谱已经在多个领域有了落地应用，产生了实际价值，但是知识图谱的作用主要还是体现在增强搜索、推荐和智能问答的效果。另外，大规模知识图谱在深度问答（特别是基于语义分析和推理的问答系统）、演化分析、对话理解等方面的应用还处于初级阶段，如何快速构建高质量指数图谱，利用知识图谱实现深度知识推理，以及提高大规模知识图谱计算效率，是当前知识图谱发展所面临的挑战。从知识图谱应用发展趋势来看，当前正在从通用知识图谱应用向领域或行业知识图谱应用拓展，如金融、医疗、公安、医疗、司法、电商等。与此同时，领域知识图谱表示与构建的标准化研究以及针对不同任务需求和应用场景而定制的特色知识图谱，使知识图谱应用的发展呈现出“标准化”、“特色化”、“开放化”、“智能化”的趋势。借助知识图谱强大知识库的深度知识推理能力和逐步扩展的认知能力，相关行业从业者将能够对特定的问题进行分析、推理、获得决策支持，从而使知识图谱在越来越多的领域找到能够真正落地的应用场景，在各行各业中解放生产力，助力业务转型。

9.5.总结及展望

知识图谱作为知识的一种承载方式，随着过去十年人工智能的浪潮，作为人工智能从感知智能走向认知智能的必要基础设施，越来越收到学术界和产业界的关注。纵观知识图谱技术、资源及产业应用的研究发展现状，以下问题将逐步成为下一步重点关注的研究问题:

“符号-数值”相结合的统一知识表示:现有知识图谱主要基于符号表示为基础，其优势是语义明确、歧义小，符号计算精确度高；但是缺点也显而易见，计算过程完全依赖于符号间的严格匹配，很难克服符号间语义鸿沟的影响，泛化能力差，不适用于大规模知识计算。基于数值表示可以将知识单元(实体、关系和规则)映射到低维的连续实数空间，方便计算，但是缺点也很明显，数值计算无法替代语义理解，其黑盒特性以及处理过程难以解释，当面对需要进行规划、推理等复杂任务（如学习求解问题的过程并进行推理）时，捉襟见肘。因此，如何利用符号表示和分布式表示各自的优点，建立数值表示与符号表示的相互转换，进而利用数值计算实现符号计算的模拟，是下一代知识图谱技术的一个有前景的研究问题。

面向大模型的知识探测：传统利用人工构造知识图谱存在需要花费大量人力，建立的知识图谱稀疏问题严重，而自动的知识图谱构建技术存在知识生成质量低而难以应用的问题。随着基于预训练语言模型的语言理解技术逐步成为研究热点，越来越多的研究者尝试采用大模型+提示学习的方式实现多类型的信息抽取。那么如何验证大模型是否可以学习到知识？学到何种知识？是一个值得探索的问题。这一问题的答案将有可能改变未来知识图谱构建的方式和手段。

常识知识获取：常识表示与获取是知识图谱构建的核心难点问题，已有工作多将现有多种类型知识（语言知识、事件知识、世界知识等）进行融合，尝试构建常识知识库。但是常识是如何定义？都有哪些类别？对于过程类（Knowing How）常识如何建模？如何抽取？是否可以结合多模态数据进行常识抽取？仍然是亟待解决的难点问题。

知识图谱共享与平台技术:随着信息技术从信息服务向知识服务的转变，知识图谱成为行业和应用领域中智能系统的基础设施，不同行业和应用表示的知识具有不同内容和特性。知识图谱虽然已经在语义搜索、问题系统、推荐系统等应用展示出一定的威力，但是基于知识图谱的应用研究远不止这些，如何进一步货站知识图谱应用场景和范围，建立知识图谱的共享与应用平台，是未来知识图谱一个研发方向。

综合上述分析，我们有理由相信，随着深度学习、自然语言处理、数据库、

语义网等相关技术的快速进展，知识图谱及其相关技术的应用前景将更加广阔。

9.6.参考文献

[Auer et al., 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of ISWC. 2007.

[Bai et al., 2021] Yushi Bai, Rex Ying, Hongyu Ren, and Jure Leskovec. Modeling Heterogeneous Hierarchies with Relation-Specific Hyperbolic Cones. In Proceedings of NIPS. 2021.

[Bellare, 2007] K. Bellare, P. P. Talukdar, G. Kumaran, F. Pereira, M. Liberman, A. McCallum, and M. Dredze, “Lightly-Supervised Attribute Extraction,” NIPS 2007.

[Berrendorf, 2021] Max Berrendorf, Evgeniy Faerman, Volker Tresp. Active Learning for Entity Alignment. In Proceedings of ECIR. 2021.

[Bizer et al., 2009] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. Journal of Web Semantics. 2009.

[Bollacker et al., 2007] Kurt D. Bollacker, Robert P. Cook, and Patrick Tufts. Freebase: A Shared Database of Structured General Human Knowledge. In Proceedings of AAAI. 2007.

[Bordes, 2013] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: Translating Embeddings for Modeling Multi-relational Data. NIPS 2013: 2787-2795

[Cao, 2019] Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, Tat-Seng Chua. Multi-Channel Graph Neural Network for Entity Alignment. In Proceedings of ACL. 2019.

[Cao, 2020] Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. Open Knowledge Enrichment for Long-tail Entities. In Proceedings of WWW. 2020.

[Chen, 2018] Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, Carlo Zaniolo. Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. In Proceedings of IJCAI. 2018.

[Chen, 2020] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, Enhong Chen. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In Proceedings of KSEM. 2020.

[Chen, 2020] Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, Ian Horrocks: OWL2Vec*: embedding of OWL ontologies. Mach. Learn. 110(7): 1813-1845 (2021)

[Dong, 2014] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In Proceedings of KDD. 2014.

[Dong and Dong, 2003] Zhendong Dong and Qiang Dong. HowNet - A Hybrid Language and Knowledge Resource. In Proceedings of NLPKE. 2003.

[Du, 2021] Li Du, Xiao Ding, Kai Xiong, Ting Liu, Bing Qin. ExCAR: Event Graph Knowledge Enhanced Explainable Causal Reasoning. In Proceedings of ACL 2021.

[Fang et al., 2021] Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. DISCOS: Bridging the Gap between Discourse Knowledge and Commonsense Knowledge. In Proceedings of WWW. 2021.

[Fu, 2019] Tsu-Jui Fu, Peng-Hsuan Li, Wei-Yun Ma. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In Proceedings of ACL 2019.

[Glaser, 2009] Hugh Glaser, Afraz Jaffri, Ian C. Millard. Managing Co-reference on the Semantic Web. In Proceedings of LDOW. 2009.

[Guo, 2019] Lingbing Guo, Zequn Sun, Wei Hu. Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs. In Proceedings of ICML. 2019.

[Gupta, 2016] Gupta A, Piccinno F, Kozhevnikov M, et al. Revisiting taxonomy induction over wikipedia[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 11-17 2016. 2016 (CONF): 2300-2309.

[Gupta, 2018] Gupta A, Lebret R, Harkous H, et al. 280 birds with one stone: Inducing multilingual taxonomies from Wikipedia using character-level classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[He, 2011] He Y, Xin D. Seisa: set expansion by iterative similarity aggregation[C] //Proceedings of the 20th international conference on World wide web. WWW 2011.

[Hearst, 1992] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992: 539-545

[Ho, 2018] Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov, Gerhard Weikum: Rule Learning from Knowledge Graphs Guided by

Embedding Models. ISWC (1) 2018: 72-90

[Huang, 2020] Huang J, Xie Y, Meng Y, et al. Guiding corpus-based set expansion by auxiliary sets generation and co-expansion[C]//Proceedings of The Web Conference 2020. 2020: 2188-2198.

[Huang, 2021] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, Jiawei Han. Few-Shot Named Entity Recognition: An Empirical Baseline Study. In Proceedings of EMNLP 2021.

[Hwang et al., 2021] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (COMET-)ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In Proceedings of AAAI. 2021.

[Ji, 2020] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, Philip S. Yu: A Survey on Knowledge Graphs: Representation, Acquisition and Applications. CoRR abs/2002.00388 (2020)

[Jiang, 2019] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V. Chawla, Meng Jiang. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. In Proceedings of KDD 2019.

[Jiménez-Ruiz, 2011] Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. LogMap: Logic-Based and Scalable Ontology Matching. In Proceedings of ISWC. 2011.

[Julie, 2014] Julie Weeds, Daoud Clarke, Jeremy Reffin, David J. Weir, Bill Keller: Learning to Distinguish Hypernyms and Co-Hyponyms. COLING 2014: 2249-2259

[Kipf, 2017] Thomas N. Kipf, Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of ICLR. 2017.

[Kulmanov 2019] Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, Robert Hoehndorf: EL Embeddings: Geometric Construction of Models for the Description Logic EL++. IJCAI 2019: 6103-6109

[Lajus, 2018] Lajus J, Suchanek F M. Are all people married? Determining obligatory attributes in knowledge bases[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 1115-1124.

[Lehmann et al., 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. Journal of Semantic Web.

2015.

[Lee, 2013] Lee, T., Wang, Z., Wang, H., Hwang, S.-W.: Attribute extraction and scoring: a probabilistic approach. In: ICDE, pp. 194–205

[Lenat and Guha, 1989] Douglas B Lenat and Ramanathan V Guha. Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project. Addison-Wesley Longman Publishing Co.,Inc., 1989.

[Lison, 2020] Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, Samia Touileb. Named Entity Recognition without Labelled Data: A Weak Supervision Approach. In Proceedings of ACL 2020.

[Li, 2020] Manling Li, et al. GAIA: A Fine-grained Multimedia Knowledge Extraction System. In Proceedings of ACL 2020.

[Li, 2020] Weizhuo Li, Guilin Qi, and Qiu Ji. Hybrid reasoning in knowledge graphs: Combing symbolic reasoning and statistical reasoning. Semantic Web, 11(1):53–62, 2020.

[Li, 2014] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, JiaweiHan. A Confidence-aware Approach for Truth Discovery on Long-tail Data. In Proceedings of VLDB. 2014.

[Li, 2019] Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, Tat-Seng Chua. Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model. In Proceedings of EMNLP. 2019.

[Li, 2019] Yuan Li, Benjamin I. P. Rubinstein, Trevor Cohn. Truth Inference at Scale: A Bayesian Model for Adjudicating Highly Redundant Crowd Annotations. In Proceedings of WWW. 2019.

[Liu et al., 2021] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. Element Intervention for Open Relation Extraction. In Proceedings of ACL. 2021.

[Liu and Singh, 2004] Hugo Liu and Push Singh. ConceptNet - A Practical Commonsense Reasoning Tool-Lit. BT technology journal. 2004.

[Liu, 2020] Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, Tat-Seng Chua. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In Proceedings of EMNLP. 2020.

[Liu, 2021] Fangyu Liu, Muhao Chen, Dan Roth, Nigel Collier. Visual Pivoting for (Unsupervised) Entity Alignment. In Proceedings of AAAI. 2021.

[Liu, 2021] Bing Liu, Harrisen Scells, Guido Zuccon, Wen Hua, Genghong Zhao. ActiveEA: Active Learning for Neural Entity Alignment. In Proceedings of EMNLP.

2021.

[Lyu, 2021] Shanshan Lyu, Wentao Ouyang, Yongqing Wang, Huawei Shen, Xueqi Cheng. Truth Discovery by Claim and Source Embedding. IEEE Transactions on Knowledge and Data Engineering. 2021.

[Mao, 2020] Xin Mao, Wenting Wang, Huimin Xu, Man Lan, Yuanbin Wu. MRAEA: An Efficient and Robust Entity Alignment Approach for Cross-lingual Knowledge Graph. In Proceedings of WSDM. 2020.

[Mao, 2021] Xin Mao, Wenting Wang, Yuanbin Wu, Man Lan. Boosting the Speed of Entity Alignment 10 ×: Dual Attention Matching Network with Normalized Hard Sample Mining. In Proceedings of WWW. 2021.

[Mao, 2020] Mao Y, Zhao T, Kan A, et al. Octet: Online Catalog Taxonomy Enrichment with Self-Supervision[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2247-2257.

[Miller, 1995] George A Miller. WordNet: A Lexical Database for English. Communications of the ACM. 1995.

[Mostafazadeh et al., 2020] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. GLUCOSE: GeneraLized and COntextualized Story Explanations. In Proceedings of EMNLP. 2020.

[Navigli and Ponzetto, 2010] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a Very Large Multilingual Semantic Network. In Proceedings of ACL. 2010.

[Niu et al., 2011] Xing Niu, Xinruo Sun, Haofen Wang, Shu Rong, Guilin Qi, and Yong Yu. Zhishi.me - Weaving Chinese Linking Open Data. In Proceedings of ISWC. 2011.

[Niu, 2020] Guanglin Niu, Yongfei Zhang, Bo Li, Peng Cui, Si Liu, Jingyang Li, Xiaowei Zhang: Rule-Guided Compositional Representation Learning on Knowledge Graphs. AAAI 2020: 2950-2958

[Pasca, 2007] M. Pasca, V. D. Benjamin, and N. Garera, “The role of documents vs. queries in extracting class attributes from text,” CIKM, 2007, pp.485–494.

[Pasternack, 2010] Jeff Pasternack, Dan Roth. Knowing What to Believe (when you already know something). In Proceedings of COLING. 2010.

[Ponzetto, 2007] Simone Paolo Ponzetto, Michael Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007: 1440-1445

[Qi, 2021] Zhiyuan Qi, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Yuejia Xiang,

Ningyu Zhang, Yefeng Zheng. Unsupervised Knowledge Graph Alignment by Probabilistic Reasoning and Semantic Embedding. In Proceedings of IJCAI. 2021.

[Ravi, 2008] S. Ravi and M. Pasca, “Using structured text for large-scale attribute extraction,” CIKM, 2008, pp. 1183–1192.

[Ren, 2020] Hongyu Ren, Jure Leskovec: Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. NeurIPS 2020

[Ren et al., 2021] Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. A Novel Global Feature-Oriented Relational Triple Extraction Model Based on Table Filling. In Proceedings of EMNLP. 2021.

[Richens, 1956] Richard Hook Richens. Preprogramming for Mechanical Translation. Mechanical Translation and Computational Linguistics. 1956.

[Ru et al., 2021] Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. Learning Logic Rules for Document-Level Relation Extraction. In Proceedings of EMNLP. 2021.

[Sainz et al., 2021] Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction. In Proceedings of EMNLP. 2021.

[Schlichtkrull, 2018] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, Max Welling: Modeling Relational Data with Graph Convolutional Networks. ESWC 2018: 593-607

[Shen, 2017] Shen J, Wu Z, Lei D, et al. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017: 288-304.

[Shen, 2018] Shen J, Wu Z, Lei D, et al. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2180-2189.

[Shen, 2020a] Shen J, Qiu W, Shang J, et al. SynSetExpan: An Iterative Framework for Joint Entity Set Expansion and Synonym Discovery[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 8292-8307.

[Shen, 2020b] Shen J, Shen Z, Xiong C, et al. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network[C]//Proceedings of The Web Conference 2020. 2020: 486-497.

[Shen, 2021] Shen J, Qiu W, Meng Y, et al. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 4239-4249.

[Shen et al., 2021] Yulin Shen, Ziheng Chen, Gong Cheng, and Yuzhong Qu. CKGG: A Chinese Knowledge Graph for High-School Geography Education and Beyond. In Proceedings of ISWC. 2021.

[Suchanek et al., 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In Proceedings of WWW. 2007.

[Sun, 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, Jian Tang: RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. ICLR (Poster) 2019

[Sun, 2020] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, Yuzhong Qu. Knowledge Graph Alignment Network with Gated Multi-Hop Neighborhood Aggregation. In Proceedings of AAAI. 2020.

[Sun, 2020] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, Chengkai Li. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. Proceedings of the VLDB Endowment. 2020.

[Tandon et al., 2014] Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. WebChild: Harvesting and Organizing Commonsense Knowledge from the Web. In Proceedings of WSDM. 2014.

[Tanon et al., 2020] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian M. Suchanek. YAGO 4: A Reasonable Knowledge Base. In Proceedings of ESWC. 2020.

[Théo, 2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, Guillaume Bouchard: Complex Embeddings for Simple Link Prediction. ICML 2016: 2071-2080

[Trisedya, 2019] Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang. Entity Alignment between Knowledge Graphs Using Attribute Embeddings. In Proceedings of AAAI. 2019.

[Vashishth,2020] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Partha P. Talukdar:Composition-based Multi-Relational Graph Convolutional Networks. ICLR 2020

[Vrandecic and Krötzsch, 2014] Denny Vrandecic and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM. 2014.

[Volz, 2009] Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov. Discovering and Maintaining Links on the Web of Data. In Proceedings of ISWC. 2009.

[Wang, 2015] Xianzhi Wang, Quan Z. Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, Xue Li. An Integrated Bayesian Approach for Effective Multi-truth Discovery. In Proceedings of CIKM. 2015.

[Wang, 2018] Zhichun Wang, Qingsong Lv, Xiaohan Lan, Yu Zhang. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In Proceedings of EMNLP. 2018.

[Wang et al., 2013] Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang,Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. XLore: A Large-Scale English-Chinese Bilingual Knowledge Graph. In Proceedings of ISWC. 2013.

[Wang et al., 2021] Chenhao Wang, Yubo Chen, Zhipeng Xue, Yang Zhou, and Jun Zhao. CogNet: Bridging Linguistic Knowledge, World Knowledge and Commonsense Knowledge. In Proceedings of AAAI. 2021.

[Wang, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen: Knowledge Graph Embedding by Translating on Hyperplanes. AAAI 2014: 1112-1119

[Wang, 2017] Quan Wang, Zhendong Mao, Bin Wang, Li Guo: Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Trans. Knowl. Data Eng. 29(12): 2724-2743 (2017)

[Wang, 2007] Wang R C, Cohen W W. Language-independent set expansion of named entities using the web[C]//Seventh IEEE international conference on data mining (ICDM 2007). IEEE, 2007: 342-350

[Wu, 2019] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, Dongyan Zhao. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. In Proceedings of IJCAI. 2019.

[Xu et al., 2017] Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and YanghuaXiao. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. In Proceedings of IEA/AIE. 2017.

[Yan, 2019] Yan L, Han X, Sun L, et al. Learning to bootstrap for entity set expansion[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019

[Yao et al., 2021] Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou,

and Maosong Sun. CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild. In Proceedings of EMNLP. 2021.

[Yin, 2008] Xiaoxin Yin, Jiawei Han, Philip S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. IEEE Transactions on Knowledge and Data Engineering. 2008.

[Yu, 2021] Donghan Yu, Yiming Yang, Ruohong Zhang, Yuexin Wu. Knowledge Embedding Based Graph Convolutional Network. In Proceedings of WWW. 2021.

[Yu, 2019] Yu J, Wang C, Luo G, et al. Course Concept Expansion in MOOCs with External Knowledge and Interactive Game[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4292-4302.

[Yu, 2020] Yu J, Wang C, Luo G, et al. ExpanRL: Hierarchical Reinforcement Learning for Course Concept Expansion in MOOCs[C]//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020: 770-780.

[Zeng, 2020] Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, Zhen Tan. Degree-Aware Alignment for Entities in Tail. In Proceedings of SIGIR. 2020.

[Zeng, 2021] Weixin Zeng, Xiang Zhao, Jiuyang Tang, Changjun Fan. Reinforced Active Entity Alignment. In Proceedings of CIKM. 2021.

[Zhang, 2019] Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, Yuzhong Qu. Multi-view Knowledge Graph Embedding for Entity Alignment. In Proceedings of IJCAI. 2019.

[Zhang et al., 2020] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A Large-scale Eventuality Knowledge Graph. In Proceedings of WWW. 2020.

[Zhang et al., 2021] Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. Fine-grained Information Extraction from Biomedical Literature based on Knowledge-enriched Abstract Meaning Representation. In Proceedings of ACL. 2021.

[Zhang, 2021]Wen Zhang, Chi Man Wong, Ganqiang Ye, Bo Wen, Wei Zhang, Huajun Chen: Billion-scale Pre-trained E-commerce Product Knowledge Graph Model. ICDE 2021: 2476-2487

[Zhang, 2019] Wen Zhang, Bibek Paudel, Liang Wang, Jiaoyan Chen, Hai Zhu, Wei Zhang, Abraham Bernstein, Huajun Chen: Iteratively Learning Embeddings and

Rules for Knowledge Graph Reasoning. WWW 2019: 2366-2377

[Zhang, 2020] Yuyu Zhang, Xinshi Chen, Yuan Yang, Arun Ramamurthy, Bo Li, Yuan Qi, Le Song:Efficient Probabilistic Logic Reasoning with Graph Neural Networks. ICLR 2020

[Zhang, 2016] Zhang Z, Sun L, Han X. A joint model for entity set expansion and attribute extraction from web search queries[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016

[Zhang, 2018] Zhang C, Tao F, Chen X, et al. Taxogen: Constructing topical concept taxonomy by adaptive term embedding and clustering[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[Zhang, 2020] Zhang Y, Shen J, Shang J, et al. Empower Entity Set Expansion via Language Model Probing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8151-8160.

第十章　医疗信息处理技术研究进展、现状及趋势

10.1. 研究背景与意义

近年来，随着以深度学习为主要驱动力的人工智能技术在各个领域得到日渐深入广泛的应用，医疗领域的智能化同样成为研究者关注的重点领域之一，同时也是国务院 2017 年发布的《新一代人工智能发展规划》的重点应用领域。包括人工智能技术以及区块链、云计算、大数据、隐私计算和 5G 等各类新兴信息技术在医疗健康领域的应用，涵盖了从医疗数据的规范化处理、数据隐私保护，到知识的获取与推理、辅助诊疗、疑难疾病会诊等各个医疗信息处理与应用层面。

在数据处理层面，医疗文本作为医疗健康领域最重要、最复杂的数据之一，其处理技术研究重点逐渐从之前简单的医疗实体抽取、实体规范化等面向数据标准化需求的技术，过渡到复杂实体抽取、实体关系挖掘与知识图谱构建等更贴近临床辅助决策需求的技术方向上。在临床辅助决策推理上，除了基于视觉处理技术的阅片等临床辅助决策系统外，直接从电子病历等医疗数据出发构建的基于深度学习的临床诊断能力也在肺炎等疾病上获得初步验证。新冠疫情的爆发，给医疗健康领域技术发展带来极大挑战，快速应对大规模流行性疾病爆发所带来的极端挑战，加速了人工智能技术在流调、基于大数据的流行性预测等技术的成熟。作为构建在敏感隐私数据基础上的智能技术，隐私保护一直是医疗领域所关注的焦点之一。除了隐私数据的去隐私化技术外，包括联邦技术等隐私计算技术也成为医疗领域智能化的研究热点。

在智慧医疗各层面技术快速发展的同时，我们也看到，智能信息处理技术在医疗领域应用的深度和广度还很有限，无论是从临床复杂数据处理，还是临床辅助决策技术的可解释性、鲁棒性等方面都还刚刚起步。在隐私数据的广泛应用方面，也还需要包括国家政策在内的各方面基础环境的构建与支撑。尽管如此，医疗健康信息处理技术的持续发展，智能技术、信息技术与医疗健康应用场景深度融合已成为大趋势。

纵观医疗健康领域的各类应用场景，总结医疗健康信息处理技术，厘清未来技术发展方向，对医疗健康领域的科学研究和产业升级具有重要的指导意义。报告将从数据、技术和应用场景等多个维度的现状、面临的主要挑战和可能的发展方向展开叙述。健康医疗数据包括多源、异构、时序、多模态、多层次的表格（如个人信息、检验结果）、文本（医学影像报告、出入院小结等）、图形（心电图、脑电图等）、图像（B 超、CT、MRI 图像等）、视频（如胶囊内镜视频图像）、组

学（基因、蛋白质等）等。医疗健康信息处理技术主要包括医疗文本分析技术、医疗知识图谱构建、医疗辅助决策、医疗领域的隐私计算、多模态数据分析等各个方面（医学影像学和生物信息学因单独属于一个较大研究领域分支，在此不做详细介绍）。由于应用场景细分类别繁多，报告中主要涉及到医疗服务、医学研究和医学教育等方面。

10.2. 领域发展现状与关键科学问题

10.2.1. 医疗健康文本分析

文本是丰富的信息和知识载体，在医疗健康领域，文本广泛存在于电子病历、临床试验报告、医学指南、医学产品说明书、医学文献、社交多媒体等不同形式的记录中。如何分析、挖掘和利用医疗健康文本中的医疗健康信息和知识，为医疗健康服务、医学研究和健康医疗知识普及等提供支撑，是医疗健康文本分析的主要目的。

近年来，医疗健康文本分析技术因医疗健康自然语言处理社区，如 I2B2（Informatics for Integrating Biology and the Bedside）、N2C2（National NLP Clinical Challenges）、CLEF (International Conference of the Cross-Language Evaluation Forum for European Languages)、CCKS（China Conference on Knowledge Graph and Semantic Computing）和 CHIP（China Conference on Health Information Processing）等，持续举办了包括不同粒度、不同维度的医疗健康文本信息抽取、知识图谱构建、医疗健康问答、语料健康文本语义相似度计算、医疗健康文本生成和病人队列选择等在内的评测任务，大大推动了医疗健康文本分析技术的快速发展。**相关技术逐渐从面向简单基础的分析任务转向面向贴合医疗健康应用场景的复杂分析任务。**以医疗健康信息抽取为例，以前往往仅关注简单的连续医疗健康实体识别任务，目前针对不同形态的复杂医疗健康实体（连续、非连续、嵌套、重叠等）识别任务的研究越来越多，医疗健康事件逐渐受到关注，复杂医疗实体/事件的关系抽取的相关研究也开始出现。

随着健康医疗文本分析技术逐步深入到实际健康医疗场景，所面临的医疗文本分析任务越来越复杂，涉及到的技术环节也越来越多，急需能打破医疗健康文本分析任务类型差异的通用型医疗健康文本分析技术。在这方面，结合医疗健康数据特点的大规模预训练“语言”模型和将不同任务进行统一表示和建模方法已开始有一些尝试性研究工作。

10.2.2. 医疗知识图谱构建

医疗领域的知识图谱可以帮助理解复杂的生物系统和病理，已开始在医学实践和研究中发挥关键作用。如果我们把知识图谱的数据来源和所面向的表示对象分为临床诊疗和医学文献、书籍等，医疗领域的知识图谱可以分为医疗知识图谱和医学知识图谱两大类。知识图谱的构建需要我们从大规模医学数据中自动抽取事实知识，并学习知识间的联系，涉及到的关键技术包括多模态异构医学信息抽取、知识的不确定性学习和知识验证等。

对于多模态异构医学信息抽取的主要挑战首先在于大规模非结构化或半结构化医学数据计算机难以理解，从而增大了自动挖掘知识的难度；其次，对于从多组学数据中获得的异构、特定领域的信息，如基因表达、化学结构等异构信息，如何融入现有知识图谱，从而在图谱中同时保留语义和生物医学特征，并互补地来解决决策问题仍然具有较大挑战。

对于知识的不确定性的学习，由于知识的复杂性质和图谱构造的特定要求，简单将图谱建模在以三元组为基础的结构关系上，虽然简化了人工智能系统对复杂知识的采集与应用难度，却同时失去了知识的复杂关联，其挑战在于很难对知识的不确定性进行量化，需要在超越三元组的更大粒度知识要素的表示与获取上积极探索与尝试，以寻找到更加有效的模型与方法。

对于知识验证任务来说，其挑战在于不同来源的知识可能对相同主题做出相互矛盾的陈述，这种分歧在研究领域是普遍存在的现象，伴随着医学数据的大量增长，知识的验证和整合逐渐成为迫切需求。

10.2.3. 临床诊疗辅助决策

运用计算机技术辅助临床诊疗决策，一直以来是人工智能领域极具挑战的研究热点。由于电子病历中记录了患者丰富的临床信息，近些年，基于医疗电子病历数据的临床诊疗辅助决策研究成为研究热点，如通过使用大规模深度预训练模型对电子病历进行表示，然后对国际疾病分类码(ICD code)进行预测，以及通过融合预先构建的医疗知识图谱增强预测的准确性和鲁棒性等。这些工作的显著特点是面向常见病，基于大规模数据，例如广州市妇女儿童医疗中心与依图医疗基于 1 亿多儿科患者转诊记录的数据，在儿科疾病上取得了重要的研究成果。当前基于深度学习和大数据驱动的辅助临床诊疗决策研究的缺点也非常明显，**主要挑战是可解释性差，鲁棒性低，缺乏推理过程，在小病种和精细诊疗上决策效果差等**。对于未来亟待解决的科学问题也逐渐清晰，主要集中在一下几个方面。

✓ 如何实现基于因果推理诊疗辅助决策，临床诊疗本质上是基于诊疗知识、诊

疗逻辑以及临床证据的，连续因果推理和决策过程。当前的模型和方法对诊疗知识图谱的使用依然不够充分，以三元组为组织形式的知识组织形式普遍缺乏复杂的因果逻辑。

- ✓ 如何实现诊疗决策过程的可解释性和鲁棒性，临床诊疗事关病人的生命健康，我们无法信任一个决策过程不透明，轻微扰动下决策结果就会改变的决策系统。当前的临床辅助决策模型高度依赖深度学习模型，而解释性差、鲁棒性低是当前深度学习模型的显性问题。
- ✓ 如何解决智能诊疗模型数据规模依赖性高，小病种决策准确率低的问题。由于医疗数据具有高度的敏感性，且不同医疗机构的数据质量参差不齐，疾病的种类也名目繁多，对每种疾病获取足够的，高质量的数据非常不现实。实际应用亟待需求低资源下的高质量决策模型。

10.2.4. 医疗领域的隐私计算

随着医疗病例的信息化和生物数据获取技术的发展，生命医疗研究的各个领域都积累了海量的临床数据。然而这些数据存储在多个部门或机构中，由于生命医疗数据高度敏感，很难在多机构或部门中共享，形成数据孤岛。目前中文信息处理技术及其他人工智能技术在生命医疗研究领域已经取得了一定的进展，但由于数据隐私导致数据无法共享的问题，一直限制整个应用的进一步发展。随着人工智能技术在生命医疗领域的不断渗透和发展，面向生命医疗大数据共享计算的需求日益增强。在数据隐私保护更为重视的今天，如何实现生命医疗大数据的安全共享计算，成为当下生命医疗研究发展主要研究问题之一。目前主流研究方向和技术包括数据脱敏、同态加密、安全多方计算以及联邦学习：

1） 数据脱敏是对数据中包含的敏感信息进行标定和处理，以达到数据变形的效果，使得恶意攻击者无法从已脱敏数据中获得敏感信息。数据脱敏技术主要方法有 k-匿名算法，差分隐私算法和基于生成对抗网络的方法。当前数据脱敏技术已经得到了初步的应用，在医疗数据的共享方面也有一定程度的探索。但是 k-匿名的泛化技术可能会使得数据过于泛化而没有区分度，从而使处理过后的数据失去了其原本的价值。差分隐私算法会对数据加入过量的噪声，从而会导致数据的可用程度下降。生成对抗网络可以自动提取数据集的重要特征并生成与原数据分布接近的新数据集，但在小数据集上不能很好的学习到数据集分布。

2） 同态加密技术允许用户直接对密文进行特定的代数运算，得到的结果仍然是相应明文进行对应操作之后得到的密文结果。可以实现医疗数据安全共享的同时能够保证病人隐私不被泄露。现在同态加密技术已经初步应用在了小体量医疗数据的共享当中。 现有的同态加密技术仍然有很多问题存在，比如使用同态加密之后密文的高效搜索比较困难，部分同态加密技术不能满足实际需求，而全同态加密存在噪音问题且计算效率低下等等。

3) 安全多方计算中，参与方将各自的秘密数据输入到一个约定函数进行协同计算，即使在一方甚至多方被攻击的情况下，安全多方计算仍能保证参与方的原始秘密数据不被泄露，并且保证函数计算结果的正确性。近年来，各种基于安全多方计算的技术都有了相应的系统实现，而且在方案的实用性和准确性方面也取得了很大的进步，但是仍然有许多问题亟待解决。**现有安全多方计算系统的 准确率和性能并不高，并且系统的安全性也有待提升。**虽然该技术已经初步应用于生物医疗数据的共享之中，但是其缺点在于节点之间的通信开销很大，在小体量生物数据面前，该方法勉强可行。

4) 联邦学习使得用户原始数据在不出本地基础上便能得到一个更优化的模型，做到“数据不动模型动”，在保证用户数据隐私安全的前提下，打破数据孤岛，充分挖掘数据中潜在价值。目前联邦学习已经初步应用于医学成像、生物组学以及医疗文本等领域。在生命科学研究中也已经有初步的研究成果。但目前联邦学习也存在一些问题：数据非独立同分：Non-IID 数据时会造成训练过程难以收敛、模型精度下降等问题。模型训练效率问题：客户端与服务端交互梯度参数，在通信中耗时连接不稳定问题容易导致模型学习低下。模型安全问题：在联邦学习的模型训练时，客户端与中央服务器之间传递梯度参数会造成隐私泄露，进而攻击者可以利用被泄露的梯度来恢复出用户的原始数据信息。

10.2.5. 复杂疾病的个性化临床诊疗决策

复杂疾病是指受多个基因调控，但不遵循孟德尔遗传定律的一类疾病。这类疾病往往存在多种中间表型、遗传模式难以确定、致病基因的外显率不全、且每一个致病基因单独致病作用十分微小，多个基因间又存在交互作用，是遗传、环境及其它非遗传因素共同作用的结果。生活中一些常见的疾病如原发性高血压、帕金森综合征、非胰岛素依赖型糖尿病以及多数恶性肿瘤等疾病都属于复杂疾病。医生很难在短时间内充分全面地考虑每一种因素，进而做出最佳的个性化诊疗方案。

复杂疾病的“个性化的诊疗”高度依赖患者从组学、细胞、器官到个体的多模态跨尺度的全面信息支持，包括个人基本信息、疾病症状、病史、生活习惯、医学检查结果、基因组数据、转录组数据，以及蛋白质组数据等。复杂疾病的个性化临床诊疗决策需要融合多种健康数据，并从中挖掘重要信息为临床诊断决策提供参考，有利于降低漏诊、误诊率，提高复杂疾病诊断的效率和准确率，以及为后续治疗提供参考。例如，疾病易感性检测需要综合考虑患者的病史、家族遗传史、生活习惯、生物组学等多种因素。对癌症病人的药物敏感性检测时，需要

利用基因组中 SNP 作为分子遗传标记，进行全基因组比较分析，识别与疾病相关的遗传变异，并利用药物基因组学数据研究基因如何影响个体对药物的反应，建立基于个体遗传标记的定制药物疗法。

个性化诊疗决策发展首先要解决的问题是构建一个包含患者的基本信息、检查检验结果、以及多组学信息在内的多模态跨尺度的个性化医疗数据库，其中的难点在于医疗健康领域的知识依赖远比大部分信息处理领域要强，跨尺度的个性化医疗数据库构建需要从医学、生物、化学等多个差异较大的学科领域抽取目标数据，涉及多种跨学科信息处理技术。

其次，个性化医疗数据和通常自然语言处理领域的文本有较大差别，通用的文本信息分析技术无法直接应用于个性化诊疗信息处理，鲁棒的跨尺度多组学数据融合分析方法仍然欠缺。另外，复杂疾病的亚型类别多、患者差异大，如何建立不同组学数据和疾病表型的关联是一个难点。

10.2.6. 基于大数据的重大流行性疾病防控

基于大数据的流行性疾病分析预测是通过分析流行性疾病的相关因素、监测早期发现流行性疾病发生的异常先兆或事件发展的不良趋势，进而对流行性疾病进行有效的预测和预警，提高流行性疾病预防控制工作的主动性和预见性，是保障公众健康和政府公共卫生管理的重要任务。涉及到两个关键的科学问题：首先，**发现并充分利用能够尽可能提前预测到特定流行病爆发前兆的新的因素**，尤其对能够反映与健康相关的群体关注点变化、能够分析社会效应的社交媒体等渠道的分析利用；第二，**丰富和完善多模态大数据融合分析的模型和技术**，从而能够充分结合社交媒体大数据、医疗监测大数据、环境气候大数据等多种因素来联合进行流行性疾病的分布和发展的分析预测。

就一般性流行性疾病而言，因疾病严重程度不高，流行性疾病的发现、监测和发展趋势预测是重点关注的研究方向；而对重大流行性疾病而言，因疾病严重程度高，防控策略对疾病发生和发展的影响的研究显得尤为重要。基于大数据的重大流行疾病防控需要更广泛的数据支持，如地理信息数据、公安数据、人口数据、城市建筑数据等。分析方法需要结合不同的应用场景，融合不同类型的模型（如 SIR（Susceptible Infected Recovered Model）模型和深度学习模型）。近两年来，因新冠病毒的仿佛和防控的常态化，相关研究工作得到了长足的发展。

10.3. 领域关键技术进展及趋势

10.3.1. 大规模多模态医疗预训练模型

定义：大规模多模态医疗预训练模型的研究旨在利用来自海量医疗电子记录设计符合数据特点的多模态数据通用表示方法、构建相应的表示模型，以支持不同应用场景下的各类医疗数据分析任务。一般来讲，医疗电子记录中包括丰富的结构化表格数据、文本、图形、图像等不同模态的数据，这些数据之间存在紧密的关联关系，如何在分别表示各种模态数据的同时，考虑各模态之间的关联关系，是大规模多模态医疗预训练模型研究需要重点关注的问题。此外，大规模多模态医疗预训练模型研究需要考虑某些模态缺失或者模态数据不全情况下的预训练方法，以及通过医疗健康领域知识与医疗数据融合来增强数据表示的方法。大规模多模态医疗预训练模型主要需要解决的问题包括：模态间数据的异质性问题和模态间数据/信息的对齐问题。

目标：利用海量医疗电子记录中的多模态医疗数据，通过充分挖掘不同模态医疗数据之间的关联关系，实现多模态医疗数据的通用表示，以支撑各种下游任务。这些表示可以是模态级的、记录级的、病人级的，甚至是人群级的。

进展：在自然语言处理领域预训练语言模型已经获得了很大的成功，以 BERT 为代表的语言模型层出不穷，也衍生出一些医疗领域的预训练模型，如 ClinicalBert，BioBert，MC-Bert，GBert，Med-Bert 等，其中 ClinicalBert、BioBert 和 MC-Bert 是基于医疗文本数据设计的预训练模型，Gbert 和 MedBert 则是基于结构化表格数据设计的预训练模型。这些预训练模型均是基于单个模态数据的预训练模型。而医疗数据天然是多模态的，为充分利用医疗数据，需要从多模态的角度进一步研究医疗领域的预训练模型。实际上，在通用领域，基于 Transformer 的多模态预训练模型已有一定的发展，如利用文本和图像数据的 VLBERT、LXMERT 等；医疗领域的大规模多模态预训练模型正虚位以待。与通用领域相比，医疗领域的大规模预训练模型需充分考虑医疗数据的多源、异构、不规则时序、冗余等多方面的特点和不同应用场景的需求特点，形成领域特有的预训练模型。

影响：大规模多模态医疗预训练模型为多模态医疗数据的融合提供了统一的范式，能大大简化医疗数据分析和处理流程和方法，将大大提供下游任务的整体性能，促进各类医疗健康数据分析技术的快速发展和落地。

发展：大规模多模态医疗预训练模型的研究大致可能存在以下几个方向：1、基于大规模双模态医疗数据的预训练模型，主要包括面向医疗文本和结构化表格

数据，以及面向医疗文本和医学图像的两种形式；2、基于大规模丰富模态医疗数据的预训练模型，将考虑超过两个模态融合的预训练模型；3、面向不同层次表示的大规模多模态预训练模型，可以从模态级的、记录级的、病人级的，甚至是人群级的进行考虑；4、融合医疗健康领域知识的大规模多模态医疗预训练模型；5、基于大规模多模态医疗预训练模型的少样本学习，以缓解医疗健康数据标注难度大和成本高的问题。随着大规模多模态预训练模型的研究工作的持续推进，医疗健康数据的利用必将上到一个新的台阶。

10.3.2. 开放的中文复杂医疗知识图谱构建

定义与目标：构建大规模开放中文医疗知识协同平台，促进国内的医产学研深度合作，构建具有因果关系的多层级中文医疗知识表示规范，构建起全国最大的多模态中文医学知识图谱，建立具有国际影响力的中文医学知识开源共享平台，由国内高水平医疗机构广泛参与、形成高质量医学知识共建关系，突破智慧医疗技术发展的知识瓶颈，在此基础上，开展大规模中文概率性、条件性、因果性医学知识图谱自动构建、知识更新、知识质量验证技术，为可解释、鲁棒性的诊疗模型提供支撑。

进展：知识图谱是包含多种关系的有向图，其中节点表示实体，边表示它们的关系。从一些高质量的人工标注的生物医学知识库开始，如 TTD、PharmGKB、OMIM 和 DrugBank，基于知识网络的研究已经迅速发展到利用更大的数据集进行下一代网络分析算法。然而，它们中的大多数只识别实体对之间关系的存在，而不包含特定的语义关系类型，因此不能被视为严格意义上的知识图谱。同时，许多最近开发的生物医学知识图谱，例如 CMeKG，CPubMed-KG 等，利用自然语言处理与文本挖掘技术，基于大规模医学文本数据，以人机结合的方式研发的中文医学知识图谱。现有的知识图谱构建往往忽略了知识的不确定性。知识与不确定性之间看起来具有不可调和的矛盾，但事实上却是一个有机融合体。在医疗领域，这种看似矛盾的融合从权威的专家构建的、医生诊疗时所需要遵循的疾病诊疗指南，到由经验丰富的专家公开发表的、用于诊疗参考的临床诊疗共识等，都充满了事实性与不确定性。这些包含了不确定性的知识，在医生给出诊断结论时具有非常重要的指导作用。之所以这种知识的不确定性在人类决策中并没有形成不可逾越的障碍，很大程度上得益于每个做出决策的人往往都具有丰富的隐式背景知识，这些背景知识能够有效地帮助人们判断在不同条件下的不确定性的度，从而做出一个相对优的选择。然而，在已有的知识表示体系中，给到人工智能系统利用的知识，往往是确定性的，缺失了不确定性特征，使得知识图谱的应用丧

失了鲁棒性。

发展与影响： 对于一个医学知识图谱来说，如何获得知识中真实有效的不确定性、如何把这种不确定性用量化的度显性的表达出来，就成为基于知识的不确定性推理所首先要面对的关键问题。知识通常以文本的形式给出，其不仅包含了知识的描述信息，同时还蕴涵着不同知识间的内在逻辑关系，这些逻辑关系对于知识的不确定性推理尤其重要。随着生物医学数据的大量增加，其中一些似乎与先前关于相同知识的断言存在不一致或矛盾。解决上述知识的确定性与鲁棒性之间的矛盾，对于增进我们对人类疾病的理解和开发有效的治疗方法至关重要。新一代的信息抽取、关系发现等技术可以通过从文献中提取声明、标记那些潜在矛盾的内容并识别可能解释此类矛盾的任何研究特征来促进此类分析。同时，多模态数据在形式上呈现异构性，而在内容上则呈现互补性。由于每一种模态都具有独特的结构，因此多模态联合学习的主要任务是将其蕴含的信息和其结构进行解耦，通过自监督学习的方式，在尽量充分保留其有效信息的同时，以一种统一的规范进行表达和建模，为知识的验证、更新与应用提供更可靠的保证。

10.3.3. 临床诊疗中的因果知识获取与融合

定义和目标：一般来说，临床诊疗的因果知识获取和融合的主要目标之一是结合先进的文本分析技术抽取包括医疗本体、临床记录等多种文本资源中的因果知识体系，为良好的智能诊疗体系打下坚实的基础。同时在获取的因果知识基础之上，结合先进的知识融合技术结合已有的因果知识，建立高效可解释性强的诊疗体系和系统。无论对于人还是机器来说，因果知识代表了一种底层的知识基础，为有效的推理和决策提供了重要的支撑，尤其在需要复杂推理的可循证临床诊疗领域，比如潜在的发病机制和病理生理学知识的因果模拟，它有着重要的意义，受到了越来越广泛的关注。

进展：为了推动该方向的发展，国内外学者已经做了一些比较丰富的研究，其中包括从医学文献或电子医疗记录中挖掘因果、建立因果图以及匹配因果链，从医疗本体中捕捉因果知识，从临床文本中建立因果图进行诊断决策等。在因果图自动构建方面，有研究者使用先进的文本分析技术从医疗文献文本中和电子病历记录的观测数据中自动构造了医疗状况（疾病）之间的因果图，弥补文本数据包含许多薄弱或不确定的因果关系的不足，为其他研究者提供了此类任务的一个开放基准。除了学习疾病之间的因果图，也有学习其他类型的因果图的研究，比如从电子病历记录中学习疾病相关症状的因果图。在知识主动更新方面，有研究者结合匹配技术和专家反馈，进行了一种基于主动迁移学习更新的从临床文本中

挖掘因果关系的识别框架研究，提供有助于总结临床文本以供重新应用医疗程序和预测分析、发现来自大量数据源的医学因果知识。在因果链映射方面，有研究者提出了一个框架 EHR2CCAS，此框架访问异构电子健康记录（HER）以估计给定时间窗口内患者异常状态的因果链（CCAS）中异常状态的存在，用于构建一个将 EHR 数据映射到 CCAS 的系统。在医疗本体挖掘因果方面，有研究者提出了一种提取和分析包含在两个权威医学本体 (AMO) 中的因果知识的新方法，并测试了专业知识的准确性。在因果知识应用诊断方面，有研究者在以动态不确定因果图理论进行眩晕、黄疸等疾病诊断，比如以呼吸困难为主诉的患者上建立疾病知识库，在具有循证医学思维和丰富诊断经验的临床医师专家的支持下，结合动态不确定因果图理论，构建并应用人工智能辅助诊断模型，弥补全科医生诊断经验的不足，缩短诊断时间。也有研究者结合贝叶斯网络和实体感知卷积神经网络，探索构建一个准确但可解释的诊断系统。

影响和发展：目前，先进的神经网络机器学习方法也已可以提供较高的准确性，那么因果知识所提供的可解释的医疗经验知识便显得越来越重要。临床诊疗的知识获取和融合的研究方法和理论已经有了一定的研究基础，在此之中因果知识已逐渐受到越来越多的关注。在注重可解释性的医疗问题中，这些方法和理论的深入研究有利于提高诊疗知识的使用效果，完善目前的智能诊疗体系，以更低的卫生服务成本更高的诊断质量辅助医生服务大众患者。

10.3.4. 可解释临床诊疗辅助决策模型

定义：在机器学习技术推动下，人工智能在医学多个领域取得了进展，尤其是在临床诊疗辅助决策方向。但由于机器模型的黑箱性质，决策支持的可解释被越来越多的临床医生和研究人员所呼吁。因此可解释技术作为为决策过程提供透明度，并尽可能的减轻模型与数据带来的偏见的手段越来越受到重视。

目标：人工智能应用于医学会带来潜在的影响，我们应该确保其在临床实践中鲁棒和可信，以能协助医生最大可能地受益于患者。可解释技术作为医疗人工智能评估的一方面，应该作为一个系统全面严格评估的工具，成为聚合分析工具和安全审查的辅助手段，而不是仅针对于单一样本决策的行为描述。例如，一个临床诊疗模型在静态固定的测试集表现不错，但可解释方法显示此模型并没有关注合理特征，则说明测试集存在信息泄露或是模型利用了虚假关联。另外，可解释技术潜在目标也包括帮助医生发现新的临床诊疗知识，提高模型性能以及识别偏差与错误。

进展：人们一直致力于通过解释机器学习模型的预测来打开模型的“黑匣子”。

一种方法是研究模型的权重，从输入中提取重要的输入和特征。例如注意力机制，其通过注意力权重可视化促进了模型可解释的研究。然而，注意力权重是否能够为模型的决策过程提供可靠的解释仍有争议。另一种方法是为模型的决策生成自然语言解释。这通常是通过在人类解释上训练模型来完成的。其他可解释方法包括事后解释技术，例如本地可解释方法 LIME 和沙普利值 SHAP; 基于概念和原型的解释，以及基于样例重要度的解释等等。尽管如此，可解释技术应用于医疗领域仍面临着非常挑战，比如认知偏差(Belief bias)。预训练模型 SciBERT 在 PubMed 等带有偏见的医疗资源上进行训练用于疾病分类等任务；尽管解释技术产生的解释往往围绕突出显示文本中有助于决策的单词，但这并没有揭示模型为这些单词学到的联想意义。模型依赖于决策捷径，例如将“医生”与“男性”联系起来，因此使用这种归因解释来为决策提供信息是有风险的。

影响和发展：目前我国基于文本信息（如电子病历）的机器学习模型可用于患者早期辅助诊断、病历质控及证据溯源的研究仍处于起步阶段，尤其是将可解释性技术应用于中文临床医疗辅助诊断方向。既往临床工作中产生的大量非结构化的自由文本（即电子病历）往往在科学研究中被忽略，丢失了可能隐藏于其中的重要信息。例如垂体腺瘤患者兼有神经系统和内分泌系统疾病的症状和体征，出现全身多种并发症，如库欣病的高血压、糖尿病、骨质疏松、心脑血管疾病、严重感染以及精神障碍等。这些并发症及其它临床特点和诊疗过程等信息都被记录在电子病历中， 隐含了关于患者及疾病的大量信息。而既往对疾病的诊断及辅助决系统策往往使用的是结构化数据及影像数据，如患者的各项生化指标，年龄，病程，性别，磁共振数据等信息，较少使用电子病历。因此，可解释临床诊疗辅助决策模型的研究可以用于两个方面的研究，一是辅助专病及其并发症进行全面且准确的诊断，二是在病历文本中对诊断进行决策证据溯源，即诊断的可解释性分析。可解释技术的应用将增加模型的可信度，为人工智能决策过程提供透明度，并可能减轻各种偏见。但由于目前可解释技术不完善和不成熟的局限性，我们提倡对模型进行多轮全面严格内部和外部的验证，将可解释性技术作为其中一种有价值的分析工具和算法审计的辅助手段，而不是全部。

10.4. 领域产业发展现状及趋势

近年来，来得益于 AI 相关技术的长足发展，医学信息处理技术也取得了很多突破，开始应用于很多业务场景，既包括单一业务场景（如医学影像自动分析和解读），也包括全场景的业务场景（如疾病诊疗、预测和管理等）。下面将从产业发展现状和发展趋势两个方面做简单介绍。

10.4.1. 领域产业发展现状

（1） 智慧医院

互联网医疗 近年来，政府开始鼓励和支持互联网+医疗发展，从谨慎探索到明确指导和实施方案，通过政策支持数字化医疗服务发展。在新冠肺炎疫情的催生下，数字化医疗获得了全面发展，也出现了一些新的业态。涉及到在线预约、问诊、药房、随访、远程医疗、医疗信息管理等多个方面。医疗及医药行业数字化创新应用已成为业界新趋势。

医疗大数据平台及其应用 医疗大数据与人工智能技术在智慧医院方面得到了广泛应用，包括临床辅助决策、医院精细化管理、病历质控、智慧科研等方面，取得了初步成果。医疗大数据以数据治理为基础，将多模态的数据整合、治理后形成高质量可计算的数据库，结合自然语言处理、医学知识推理、机器学习、大数据统计分析等人工智能技术，助力医院临床、管理、科研及随访等业务。

国内开展大数据平台及人工智能应用等业务的企业包括主要包括医渡云、森亿智能、惠每医疗、生命奇点、大数医达等医疗大数据人工智能的新兴企业，其中医渡云于 2021 年 1 月港股上市，意味着医疗大数据行业进入爆发期。除这些新兴企业外，一些互联网企业如腾讯、阿里、百度、讯飞等、一些老牌的医疗行业企业平安、杭创、东软、北大医信等和信息服务提供商华为、中国电信等也在布局医疗大数据平台及人工智能相关业务。

（2） 肿瘤早筛早诊

早筛、早诊、早治一直都是全世界普遍达成共识的降低癌症发病率、死亡率的有效手段。在“健康中国”时代背景下，国家不断宣传肿瘤早发现、早诊断、早治疗的防治理念，肿瘤早筛越来越受到国人重视。2018 年，国家卫健委提出了针对原发性肺癌等 18 个肿瘤的高危人群进行早筛、早诊、早治的要求。国内从事肿瘤早筛的公司包括华大基因、泛生子、诺辉健康、和瑞基因、吉因加等。

（3） 药物研发

真实世界研究 真实世界研究针对临床研究问题，在真实世界环境下收集与研究对象健康有关的数据（真实世界数据，Real World Data， RWD）或基于这些数据衍生的汇总数据，通过分析，获得药物的使用价值及潜在获益—风险的临床证据（真实世界证据，Real World Evidence， RWE）的研究过程。2020 年之后，国家药品监督管理局发布了真实世界证据相关的一系列文件，指导真实世界研究应用于药物研发。真实世界证据应用于支持药物监管决策，涵盖上市前临床研发以及上市后再评价等多个环节。

AI 制药 近年来，人工智能技术在海量数据中筛选新的治疗靶点和新药分子结构中获得了成功应用，有望大大缩短药物发现所需的时间和降低药物研发成本。

国内外一些制药公司陆续启动 AI 辅助制药流程，并取得了一些喜人成果。如总部位于中国香港的国际知名 AI 制药公司 Insilico Medicine （英矽智能）通过人工智能发现了治疗肺纤维化的新靶点，并设计了一个新的药物分子来靶向这个靶点。整个研发过程仅在不到 18 个月内完成，花费仅为大约 200 万美元。

（4） 数字疗法

数字疗法（Digital Therapeutics，DTx）是以循证医学为基础，由软件程序驱动的干预方案，用以治疗、管理或预防疾病。目前国内外的数字疗法产品可以按照适应症分为呼吸系统类、精神类、内分泌类等十二大类。国内外数字疗法已经在糖尿病、代谢综合症、轻度认知障碍等疾病的管理和治疗中得到应用。

（5） 疫情防控

大数据、人工智能等技术在疫情等防控中起到了重要的作用。包括在疫情期间的自我筛查工具、新冠疫情趋势预测及政策仿真模型、智慧流调和智慧溯源工具、风险评估工具；非疫情期间，基于多维度数据的监测进行传染病与症候群的智能监测与预警。

（6） 创新保险

大力发展人工智能技术支持的创新型健康保险业务，为国民提供更加优质的保障。在丰富商业健康保险产品的同时，开展多样化健康保险服务。在完善基本医疗保障制度、稳步提高基本医疗保障水平的基础上，保险公司要提供多样化、多层次、规范化的产品和服务。作为一种普惠型补充医疗保险，它既可以可以对传统的社会医疗保险与商业医疗保险进行有效补充，也可以在发展多样化健康保险服务方面，建立保险公司与医疗、体检、护理等机构合作的机制，加强对医疗行为的监督和对医疗费用的控制，促进医疗服务行为规范化，为参保人提供健康风险评估、健康风险干预等服务，以此为基础探索健康管理组织等新型组织形式。

（7） 智慧中医

中医药在“治未病”、养老、康复等方面作用突出，受到了广泛重视。提高中医药服务能力，完善体系，对切实发挥中医药的优势与特色，充分发挥中医药在医疗保健服务中的重要至关重要。在新的形式下，构建现代化中医院是目前发展的趋势。

10.4.2. 领域的产业发展趋势

医疗大数据的趋势是构建产业生态

医疗大数据的趋势是构建新一代医疗数据开放中心，一方面，对医院的多模态数据进行接入、治理和管理；另一方面，打破院内院外、行业内行业外的数据“交换和共享”壁垒，最大限度地发挥医疗大数据的价值。例如行业龙头公司医

渡云新研发的 EYWA 数据平台，该平台可以通过数据治理帮助医院将数据形成高质量的、可计算的数据，对任意第三方提供基于私有云的数据和计算资源的开放，供其进行数据分析和挖掘，并形成统一的应用开放平台，推动医院数据应用生态的发展。

联邦学习与多方安全计算有望解决数据共享和安全问题

鉴于多中心研究时的数据共享难、数据安全难以保证等诸多问题，基于联邦学习与多方安全计算技术的医疗数据联合应用的技术将会成为医疗数据应用的新范式。

数字疗法会成为新的市场热点

计算成本和传感器成本的大幅下降，以及大数据和人工智能技术的快速迭代推动了数字疗法的高速发展。软件程序能够大规模提供高质量、经临床验证的治疗干预措施。新技术和政策引导可能会使得针对更广泛健康状况的数字健康干预措施激增。

AI 在药物研发中发挥重大作用

新药研发，需要经过药物发现、临床前研究、临床研究和审批上市等多阶段，往往需要耗费十几年乃至数十年的时间，以及数十亿美元的成本，并伴随着高达90%以上的失败率。AI 制药必将是一个具有广阔应用前景的新兴研究领域。

10.5. 总结及展望

在十三五期间，人工智能技术获得了飞速发展，并迅速应用到各个领域。医疗健康领域事关人民生命健康，对基于人工智能技术的创新性疾病诊断、治疗、药物研发等应用既有强烈的需求，又保持高度的审慎。这种审慎既体现在诊疗数据开放性上，也体现在诊断和治疗技术的实际临床应用上。数据的开放性首先亟需国家相关数据隐私保护与应用的明确政策支持，另一方面，隐私计算、联邦计算、区块链等数据隐私保护计算技术的大力推广应用，也能对临床数据在医疗健康的人工智能技术发展上获得充分利用提供支持与保障。对于人工智能技术在临床诊疗上的应用，审慎的主要原因还是在于现有技术的解释性与鲁棒性不足。另一方面，也和技术成熟度不够，应用范围较窄，且成本高昂有关，IBM 沃森健康在肿瘤疾病诊疗中与 MD Anderson 不算太成功的应用合作历程是一个值得认真总结的典型例子。2020 年，作为“十三五”的收官之年，席卷全球的新冠疫情，给医疗健康领域带来巨大的挑战，同时，也让人工智能技术在医疗健康领域的应用快速发展，从智能流调、流行趋势预测，到非接触式互联网医院的逐步改革推进，都释放出了大量的技术需求。

“十四五”期间将是智能信息处理技术在医疗领域扎根落地的关键时期，国家智慧医疗的支持力度进一步加大，从智能医生助手，到疑难疾病会诊等课题的推进，无疑都在推动智能信息处理技术向临床应用靠近。为了切实推进技术的临床应用，我们也必须持续研究并解决好模型与方法的鲁棒性、可解释性，数据的隐私性与共享的矛盾等问题。为此，在“十四五”期间，智能信息处理技术在医疗领域应用的工作重点将包括：1）临床多模态数据的标准化、去隐私化化政策的完善与技术的发展；2）中文临床医疗知识知识的规范化标准的制定、大规模中文临床知识的建设与开放共享；3）构建鲁棒的、可解释的临床诊疗决策模型与方法；4）促进智能信息处理技术在医学研究、药物研发等产业领域的深入应用；5）进一步促进隐私计算、联邦计算、区块链等技术与医疗领域的融合应用。

接下来的五年，将成为智能信息技术真正走向临床应用的五年，我们也期待智慧化的诊疗系统经过这五年的发展能够真正走到医生和患者身边，切实以技术来提升我国的医疗服务水平，为解决优质医疗资源的紧缺和分布的不均衡做出重要贡献。

第十一章　网络空间大搜索技术研究进展、现状及趋势

11.1. 研究背景与意义

IT 技术是推动各行业发展的催化剂，而搜索技术被誉为**IT 技术皇冠上的明珠**。网络空间由互联网空间，扩展到了信息、物体和人关联的泛在网络空间，伴随着网络空间的发展，大搜索需求也随之产生，搜索范围从传统的互联网信息搜索，到泛在网络空间的人、物以及知识等的泛在搜索，数据范围从基于特殊通道的特殊数据处理，到结构化数据分析挖掘，正朝着面向大数据的智能搜索方向发展。传统搜索引擎基于关键字匹配和排序、关联信息推荐、服务搜索等，是一种存在性搜索。随着搜索空间从面向信息的互联网扩展到了人、机、物互联以及服务、应用的泛在网络空间，网络应用模式从 Web1.0 发展到了 Web3.0，以及大数据技术的广泛应用，传统的搜索引擎系统已经不能满足时代的需求。网络空间大搜索技术是综合利用大数据分析、自然语言处理和人工智能等技术，对网络空间中的大数据进行获取与分析，获取其中蕴含的有价值的信息和知识，针对用户的搜索需求，将全面准确的知识解答尽快返回给用户。**网络空间大搜索支持泛在网络空间对人、物、知识的搜索，是新一代搜索引擎**。

网络空间大搜索是指面向泛在网络空间中的人、物体和信息，在正确理解用户意图的基础上，基于网络空间大数据知识获取，给出满足用户需求的智慧解答。网络空间大搜索具有**“5S”特点**，即：**泛网获取（Sourcing）**，从多通道获取信息；**用户感知（Sensing）**，理解用户的搜索需求;**多源综合（Synthesizing）**，构建由巨规模实体及关系的知识图谱，并进行综合；**智慧解答（Solution）**，产生智慧解决方案；**安全可信（Secure）**，能够保护用户隐私。与之相对应，网络空间大搜索主要研究内容包括：（1）泛在网空间中知识获取与验证方法；（2）搜索真实意图的理解与表示方法；（3）泛在网络空间中知识表示、演化与管理方法；（4）智慧解答的快速匹配求解方法；（5）搜索结果可信与隐私保护机理。在此基础上，搭建面向网络空间的通用大搜索引擎支撑平台及环境，在公开信息价值搜索和智能政务大搜索等方面开发一组关键示范应用，并取得显著的社会和经济效益。

网络空间大搜索技术是一个有着自身特点的学科领域，其发展需求同国家需求与社会发展需求高度吻合。发展网络空间大搜索技术，对于我国信息技术的发展具有重要意义，包括：1）**抢占 IT 技术高地**：如果说大搜索是过去二十年互联网时代皇冠上的明珠，那么网络空间大搜索是将来二十年物联网和人工智能时代皇冠上的明珠；2）**保障国家战略需求**：自主的搜索引擎是国家安全的基础。俄

国就提出建设全球首个国家搜索引擎，加入安全接入、过滤内容等；3）**促进社会经济发展：网络空间**大搜索与各种现实生活具有深度关联，将在环保、医疗、教育、交通等各种领域有深入应用，在可以预测的将来，网络空间大搜索将重新定义我们的生活，服务于更多的大众，全方位提高人民的生活质量，并推动 IT 技术的发展。

11.2. 领域发展现状与关键科学问题

11.2.1. 领域发展现状

随着泛在网的普及和发展，以及 Web2.0 鼎盛期和 Web3.0 萌芽期到来，网络搜索技术也正面临着巨大的机遇和挑战。在 Web1.0 时代，以 Google 和百度为代表的搜索引擎对于互联网的发展起到了巨大的推动作用，时至今日依旧发挥着重要的作用。但在 Web2.0、Web3.0 时代，泛在网中包含的具体内容是人物、信息和物体，被搜对象具有动态演绎特性，传统的搜索技术已不能满足泛在网环境下用户智慧搜索的实际需求，单纯在技术上进行微创新和改进并不奏效，具体原因包括如下方面：

（1）用户搜索意图的正确理解方面。目前的搜索引擎主要采用关键字匹配技术，如用户在搜索“人大”时，不能确定用户是具体想搜索“中国人民大学”还是“中国人民代表大会”；在搜索“天涯海角”时，不能确定用户是想搜索成语、景点还是车站。泛在网智慧搜索需要对用户真实搜索意图进行分析，结合社交网络与上下文场景等语义信息，实现智能化的个性搜索。如何准确理解用户的真实意图，目前的搜索引擎未能有效解决。

（2）内容和时间相关的查询方面。传统的搜索引擎在时间相关性方面支持不够，例如，用户在查询“今天去哪里看红叶”或者“现在去哪里急诊最快”时，现在的搜索引擎只能给出有关红叶的网页介绍，和有关急诊的一些历史信息，而不能给出实时的正确答案。下一代泛在网搜索引擎需要支持查询结果与时间敏感的查询。

（3）内容和空间相关的查询方面。目前的搜索引擎与用户搜索时所处的位置关系不大，而泛在网络智慧搜索引擎需要将内容和空间需求在终端设置，例如，用户搜索“我的位置半径 500 米以内，哪里可以买到 NIKE 的 T 恤衫”时，现在的传统搜素引擎只能给出 Nike T 恤衫的网店、以及 Nike 官方的关于 T 恤衫的介绍的网页，并不能返回用户真实需要的信息。

（4）关系与关系演化相关的查询方面。在 Web2.0 应用中，需要在社交网络

中发现符合用户输入的人群、关系、变化过程等，譬如查询符合某些特性的社区查询，返回人与人之间的关联关系的关系查询，分析特定社区演化规律的社区分析，及人与人之间关系发展变化规律的查询等，现在的 Web1.0 搜索引擎显然不满足上述需求。

（5）物体/设备相关的查询方面。尽管在针对物联网设备相关的查询方面出现了一些类似 SHODAN 之类的搜索引擎，但是在发现物体/设备的位置、轨迹、状态等信息方面，尤其是针对浩如烟海的终端设备进行实时地位和轨迹回放查找方面，已有的传统搜索引擎不能满足对物体/设备相关的实时查询。

（6）多精度的可控查询方面。随着泛在网络中被查询对象类别和数量的急剧增加，从安全和隐私保护角度考虑，内容所有者也应根据查询者不同的身份或权限，返回不同精度的信息。例如我现在的具体位置信息，针对陌生人、同事、好友、家人，在不同的时间段，公布的精度范围需要从不公开、2K 米、500 米以及 10 米不等的信息进行精确可控，目前的搜索引擎显然不支持该需求。

综上所述，为了满足新型网络和应用环境下用户搜索的新需求，网络空间大搜索必须具有“智慧”的特征。泛在网智慧搜索是能够洞察理解用户的搜索意图，在海量、多源、异构、多态、不确定的数据中，实现对与人物、物体和内容等相关的信息的对象级搜索，提供最贴合用户需求的搜索结果。在泛在搜索中，搜索的对象不仅包括传统搜索中的信息，而且涵盖了更为广泛的实时变化的人、物体、事件等信息。在泛在搜索中，“搜索”需要准确了解用户的需求，在可选择的范围内，帮助用户预先判断和选择，用户拥有最终的选择权。“智慧”是基于海量数据的融合与交互，帮助用户做选择。从不同维度，提炼用户行为组成用户数据，通过用户数据分析用户的行为特征，洞察用户的真正需求，支持对对象实时状态搜索和演化趋势的关联分析，使搜索内容从虚拟世界走向现实世界。

11.2.2. 关键科学问题

网络空间大搜索面临的科学问题主要包括如下四个方面：

科学问题 1、如何在大数据的海洋中，对多模态、多层次、多粒度的知识进行提取？

面对泛在网络空间大数据中数以万亿计的对象及关系，其种类繁多，包括音视频、图片、文本等结构化、非结构化的数据，且其属性在不断演化，如何对多模态的数据进行多层次、多粒度的准确知识抽取？

科学问题 2、如何实现对海量用户的搜索意图准确理解与表示?

如何对用户用自然语言、语音、手势、历史偏好和个人情感信息等表达的搜

索意图，以及用户所处场景及上下文，进行准确的理解与表示？如何在大数据环境下进行更加准确的语义级意图理解？

科学问题 3、如何对海量、分布、异构、演化的知识进行管理，及面向搜索任务进行融合推理？

如何对海量、分布且不断演化的知识库进行有效存储管理，并对泛在网络空间上的动态异构对象间建立相关联的高效索引知识库？如何将用户搜索意图，在海量分布的知识库上进行快速的匹配和融合推理？

科学问题 4、如何对搜索结果的可信度进行准确评价？如何保护用户的隐私不被泄露？

如何保证用户搜索过程中采集的数据是可信的，而且经过知识推理的结果是可信赖的？如何保证合适的搜索结果只返回给合适的用户而不被滥用？如何保证用户的个人隐私不被泄露？

11.3. 领域关键技术进展及趋势

11.3.1. 泛在网络空间信息获取与发掘

泛在网络空间信息获取与发掘的是基于泛在网络获取的大数据，面向泛在网络空间的海量实体及关系进行知识挖掘，通过关联、融合、统计、推理、众包等知识获取和推演技术，发现和获取数据中蕴含的各类知识和智慧的过程。

泛在网络空间信息获取与发掘目标是以一定的策略和方法、面向给定任务目标在网络空间中采集、获取和挖掘相关数据和信息。主要技术包括：

（a）面向目标任务的多来源、多模态数据获取方法：面向各种应用领域和不同数据模态的目标任务表示、匹配及获取技术；面向实时数据流的目标信息采集技术；目标驱动的异构、异质数据的协同采集技术；巨规模采集任务并行计算和管理平台技术；目标采集数据的完整性和精确性评估模型等。

（b）面向目标任务的关联数据发掘方法：数据关联推演知识的表示、管理，及基于推演的间接数据获取方法；基于上下文的多模态数据关联挖掘方法；场景、时空感知的关联数据挖掘分析方法；基于众包、标注等方法的关联数据挖掘方法等。

（c）巨规模、多模态实时数据流的清洗方法：基于滑动窗口数据摘要及优先队列的重复数据删除技术；基于编辑距离算法的异构相似数据匹配技术；基于情景语义描述模型的噪音数据清洗技术等；

（d）泛在网络空间数据融合与冲突消解方法：基于数据依赖关系图的多模

态异构数据的融合计算模型；情境驱动的多层次融合和情景语义描述模型；基于本体论的多层次（数据级、特征级、决策级）数据冲突消解方法等。

下面从文本知识获取、图片知识获取和视音频知识获取三个方面对泛在网络空间信息获取与发掘的进展影响和发展进行展开。

11.3.1.1.文本知识获取

文本的知识获取从所抽取的内容上看主要包括实体知识抽取、事件抽取、属性抽取，下面分别从这三个方面进行介绍。

在实体知识抽取方面。实体知识抽取是面向信息提取、问答系统、句法分析、机器翻译、语义网络（Semantic Web）元数据标注等应用领域的重要基础研究。通常而言，早期实体的任务旨在识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）实体。在实体知识抽取研究发展初期，针对西方语言的实体知识抽取一般都是基于手工编制规则而构建规则系统的人工方法。其中具有代表性的工作包括 20 世纪 90 年代纽约大学的 Grishman 等人开发的参与 MUC-6 评测的 Proteus 系统和 IsoQuest 公司的 Krupka 等开发的参与 MUC-7 评测的 NetOwl 系统是最早针对西方语言的实体知识抽取模型，相关理论研究及应用系统大多通过分析实体的词型特点、上下文信息、语法成分特征以及用词规律等特点来构建相关规则。随着近几年深度神经网络在人工智能相关领域应用的不断深入，自然语言处理中的很多任务利用深度学习都得到了不错的结果。2020 年卡耐基梅隆大学的 Lample 等人在此基础上提出了语言无关的实体抽取模型，在英语、德语、荷兰语和西班牙语这四种语言上都取得了不错的成绩。

目前发展趋势是采用深度学习模型的命名实体识别技术进行命名实体识别。主要可以分为两大类：一类仍采用序列标注的方法，结合 CRF 模型与深度神经网络，如 RNN（Recurrent Neural Network，循环神经网络）和 CNN（Convolutional Neural Network，卷积神经网络）等，通过标签分类方法进行实体识别；另一类利用 Span（跨度）的方法，枚举句子中所有可能为实体的词组（N-Gram），通过语言模型或者其他神经网络的方法计算每个词组的表示，利用分类的方法进行实体识别。当前命名实体识别在公开数据集 ACE2003 上的效果已经能达到 94.3%。

在事件抽取方面。事件抽取技术是从非结构化的信息中抽取出用户感兴趣的事件，并以结构化的形式呈现给用户。根据事件的相关定义，事件抽取任务可分为元事件抽取及主题事件抽取。当前的事件抽取研究主要面向元事件，而主题事件抽取的研究成果较少。元事件表示一个动作的发生或状态的变化。针对事件抽

取任务，主要包括事件类别的识别与分类以及事件元素识别两大核心任务，以往研究工作的重点在于预先规定好事件类型并定义事件模板。2019 年，华盛顿大学的 Ritter 等人将机器学习与模型匹配相结合，克服了单一模型的局限性。基于事件框架的主题事件抽取方法通过定义结构化、层次化的事件框架来指导主题事件的抽取，利用框架来概括事件信息，表达主题事件的不同侧面,目前相关研究比较少。典型工作主要应用于主题新闻事件文档检索与分析。

在属性抽取方面。属性抽取是指抽取事物本身所固有的性质和事物的一些基本特性。事物的属性通常是从多个方面和多个层次来表现的，因此事物的属性是多样性的。研究事物需要识别出这些事物的属性特征，可以深入了解这些事物的特征和内涵。属性抽取与应用的联系十分紧密，目前的研究热点在人物属性抽取、企业属性抽取和概念属性抽取上。2002 年，英国南安普敦大学的 Alani 等人把属性抽取和本体结合起来，开展了 ArtEquAKT 项目实现自动从网页中抽取艺术家的信息，并生成人物传记。2019 年，上海海事大学的 Zhong 等人将神经网络模型应用于属性抽，克服了传统方法对文本表示能力较差的问题。2021 年，中科院软件所的 Zhang 等人将属性抽取与实体抽取进行结合，使这两种抽取相互促进，同时提升了实体抽取与属性抽取的效果。

在关系抽取方面，监督方法一般通过对无标注语料中实体关系特征的学习进行聚类，然后依据聚类的结果给定关系。但是受限于聚类结果本身难以规则化和低频率实例召回率低等问题，抽取效果一般较差，且难以直接由于构建知识图谱。半监督方法处理只有少量标注的情况，利用 Bootstraping 以及远程监督学习的方法，依据已有标注信息抽取新实例来丰富训练数据。该方法不需要人工标注，但是需要依赖已有知识图谱，并且语料汇总噪音较多。有监督方法依赖于高质量的标注好的数据语料，采用分类方法解决。目前针对有监督的关系抽取方法是研究最为广泛充分的，该类方法对于高质量数据的性能（F1 值）能达到 90% 以上，但该类方法只能抽取固定的关系类别，模型迁移性较差。

11.3.1.2.图片知识获取

当前图片知识获取主要包括基于人工设计的特征表达方法和基于深度学习的特征提取方法和基于图片的视觉关系检测。

在基于人工设计的特征表达方法方面。最初面向图片的知识获取主要从图像分类、物体检测等技术中获取图片中的概念信息，主要是基于人工设计的特征，主要包括基于局部 SIFT 方法、基于直方图 HOG 的方法和基于全局 GIST 的方法。例如，研究人员采用梯度方向直方图 HOG 特征预测图片中存在的概念，研

究人员提出了一种尺度不变描述 SIFT 特征作为图片特征用来预测图片中的概念。

在基于深度学习的特征提取方法方面。自从 Alex Krizhevsk 等学者在 2012 年提出一个 8 层的深度模型并在 Image-Net 竞赛上取得非常好的效果后，卷积神经网络(Convolutional Neural Networks，CNN)在图像分类与识别领域受到了广泛关注，并取得了巨大成功。将卷积神经网络用于图像识别与分类，可以归纳为三种途径：一是直接在待分类的数据集上训练一个深层的网络。随着 CNN 深度和宽度的增加，CNN 的分类性能有着明显的提升。二是在训练好的网络上直接提取特征。训练好的 CNN 模型可以直接用来当特征提取器，提取的特征可以用做其它的后续操作。三是在目标数据集上对现有深度模型进行“精细化”（Fine-tune）改进。在特定数据集上训练好的模型有很强的泛化性能，但是 Fine-tuning 能够进一步提升分类性能。Fine-tuning 是在目标数据集上重新调整网络参数，从而使深度模型能够捕获针对目标任务更具有区分性的特征。

在基于图片的视觉关系检测方面。更全面的图片知识获取不仅需要能辨别视觉信息中包含的物体和场景等概念，还需要考虑物体与物体之间的关系，进行视觉关系检测乃至对图片的语言描述。学习关系是通用智能行为的一个重要组成部分。在图像研究中，学习图像中物体之间的关系也是深层次理解图像的重要表现方式。在图像研究与发展的过程中，一些基本的关系被用来辅助其他任务。如空间关系用于图像分类和图像分割，物体的共生关系用于场景分类。目前研究通用视觉关系检测的方法主要分两步：1）物体区域的检测，2）两区域之间关系的预测。物体检测采用现有的物体检测技术（如：Faster-RCNN）。在两区域之间关系的预测中，先分别得到物体和谓词的类别，最终预测的关系由两物体和谓词共同决定。由于关系中物体与谓词高度依赖，利用这种依赖可以提升关系的检测。图像内容描述输入为一幅图像，输出为描述该图像的句子。近年来，随着深度学习的盛行，Karpathy 等提出了基于 CNN-RNN 的方法来产生语言描述。其中，CNN 用来提取图像的特征，RNN 将该特征解码为语言描述。此外，一些工作集成了从自然图片中产生场景图。

11.3.1.3.视音频知识获取

面向视音频的知识获取涉及到视音频的表示、视音频与语言的关联两个方面的内容。得到视音频的语言描述后，我们可以进一步基于文本的信息进行结构化的抽取。

在视音频的表示研究方面。数据表示是视音频分析、识别、理解与搜索等任务的基础性核心问题，长期以来受到广泛的关注和重视。相关工作主要从两个方

面开展。传统的方法依然是基于人工设计的特征表示,包括主要基于局部 SIFT 的、基于直方图 HOG 的和基于全局 GIST 的方法等。从另一个方面来讲，视音频的表示具有复杂的语义属性，包括物体，场景和事件等。近年来，在基于深度学习的自动表示方面取得了较多成果。2015 年《自然》、《科学》相继出版了“深度学习”相关专辑，探讨机器智能的动态与未来。近年来深度学习也引领了视音频的特征表示与概念识别研究方向，得到了研究者们的广泛关注，包括面向 CNN 和 LSTM 的方法。

在视音频与语言的关联研究方面。在视音频有效表示的基础上，接下来通过视音频和语言的关联获取视音频的知识。涉及到基于单个句子的视音频描述和基于多句子的视音频描述。传统的方法对于基于单句的视音频描述，主要采用基于神经网络的编解码框架进行实现。近来的一些方法侧重挖掘视频的注意力机制，进一步考虑网络视频的不同主题，产生主题引导的视频描述。视频中蕴含的知识丰富，一句话很难全面的给予描述。为此，研究学者提出面向多个句子的视频描述方法。该方法的思路对视频进行分割，对每个视频片段都给予相应的描述。

11.3.2. 超大规模知识图谱的构建与管理

超大规模知识图谱的构建与管理围绕基于异构信息网络的统一对象实体关联表示这一问题展开研究，针对现实世界体现出不确定性、模糊性、多层次性、多维度性，研究基于异构信息网络知识的统一对象实体关联表示模型，解决传统的同质化建模难以准确描述关联数据的异构化形态等问题；针对传统的以关键字为目标的搜索引擎不能满足泛在网络空间中物体、信息和人物的搜索需求的情况，研究实现对包括人、物、信息等在内的泛在网空间数据与知识的获取、表示与组织。

超大规模知识图谱的构建与管理的目标是面对泛在网络空间中数以万亿计的对象及关系，种类繁多，且属性在不断演化，对其进行准确建模，将泛在网络上的动态异构对象间建立相关联的高效索引知识图谱。主要技术包括：

（a）巨规模实体关系的表示模型和方法：基于超图的统一实体关系表示模型；实体间巨复杂关联关系及其演化的表示方法；实体多维属性的及其时空变化的表示方法；基于实体关系表示模型的实体查找、关联、推演等演算方法。

（b）基于实体关系模型的知识图谱组织和管理：面向概念、事件、人物等目标的巨规模知识组织管理方法；多维度、多尺度的知识高效匹配和查询技术；高可扩展、可演化的知识图谱体系架构；知识图谱的支撑计算平台技术等。

（c）知识图谱的实时演化和更新：基于概率统计的巨规模关联知识推演方

法；基于大数据关联分析的知识挖掘方法；面向知识图谱的规则推演的知识发现方法；基于众包的知识冲突消解方法；知识图谱质量的评价方法等。

下面从知识表示管理、知识融合协同推理和基于 MDATA 多维关联模型的知识表示三个方面综述当前国内外发展现状。

11.3.2.1.知识表示管理

知识表示和管理的研究已经有很长的历史，迄今为止，知识表示和管理的模型可以分为基于逻辑的表示模型、基于框架的表示模型、基于语义网的表示模型、基于本体的表示模型和面向机器学习的表示模型。

在基于逻辑的表示模型方面。专家系统已经有很多年的历史，逻辑规则和模糊规则是专家系统中常见的知识表示模型。美国加利福尼亚人工智能研究中心的 Moore 在 1982 年的 AAAI 上首先提出了基于逻辑的知识表示。早期的逻辑规则依靠专家构建，后来研究者们一直探索自动构建逻辑规则的方法。在基于模糊规则的系统中，知识是通过模糊集来表示的。和逻辑规则相比，模糊规则能够更好地表示不确定性和连续变量。早期的模糊规则构建是通过专家的方式。后来人们不断提出新的方法来从数据中提取模糊规则，如波兰哥白尼大学的 Duch 等在 2017 年提出的特征空间映射和 C4.5 分类树方法、加州大学伯克利的 Nurnberger 等在 2021 年提出的基于遗传算法的方法。基于逻辑的知识表示方法的特点是善于表达因果关系，具备很好的推理能力，但在知识表示的灵活性方面有很大不足。

在基于框架的知识表示方面。麻省理工的明斯基于 1975 年首先提出了基于框架的知识表示方法。框架模型能够把知识的内部结构关系以及知识之间的特殊关系表示出来，并把与某个实体或实体集的相关特性都集中在一起。当前在框架知识表示方面仍然有很多研究工作，如加拿大蒙特利尔大学的 Azoulay 在 2017 年 AAAI 上发表的面向大规模专门语料库的基于框架的知识表示方法。基于框架的知识表示最突出的特点是善于表示结构性知识，框架系统的数据结构和问题求解过程也与人类的思维和问题求解过程相似。但框架知识表示缺乏形式理论，没有明确的推理机制保证问题求解的可行性和推理过程的严密性。同时由于许多实际情况与原型存在较大的差异，因此适应能力不强。

在基于语义网的知识表示方面。剑桥大学语言研究中心的 Richens 在 1956 年首先提出了语义网的概念。语义网利用节点和带标记的边结构的有向图描述事件、概念、状况、动作及客体之间的关系。带标记的有向图能十分自然的描述客体之间的关系。加利福尼亚 SDC 公司的 Simmons 等 1960 年在 SYNTHEX 项目中单独进行了语义网方面的开发。WordNet 由普林斯顿大学从 1985 年开始开发，

当前的最新版本是 3.1。FrameNet 由加州大学伯克利的 Fillmore 等人在 1997 开发，是多层次的框架构成的网络。语义网具有广泛的表示范围和强大的表示能力，用其他形式的表示方法能表达的知识几乎都可以用语义网络来表示。但基于语义网的知识表示也有一些缺陷，其推理规则不十分明了，不能充分保证网络操作所得推论的严格性和有效性。一旦节点个数太多，网络结构复杂，推理就比较困难。同时语义网页不便于表达判断性知识与深层知识。

在基于本体的知识表示方面。本体概念的产生可以追溯到古希腊时代。本体模型把知识表示为一个概念的分类系统，其中概念包含属性、值和关系。本体知识表示模型的主要目标是提供一个知识共享与重用的平台。一个本体至少包括三个部分：类（领域概念）、关系及实例。本体的常用描述工具是 Web 本体语言（OWL）和资源定义框架（RDF）。通用本体的构建很难实现完全自动化，一般都是半自动化的方法。通用本体当前面临的挑战是构建困难和难以验证效果。表示本体不面向特定领域，这种本体里面的实体并没有确切地说明应该表示什么，主要应用于语义网。韩国庆熙大学的 Khanet 等 2015 年提出 MBO（Mediation Bridge Ontology）表示本体来支持语义网上的交互。表示本体的构建大都采用人工的方法。基于本体的知识表示方法有很强的表达能力，但本体的构建代价比较大，并且本体表示的复杂性导致基于本体的知识表示在产业界应用的比较少。

在基于机器学习的知识表示方面。和前面的知识表示不同，面向机器学习的知识表示侧重于如何通过机器学习的方法从数据中自动获取知识。在前面提到的几种知识表示方法中，也都或多或少用到了基于机器学习的知识获取。近年来，随着大数据的发展，知识图谱及相关的基于图的知识表示模型得到广泛应用。知识图谱这一概念 2012 年被 Google 公司提出，而后诞生了很多其它的类似模型。这类基于图的模型可以说是语义网的一种，但更侧重实用、大规模和自动化。卡耐基梅隆大学的 Gardner 等提出了一种结合语义向量和随机游走的关系路径推理方式，首次尝试采用语义向量结合路径特征的方式来建立推理模型，获得了比较好的效果。基于图的知识表示需要基于图数据库的管理。现有系统所定义的算子大部分基于图遍历运算和迭代操作，对于图上结构操作（如子图匹配运算）的管理与分析效率不高。此外，这些常用的图数据管理系统都是主要面向一些静态图上的运算。对于不断更新的动态图，现有系统的查询处理的效率不高，尤其是针对大搜索应用中对于异构数据的处理和基于 agent 的灵活配置服务还需要进一步深入研究。

总的来说，知识表示的研究有很长的历史，发展出了多个分支，每一种知识表示方法都有其优缺点。由于大搜索应用中知识的多层次、多粒度及多样性，单独的任何一种已有知识表示都难以直接适用。

11.3.2.2.知识融合协同推理

虽然当前搜索技术已经相对比较成熟了，但是目前主流的搜索技术都是针对单一数据类型（如文本、图像、音频、地图等）或单一目标（如酒店、机票等）的，很少有系统能实现复杂的搜索任务，如需要多个搜索目标通过业务逻辑、时间顺序、物理位置等的组合，完成一个复杂的任务。通过简单任务的组合实现复杂任务这一思路在相关领域已经有了广泛的研究，在 Agent 和智能 Agent 领域，通过在多个 Agent 之间建立通讯机制，可以实现多个代理之间的协同工作。作为多 Agent 系统的研究热点，多 Agent 协同及任务规划问题已经有了多年的研究。可以把协同模式分为三类：集中式、分布式和分散式。

在集中式规划结构方面。集中式规划结构指多 agents 算法能在单台机器上运行，算法不同模块间的信息交互通过共享的存储器进行。多 agents 任务分配问题有很多经典模型，包括旅行商问题 TSP、车辆路径问题 VRP 等。早期的方法是精确方法，但精确方法只是对目标函数和当前系统模型最优，而这些模型往往是近似的。为解决这些复杂问题，研究者们提出了大量的近似方法。为进一步减少求解此类问题的计算时间，波士顿大学的 Castanon 等采取了一种通过限制持续时间来减少问题规模的方法—滚动时域规划。除混合整数线性规划框架外，求解自主多 agents 编队规划问题应用框架还包括马尔科夫决策过程（MDP）。近期对 MDP 的重大改进是谷歌的 Mnih 等在 2019 年提出深度强化学习，并在视频游戏领域取得突破性进展。在此基础上，最近几年取得很多重要进展，如双重深度 Q 网络和深度注意力 Q 网络。快速的集中式算法经常使用并行计算实现，其充分利用了现代计算机系统多处理器的优点。由于算法中所有模块有权快速获取当前全局共享存储状态，模块之间的通信代价可以忽略不计。在某些环境下，agent 应具备更自主的能力，如在规划参数（agent 状态，任务参数，环境变量）变化很快，需传输大量数据给中心处理单元的情况下，集中式的结构并不理想。

在分布式规划结构方面。分布式规划结构中，分布式算法运行起来像分离的模块，这些分布的算法模块使用其自身的存储分区来保存与规划过程相关的数据，模块间相关的信息通过可靠的通信信道共享。分布式算法信息共享带来了通信代价和延迟，在通信层面引入了比集中式算法复杂的附加层。分布式算法依赖牢固的通信设施，每个分布式节点熟悉能与其通信的所有其他节点，并假设节点之间收、发的消息是可靠、低延迟的。为解决分布式的混合整数规划问题，很多研究者提出了不同的方法。一种常见的方法是卡耐基梅隆大学的 Dias 等 2016 年提出的基于市场机制的算法，通过设计共享 Agent 分配信息而非态势感知去完成任务空间的一致性，有效解决了分布式或分散式环境下的混合整数协同分配问题。马

萨博弈论为解决多 agents 规划问题另辟蹊径，将 agent 之间的交互视为博弈。博弈论的基本思想是，agent 是单独的决策实体，基于对环境和其他 agents 的知识，采取行动以最大化其局部效益。

在分散式规划结构方面。分散式规划结构由通信基础设施不可靠且零星分布的环境中，自主规划的独立 agent 组成。这种环境下，消息延迟、网络连通性、程序执行率或消息到达可靠率方面都没有严格的约束和保证，算法不能依靠模块间恰当的信息共享，从而限制了 agents 间可实现的协作和协同量。当实际的通信条件较好时，相对分布式结构，分散式算法可能是保守的，性能亦差于分布式算法，但它们在通信环境激烈变化的情况下比分布式结构稳定。由于不给 agent 设置严格的交互规则，完全分散式的算法使得稀疏链接的 agent 具有更高的自主程度。在弱通信环境中，分散式算法使得大规模编队能有效地交互，而不会使整个网络陷入限制性的消息传输需求的困境中。在规划协同语言方面，耶鲁大学的 McDermott 等在 1998 年提出了规划领域定义语言 PDDL。目前 MA-PDDL 已经被很多很多人使用，成为多 Agent 协同和规划的一个事实标准。

总的来说，多 agents 协同和任务规划问题经过多年的发展积累了很多成功的技术。但在多 agents 的一致性保证、规划的优化等方面，仍然面临着很大的挑战，特别是当 agent 数量众多，环境又比较复杂的时候。

11.3.2.3.基于 MDATA 多维关联模型的知识表示

在大数据时代，信息过载，人们面临着数据爆炸问题，从原始的数据到有效的信息之间存在着一个鸿沟，知识图谱从一定程度上缓解了这种鸿沟问题。但是，传统知识图谱通常还面临着多模态、知识动态变化、多领域知识统一表示困难、知识融合难度大等问题，迫切需要新的建模方式。大搜索专委会国防科技大学贾焰教授团队提出了支持时空属性和演化的超知识图谱表示 MDATA 模型，在统一的知识本体体系框架的基础上，通过引入标量属性和向量属性，以及时空标签等方式，支持知识的“超语义图”表示方法，突破了巨规模关系结构、复杂语义内容及其时空属性融合表示的世界性难题。

（1）时空特性缺失。现有的知识图谱三元组模型只能表达一些简单的关联事实，但很多领域应用的需求已经远远超出了三元组所能表达的简单关联事实，实际应用日益对于利用更加多元的知识表示丰富和增强知识图谱的语义表达能力提出了需求。这一趋势首先体现在对于时间和空间语义的拓展与表达方面。有很多知识和事实是有时间和空间条件的，比如说“美国总统是特朗普”这个事实的成立是有时间条件的，十年前美国的总统不是特朗普，十年之后应该也不大可

能是特朗普。还有很多事实是有空间条件的，比如“早餐是烧饼与油条”这件事，在中国是这样，但是在西方并非如此，西方的早餐可能是咖啡、面包。从时空维度拓展知识表示对很多特定领域具有较强的现实意义。

（2）无法区分关系与属性。当前的知识图谱用<头实体，关系，尾实体>这种三元组来表示，但是这并不符合实际情况，尤其是不能区分关系和属性。对于属性而言，“尾实体”往往只是一个字符串/数值常量，而不必是一个实体节点。实体就是做精确匹配，可以基于其关联的属性等来做进一步分析；而属性值是无法作为源头关联更多节点和边，但是可以使用数值计算等方法。对于后续的检索和匹配任务而言，不能区分关系与属性会造成较大的性能影响。

（3）不能处理向量化属性。对于一个实体而言，某种属性往往是具有向量化特征的，尤其是与时空相关的属性。例如，一个人的身高会随着时间的变化而变化，一个人的工作地点也往往会随着工作变迁而改变，这些属性都可以看成是向量化属性。但是，当前知识图谱只能表示标量属性，在面临这种情况时把这些属性拆分成多条三元组，这会对具体的存储模型造成很大的负担。而采用向量化的表示方式可以避免这种问题。

11.3.2.3.1. MDATA 模型概述

MDATA 模型包括关联表示，关联构造，关联计算等部分，如图 1 所示。其中，关联表示形成一张超语义图，可以用如下四元组来表示：

$$< Concept, Entity, Relation, Property > \quad (1)$$

其中$Concept$代表概念集合，表示为$\{concept_i | i = 1, \dots n\}$；$Entity$代表概念的实体集合，表示为$\{entity_i | i = 1, \dots m\}$；$Relation$是实体的关系集合，表示为$\{relation_{ij}\,[, TZone] \,|\, relation_{ij} = < entity_i, entity_j >\}$，其中可选项$TZone$为时间范围；$Property$为属性集合，包含实体标量属性集$PScalars$、实体向量属性集$PVectors$，其中，$PScalars = \{< entity_k, Ps_j > [, TZone]\,[, Loc] \,|\, j = 1, \dots, l\}$，其中可选项$Loc$为位置标签，$PVectors = \{< entity_k, < \{Proi_1, Val_1, w_1\}, \dots, \{Proi_m, Val_m, w_m\} >> [, TZone][, Loc][, \emptyset]\}$，其中$\emptyset$表示概率。

MDATA=<关联表示，关联构造，关联计算>

关联表示:

超语义图 = <*Concept*， *Entity*， *Relation*， *Property*>

$Concept=\{concept_i|i=1,\ldots,n\}$，概念集合

$Entity=\{entity_i|i=1,\ldots,n\}$, 概念的实例集合

$Relation=\{R_{ij}[,TZone][,Loc]|R_{ij}=<entity_i,entity_j>\}$，关系集合，*TZone*, Loc为时空属性

$Property=\{PScalars, PVectors\}$，属性集合，包含标量属性集和向量属性集

$PScalars=\{<entity_k,Ps_j>[,TZone][,Loc]|j=1,\ldots,m\}$，实体标量属性集，可选项*Loc*为位置标签

$PVectors=\{<entity_k,<\{Pro_{i1},Val_1,w_1\},\ldots,\{Proim,Valm,wm\}>\},[,TZone][,Loc][,\Phi]\}$，实体向量属性集空间，Φ为概率

关联构造:

构造算子 =$\{ConsOps|$ NER, RE, PE, InferOps, Verify, ...$\}$ ，包括概念构造，实例识别，关系抽取，属性抽取,知识推理，知识验证等算子

关联计算:

计算算子 =$\{QueryOps|$ SubG, ScopeQ, EQ, PQ, PathQ, kNN, ...$\}$ ，包括子图匹配，范围查询，实体/关系/属性查询，路径查找，k近邻计算等算子

图 1 MDATA 数据模型

同时，MDATA 还包含两个方面的算子：关联构造和关联计算。其中，关联构造用构造算子表示，可以表示为{ConsOps | NER, RE, PE, InferOps, Verify, ….}，包括概念构造，实体识别，关系抽取，属性抽取，知识推理，知识验证等算子；关联计算用计算算子表示，可以表示为{QueryOps | SubG, ScopeQ, EQ, PQ, PathQ, kNN, ….}，包括子图匹配，范围查询，实体/关系/属性查询，路径查找，k 近邻计算等算子。

图 2 是一个 MDATA 知识表示的一个具体示例。

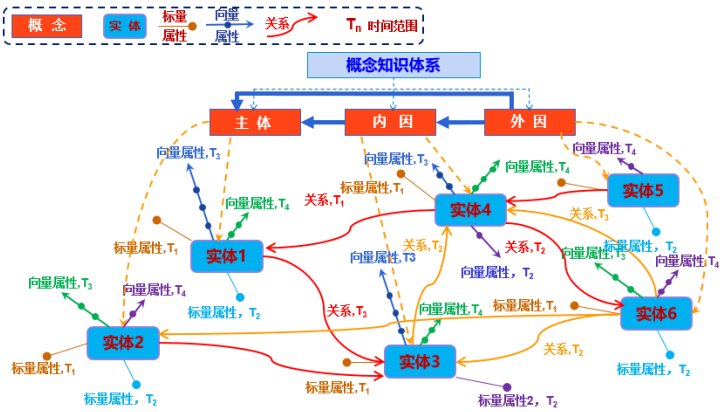


图 2 MDATA 知识表示示例图

11.3.2.3.2. MDATA 模型特性

为将 MDATA 进一步细化，能完成从理论到技术到工程实现落地，贾焰团队在原有的知识图谱的本体理论基础上，做出进一步改进。相关符号定义如下：

1、本体分为主要本体(Primary Entity，PE)和次要本体(Secondary Entity，SE）两类，其中 PE 表示在知识体系中最主要的本体，可以和知识图谱的本体对应，在知识体系中 PE 的名字是唯一的，具有相同名字的两个 PE 需要

合并为同一个；而 SE 表示次要的实体，例如属性值等，SE 一般不会参与本体的计算中，所以单独列出；

2、关系（Relation，简写为 R），关系描述的是 PE 和 PE 之间的联系，即主要本体之间的关系，与知识图谱的关系一致；

3、属性（Property，简写为 P），属性描述的是 PE 和 SE 之间的联系，即主要本体和次要本体的关系，可以理解为描述主要本体 PE 的属性。

通过这些定义，可以实现 MDATA 的如下特性：（1）通过区分 PE、SE、R 和 P，可以将关系和属性区分开，这是知识图谱尚未实现的；（2）PE 的名字是唯一的，也就是知识库中只有一个唯一的 PE_{id}；而 SE 的名字可以重复，比如 PE_1 和 PE_3 都可以有相同的 SE_1；（3）关系 R 可以分为单向、双向，表示单向关系和双向关系；（4）属性 P 仅单向，由 PE 指向 SE，表示 PE 具备 SE 的属性值；（5）PE 和 SE 之间仅存在唯一的边，不会出现重边的情况；（6）对于时间属性，时间区间用 TimeZone=[t_{begin}, t_{end}]来表示，时间点用 TimeZone=[t_{begin}, ∞]来表示；而空间属性以单一属性值 Spatial 来表示;（6）PE_i 与 PE_j 存在的关系 r_k，只会出现一次，即某种关系是唯一的；（7）PE_i 与 SE_j 存在的属性 P_i，P_k 可以出现多次，例如 PE_i 的属性 P_k 是 SE_j，PE_i 的属性 P_i 还可以是 SE_l，一个 PE 的属性可以有多种属性值。例如一个人的兴趣可以是足球、也可以是篮球，那么足球、篮球是两个不同的 SE，而属性 P 则是兴趣；（8）关系 R 不存在值域，而属性 P 可以理解为存在值域；（9）对于每个属性 P_k，其值域可以看作是 SE 的集合，即 P_k 的值域可以记为 $S(P_k)=\{SE_i,SE_j,\ldots,SE_p\}$;而不用属性的值域可以相交，即 $S(P_i)$和 $S(P_j)$的交集可以不为空。例如属性（出生地）为四川，属性（常驻地）也可以为四川。

PE 与 PE 之间可能存在多个关系 $R_1,\ldots, R_n$，一种方法是做成关系集合，另一种方法是每个关系一条边。下一节将介绍关系如何实现。

11.3.2.3.3. MDATA 模型整体表示

在 MDATA 模型中，每个关系表示为一条边，这个方法可以有效将关系 R 的单向、双向特性表示出来，其整体表示形式如图 3 所示。

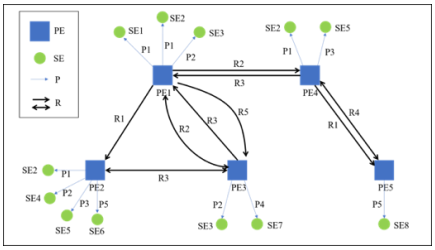


图 3 MDATA 整体表示形式

在对 MDATA 的整体表示中，需要引入 MDATA 中的时空特性，主要体现在两个方面：

1、关系 R 上存在时间、空间特性，即 PE 和 PE 之间的关系有可能随着时间、空间发生变化，因此需要描述这种变化；

2、属性 P 上存在时间、空间特性，即 PE 的属性值 SE 也是有时间、空间特性限制的。

下一小节将对时空表示细节进行介绍。

11.3.2.3.4. MDATA 模型时空表示

为了更好地进行可视化展示、借鉴成熟的数据管理方法，下面介绍两种方法对关系 R 进行展示，都采用（E,R,E）（实体，关系，实体）三元组的形式。

1、MDATA 中结合时空特性的关系 R 表示方法

如图 4 所示，对于图 3 中的单向的关系 R_1，表示为一个新类型的节点（图 4 中的六边形），并通过六边形引出时空特性。具体而言，对于原本的关系 (PE_1,R_1,PE_2)，拆为两部分：$(PE_1,null,PE_1\text{-}R_1\text{-}PE_2)$和$(PE_1\text{-}R_1\text{-}PE_2,tail,PE_2)$，其中 $PE_1\text{-}R_1\text{-}PE_2$ 为增加的表示关系 R_1 的节点。

在新增加了关系节点以后，三元组为（PE,X,PE-R-PE）、（PE-R-PE,X,PE）两种，其中 X 有三个属性，分别为 null、tail、head，null 表示不需要箭头，tail 表示在连接尾部实体的地方添加箭头，head 表示在头部实体的地方添加箭头。箭头设计方式主要是方便后续的可视化。

属性节点 $PE_1\text{-}R_1\text{-}PE_2$ 引出时空特性，分为三类：（1）TimeZone：表示单纯的时间区间，如[2001,2004]，[2001, ∞]等；（2）Spatial：表示单纯的空间信息，如北京、某 IP 地址等；（3）T-S：表示结合时间区间、空间的特性，描述为[2019,2020, 广州]（和 SE 类似，TimeZone/Spatial/T-S 三类特性都可以出现多次）。通过时空特性的表示，当时空知识发生变化时，直接修改时空特性数值即可，不用修改过多内容，包括图的结构等均不需要修改。

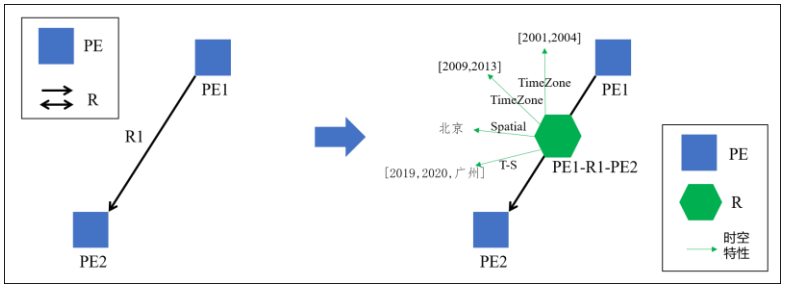


图 4 MDATA 中单向关系 R_1 的表示方法

图 4 的转换过程可用表 1 中的三元组进行表示：

表 1 转换的三元组

PE1，null，PE1-R1-PE2
PE1-R1-PE2，tail，PE2
PE1-R1-PE2，TimeZone，[2001,2004]
PE1-R1-PE2，TimeZone，[2009,2013]
PE1-R1-PE2，Spatial，北京
PE1-R1-PE2，T-S，[2019,2020,广州]

对于图 3 中的双向的关系 R_2，可以拆分为(PE_1,head,PE_1-R_2-PE_3)和（PE_1-R_2-PE_3,tail,PE_3），如图 5 所示，其他与上述表示形式一致。

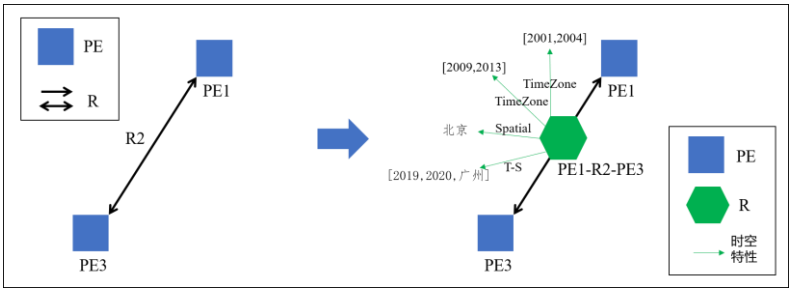


图 5 MDATA 中双向关系 R_2 的表示方式

该方法的不足之处在于：由于需要增加新的关系节点，就一个关系 R_k 而言，如果有 N 个节点，有可能存在 N^2 个新增的节点，而对于这些新增加的节点，需要一个很简单的方式进行快速检索，一种思路是根据 PE_i-R_k-PE_j 中的 R_k 进行检索，为了增加检索效率，可以将 PE 和 PE 图之间的三元组改进为四元组，如表 2 所示：

表 2 三元组转换成四元组

(PE_1,X,PE_1-R_2-PE_3)->(PE_1,X,PE_1-R_2-PE_3,R_2-head)

$$(PE_1\text{-}R_2\text{-}PE_3, X, PE_3) \text{->} (PE_1\text{-}R_2\text{-}PE_3, X, PE_3, R_2\text{-}tail)$$

这样相当于将原来的关系 R_2 扩展为 R_2-head、R_2-tail 两个，然后用他们建立检索，可以提高检索效率。

2、MDATA 中结合时空特性的属性 P 表示方法

下面介绍如何在 PE 和 SE 之间的属性 P 加入时空特性。由于 SE 节点可以当作是比较基本的节点，故不在 P 上新增节点，而是直接在属性 SE 的属性值上增加时空特性，即 SE 节点本身就表示属性值，而由属性值引申出相关的时空特性。

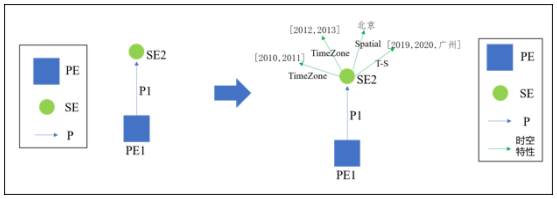


图 6 MDATA 中属性 P 的表示方法

如图 6 所示，属性 P_1 的属性值为 SE_2，但是 SE_2 存在时空特性，如居住地、喜好等可能存在时空特性，时空特性也和关系 R 一样，分为三类：（1）TimeZone：时间特性，表示时间区间；（2）Spatial：空间特性，表示空间信息，如位置、经纬度、IP 地址等；（3）T-S：结合时间区间、空间的特性。这种表示方法很容易可以看出，图 6 也可以用三元组进行表示,展示方式如上一小节所示，此处不再赘述。

11.3.2.3.5. MDATA 模型表示架构

结合上一节的描述，下面探讨两种针对 MDATA 的表示架构，一种是将所有表示都放入一个图里面，生成一个大的图，包含了具有时空特性的关系、属性等；另一种将 MDATA 知识库分为大图、小图等多个不同层级的图，这样可以提高图计算的效率。下面将详细描述两种架构并分析相应的优缺点。

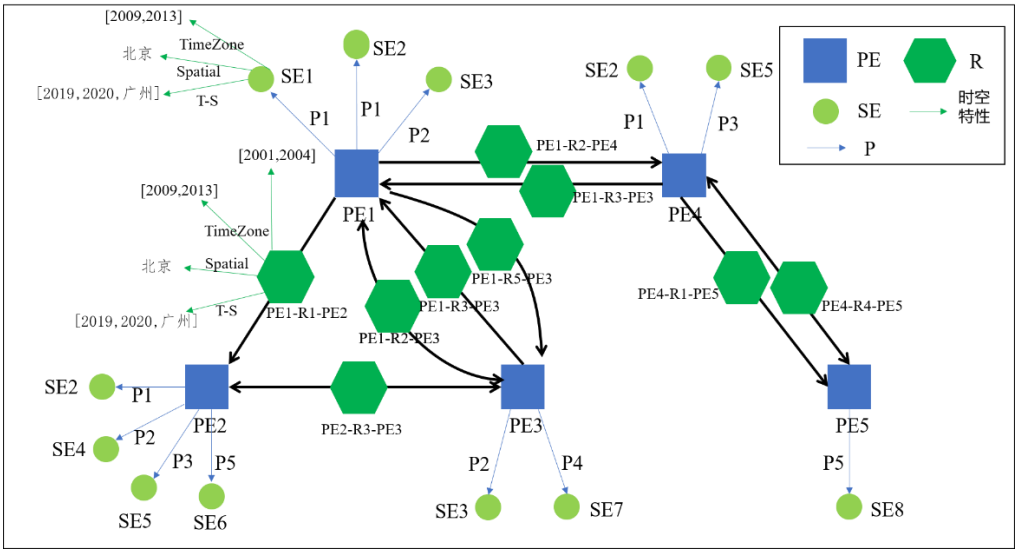


图 7 基于大图架构的 MDATA 表示

图 7 展示了大图架构，图中的节点包括以下四种类别：(1)PE 节点，表示主要实体；(2)SE 节点，表示次要实体，代表的是属性值；(3)改进后的 R 节点，表示为 PE-R-PE，即 PE 和 PE 之间的关联；(4)时空特性的数值节点，由于时空特性的数值节点为子节点，可以不展示节点，直接展示时空数值即可,其中时空特性的数值包括三种：TimeZone（时间区间）、Spatial（空间数据）、T-S（包含时间区间和空间数据）。

图中的边包括以下三种类别：(1)PE 之间相连的边，已在大图中改为 PE 和 R 之间相连的边，如图 7 中的黑色边，包括两边无箭头、一边有箭头两种；(2)表示属性 P 的箭头，如图 7 中的蓝色边表示；(3)时空特性的箭头，如图 7 中绿色边，即包括关系上的时空特性，也包括属性上的时空特性，其中时空特性的边分为三类：TimeZone、Spatial、T-S。

大图架构具有如下特点和优势：（1）能直观地将时间、空间、时空融合的特性表示在图中，这是当前知识图谱所不具备的；（2）通过定义 PE、SE 的概念，将关系、属性进行很好的区分，这样在计算关系、属性的时候能提高效率；（3）能继续沿用三元组的表示方法，这样可以很好地利用已有的成熟技术进行存储、管理以及可视化展示。

同时，大图架构也存在一些不足之处：（1）原本的关系 R 在 MDATA 描述中增加了复杂度，比如原本只有 M 种关系，在图中需要增加很多节点，最多可能增加 N2M 个节点，增加的节点个数和边的数目一致；（2）节点数目较多，加入时空特性以后图的规模更大;（3）当前的成熟技术不能完全采用，需要重新规划。

多级图架构表示成多种子图分级展开，主要分为以下三类子图：（1）PE 主图，仅包含 PE 节点和关系 R 的主图，对于知识库中的 PE 主库；（2）属性图，

点击某 PE 节点，可展开该 PE 节点的所有属性节点 SE 以及属性节点 SE 对应的时空特性，对应 PE 属性知识库；（3）关系图，点击某关系 R，可展开该关系 R 的时空特性，展开时可以出现六边形的节点形式即可，对应关系知识库。

下面将给出具体介绍。

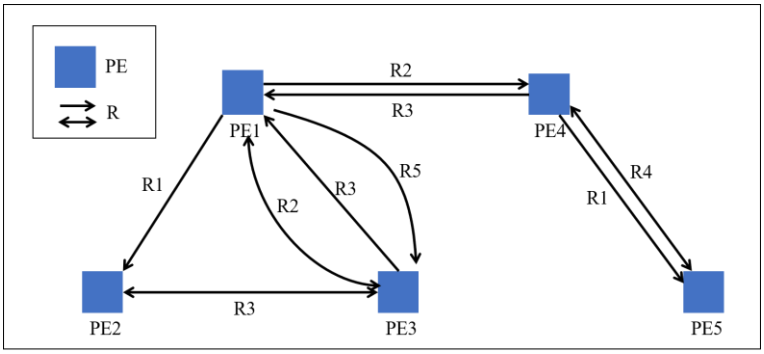


图 8 PE 主图

如图 8 所示，第一级图为 PE 主图，是整个知识库的核心，记录不同 PE 之间的知识联系和关系。此图可以很好地使用已有的管理和存储方式，比如三元组（PE，R，PE）。

知识库关联可以理解为 PE 之间形成了知识库，记为 PE 主库，PE 主库保存的数据最多为 N2M 条，其中 N 为 PE 个数，M 为关系个数。

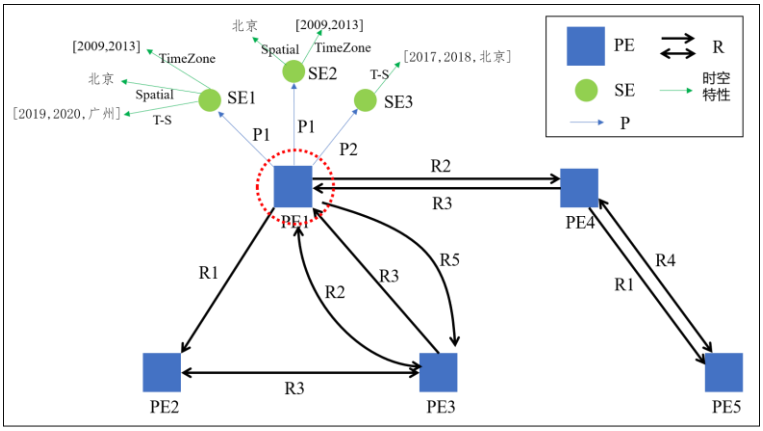


图 9 具有时空特性的属性子图

点击任何一个 PE 节点，可展开属性子图，如图 9 所示，点击 PE1 以后，可展开所有属性对应的属性图，包括每个属性值包含的时空特性。该图的形成方式也容易实现，可通过三元组方式进行保存即可。

知识库关联相当于对每个 PE 存储一个属性知识库，而每个 PE 的属性知识库可以看做是 PE 主库中的 PE 名字关联得到的知识库，因此总共有 N 个属性知识库，每个属性知识库对应一个 PE。

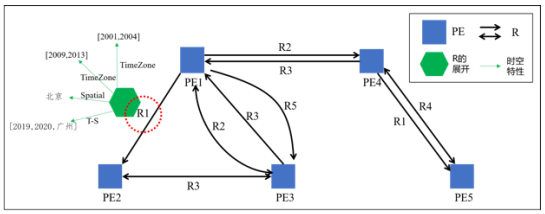


图 10 具有时空特性的关系子图

类似地，点击主图中的关系节点，可以展开关系的时空特性，如图 10 所示，点击关系 R1，可展开 R1 的时空特性，可视化的时候可以将 R1 扩展成一个六边形节点，然后扩展出时空特性。同样的可以通过三元组的形式实现。

对于每种关系建立一个关系知识库，因此总共会建立 M 种关系知识库。关系知识库需要单独设计，例如对于关系 R 的每一条实例（PE1-R-PE2），建立一条知识，如表 3 所示：

表 3 知识建立过程

(R-head,R-tail)：(PE_1,PE_2)
R-direction：(单向或者双向)
R-TimeZone:记录时间区间的集合
R-Spatial：记录空间特性的集合
R-TS：记录结合时间、空间融合特性的集合

多级图架构有如下特点和优势：（1）表示方式更直接，将 MDATA 知识库分为主库、属性库、关系库三种，分别对应可视化时的 PE 主图、属性图、关系图；（2）表示简单，逻辑性强，主图、次图的关系十分清晰；（3）关系 R 的处理上不像第一种表示方法，需要将关系扩展为很多个扩展节点，这种方法不需要扩展多的节点，在进行关系查询、关联、管理的时候更容易；（4）知识动态变化时更新更容易，例如时空特性数值变化时，不需要对主图进行修改，只需要对子图（如属性图或关系图）的数值进行修改即可。但是，多级图架构需要单独保存 N 个属性库和 M 个关系库，如果 PE 个数太多，需要保存的库的数量也会很多。

随着数据的爆炸式增长以及应用的不断拓展，传统的知识图谱表示模型面临时空特性缺失、无法区分关系和属性、不能处理向量化属性等问题。本节提出一种新的多维关联超知识图谱表示模型 MDATA，给出了具体定义、分析了相关特

性，并且提出了两种架构，分析了相应的优缺点。下一步需要研究在这样的方式下如何实现各种计算，如关系查询、搜索、关联计算等。另外就是如何使用MDATA表示方法解决多领域知识的统一表示、已有知识图谱知识库等方法的映射、多领域知识关联、知识动态优化、知识准确性验证等难题。

11.3.3. 用户搜索意图准确理解与表示

用户搜索意图准确理解与表示定义是基于用户查询输入的关键词、语音、手势等内容，在语义级上准确理解用户的意图，并以支持高效查询推演的统一模型进行表示。通过将搜索输入内容转换为机器可识别的表示语言，深度学习用户思维，统一搜索查询视图，从而将用户搜索转换为机器可识别的语言模型，便于机器理解搜索意图。涉及的关键技术包括：基于时空特性的用户意图理解；基于统计分析的用户意图理解；基于形体动作的用户意图理解；基于情感分析的用户意图理解；交互式用户意图理解。

用户搜索意图准确理解与表示的目标是基于用户查询输入的关键词、语音、手势等内容，以及用户所处场景及上下文，在大数据环境下进行更加准确的意图理解，并采用支持高效查询推演的统一模型进行表示。主要技术包括：

（a）搜索意图的统一表示和语义建模：面向多模态数据的语义级用户意图的统一表示方法；用户意图时空特性的表示方法；用户意图的场景相关特性表示方法；用户意图的情感相关特性表示方法等；

（b）语义级用户意图准确理解方法：基于上下文感知的用户意图理解方法；基于时空特性的用户意图理解方法；基于统计分析的用户意图理解方法；基于情感分析的用户意图理解方法；基于事件推演的用户意图理解方法；多维度综合的用户意图理解方法；用户意图理解评价模型和方法等。

下面从包括单来源用户真实意图理解、协同式用户意图理解方面两个方面对研究现状进行综述。

11.3.3.1.单来源用户真实意图理解

在基于文本的用户真实意图理解方面。2004 年美国华盛顿大学 Oren Etzioni 等人提出了基于规则模板抽取实体/概念之间的关系来描述和理解搜索意图。2007 年美国华盛顿大学 A Yates 等人提出了无监督学习方法改善了 Oren Etzioni 方法需要人工定义规则的缺点。为了进一步提高关系抽取的准确性，2012 年，威斯康辛大学麦迪逊分校 W. Wu 等人采用基于语义的迭代方法抽取出更多更准确的实体间的语义关系。2011 年，印度尼赫鲁科技大学 G.Madhu 等人利用语义网

工具和技术提供分层模块的方法解决搜索引擎对语义内容的理解。2011 年，由卡耐基梅隆大学和阿默斯特大学持续研究开发的开源 Indri 搜索原型系统，利用 Petri 推理网络模型的优势来支持较复杂的结构化查询，利用语言模型及平滑技术对推理网络中的节点进行有效评估，从而达到良好的查询效果。2020 年，美国马里兰大学 Han Lushan 等人提出了基于本体的智能信息检索理论，结合统计学和语义相似度判读理解用户的意图。慕尼黑大学计算语言学系持续研究的 Scirus 项目建立了涵盖所有专业科学领域的超过 50,000 个叙词的科学叙词表，以保证检索效率。系统对每次搜索到的信息内容会自动抽取反映主题内容的关键词，以提高搜索的专指性，这是一般的搜索引擎所无法比拟的。

在基于图片和视音频的多模态意图理解方面。现实世界是多模态交互式的，因而查询的对象也应该是多模态的。由于多模态信息的异构性，基于多模态查询的交互协同检索的用户意图理解更具有挑战性。受益于计算机视觉技术的迅速发展，传统的方法可以采用计算机视觉方法将多模态信息转化为语义信息，这些语义信息通常都是以文本词或者句子表示。此外，引入直接或者间接的反馈可以进一步改进用户意图的理解，缓解用户查询过程中面临的 Gap。例如，Tuukka Ruotsalo等人提出交互意图建模通过计算建模（针对交互进行可视化呈现）增强人类信息探索能力，同时通过用户界面帮助用户进行搜索和探索。然而多模态信息无法全部的通过文本语义信息表示。因为这样的语义信息无法准确的传达多模态信息其他很多方面的信息，比如视觉信息所蕴含的视觉风格，心理视觉因素等。因此语义信息和多模态内容需要综合考虑，需要在包含文本和视觉等的多模态信息与用户行为和意图之间建立一种映射。

11.3.3.2.协同式用户真实意图理解

在基于用户个人画像的意图理解方面。社会媒体数据包含有丰富的（个人或群体）用户的行为及属性信息，基于社会媒体进行个人画像，然后进行意图理解是一个重要研究方向。社会化检索和推荐可以整合内容和社会网络信息来提高相关信息的发现与推荐的用户满意度。通常，这些社会网络信息包括：个体用户信息，例如用户过往浏览打分纪录、时空信息、阅读水平等；以及群体用户信息，例如同社区用户和相似用户的信息以及社会化标注。从社会媒体上获取情感（sentiment）也成为了当前的一个研究热点。如何能根据用户实时的互动行为（如根据用户对某些新闻的评论、用户之间的对话）实时得总结出当前用户的情感极性，亦会对理解用户搜索意图起到非常重要的作用。Zhang C.等人提出了“社会化搜索引擎”的概念，并实现了一个融合探索式信息检索功能和在线社会化聊天功

能的社会化搜索引擎原型系统。工业界，例如谷歌，也开始尝试将聊天功能和搜索功能做简单的结合，如某融合了搜索引擎功能手机输入法可以将置信度较高的搜索结果摘要（人物，地点等）直接发送到聊天窗口，一些即时聊天工具会根据当天聊天内容推荐一些用户可能感兴趣的内容或广告。但是，学术界还没有针对社会化信息探索背后的认知特性展开深入的研究，我们还不知道用户在这个过程中认知状态如何变化，因此，我们应该进一步推进社会化信息探索的应用，以更好的满足用户的信息需求。

在交互式意图理解方面。在信息检索研究中，通过分析用户交互模式特征进行隐相关反馈进而提高文档排序准确性的研究日益增多，例如基于查询历史和文档点击历史，也包括其它交互特征如停留时间、显示时间和滚动条、视线追踪和面部表情等等。这些交互信息可以用来预测用户的潜在需求，Dupret G.等人提出了一个模型，基于过去点击历史和一些用户浏览行为的简化假设以预测在初始搜索结果中的文档被点击的概率。更进一步的是，最近 White R. W.等人研究了利用搜索/浏览路径（从每个点击的初始搜索到路径终点的群体用户的浏览路径）来建立预测模型以寻找每次搜索结果的最优路径。West R.等人提出建模用户交互行为的动态性以根据以往的交互历史来预测进一步的用户行为。Radinsky K.等人采用眼动实验和鼠标痕迹等对用户的阅读和点击行为进行了深入分析，扩展了点击模型进行行为建模，在搜索引擎排序过程中取得了很好的效果。Liu C.等人从用户认知判断着手，关注于查询意图理解及相关性判断，提出 DeepRank 框架去模拟人类判断过程，并取得了更好的实验效果。

在语义级用户意图理解方面。最早的提出智慧搜索引擎概念的是 2009 年美国沃尔夫勒姆研究公司开发的 Wolframalpha 搜索引擎，其创新之处在于从公众的授权的资源中发掘、建立起一个异常庞大的、经过组织的数据库，再利用高级的自然语言算法进行处理，最终能够马上理解用户问题，并直接给出答案。2012 年 5 月 Google 发布了整合大量开源知识库并加入用户数据沉淀的知识图谱搜索功能。他从三方面提升 Google 搜索效果：①一个搜索请求可能代表多重含义，找到用户最想要的那种含义；②可以更好的理解用户搜索的信息，并总结出与搜索话题相关的内容，提供最全面的摘要；③让搜索更有深度和广度。Google 发布知识图谱半年之后，2012 年底搜狗发布了第一个中文的知识图谱搜索引擎——知立方。知立方通过整合海量的互联网碎片化信息，引入“语义理解”技术，试图理解用户的搜索意图，对搜索结果进行重新优化计算，直接给出用户想要的准确答案。2013 年，百度成立了百度深度学习研究院（Institute of Deep Learning，IDL），开始将知识图谱的技术应用到搜索技术中，融入了人和人之间的关系、物和物之间的关系，推出了下一代智慧搜索引擎雏形——“知心搜索”，在给出常规的搜

索结果的同时，给出相关的百度百科等内容。2013 年，微软（Bing 搜索引擎）也从五个方向对未来搜索引擎进行了的战略性思考。第一是从组织所有相关的网页信息，到直接关注用户的搜索目的；第二是建立知识库，利用各式各样的挖掘技术，把结构性 Web 中的对象关系抽出来，以知识的方式来表示；第三是语义的检索与任务完成；第四是从搜索内容走向搜索应用和服务；第五是云平台和建立生态系统。试图在用户的问题上增添更多内容以便直接得出用户想要的答案。

在基于知识图谱的用户意图理解方面。早期的基于知识图谱的用户意图理解研究主要以基于形式文法的 Pythia 系统为代表，该系统通过一系列语法规则构造一系列包含句法及语义信息的字典，通过字典及语法规则将问题解析成 SPARQL 查询语句，从图数据库中返回答案。TBSL 及 LODQA 等系统试图通过基于模板的方法来解决面向知识图谱的 QA 问题，这些系统首先将问答语句转化为一个 SPARQL 查询模板，模板中的一系列命名实体用变量代替，模板构造完成后，根据语法分析找出语句中的动词、名词、形容词等，并将其与知识图谱中的实体概念匹配，并生成 SPARQL 查询语句。

总的来说，当前的用户搜索意图理解还存在以下问题和挑战：（1）大部分搜索意图理解技术还是基于单通道/模式的，已经到了极限无法提高；（2）缺乏有效的模型和技术来融合搜索行为、用户偏好、社会关系等信息来提高意图理解的水平；（3）无法充分有效地利用搜索上下文，特别是时空特性。

11.3.4. 用户意图高效准确匹配与推演

用户意图高效准确匹配与推演是大搜索的关键步骤，是指运用统一表示对用户意图在知识聚合中进行匹配，求解问题，并给出一组有序的推荐解答方案的过程。大搜索引擎中，由于用户表达意图的信息已经不仅局限于简单的文本串，而是一个复杂的关联图结构，与之相适应，其模式匹配方法也不同，例如通过文本关键词匹配和子图查询等来实现搜索意图与搜索空间中目标项的匹配等方法。涉及的关键技术包括：基于实体关系模型的知识聚合、管理与更新、基于文本模型的匹配技术、基于图模型的匹配技术、基于音视频数据的匹配技术、面向搜索目标的解答排序与评估技术等。主要可以概括成以下两点：

用户意图的高效匹配和推演方法的目标是指运用统一表示的用户意图在知识图谱中进行匹配推演，求解问题，并给出一组有序的推荐解答方案的过程。主要技术包括：

（a）面向用户意图的解答排序与评估技术：研究异构信息聚合搜索评价技术，分析服务信息源和用户意图的关系，评价返回的各种类型的信息之间的相互

作用、信息源的排序来综合评价整体结果质量；研究搜索结果评估体系，主要实现不同设备上的搜索体验的评估；针对大搜索下的用户行为分析与建模，建模评价需求和目标的用户满意度等。

（b）基于图模型的搜索意图匹配技术：大图的高效索引和分布式组织管理技术；大图划分和分布式缓存理论与方法；面向大图结构的特性分析技术，基于大图的高效查询及其优化技术；基于大图的用户意图高效推演技术等。

下面从包括关键词倒排索引与匹配、大图计算平台的支撑方面两个方面对研究现状进行综述。

11.3.4.1.关键词倒排索引与匹配

信息检索方法。Yao X.等人提出的 pagerank 主要是根据网页间的相互链接关系，对网页进行重要性评估，从而对候选结果进行有效排序；Bast H.等人采用 learning-to-rank 方法更好地解决固有的实体识别问题，用三个预定义的模板来产生答案候选集，具有更加丰富的语言学特征，如重叠单词的数量，派生单词，单词向量嵌入余弦相似性被用于训练问题关系的对齐模型。Dubey 等提出了一个名为 EARL 的框架，它将实体链接和关系链接作为一个联合任务执行。EARL 实现了两种不同的解决方案策略，第一种策略是联合实体的形式化和将任务链接起来作为广义旅行商问题（GTSP）的一个实例；第二种策略使用机器学习来利用知识图中节点之间的连接密度，它依赖于三个基本特征和重新排序步骤，以预测实体和关系。Zheng 提出了一种数据+oracle 方法来回答知识图上的 NLQ，让用户在查询理解期间验证模糊性，同时正式化了交互问题并设计了一个有效的策略来解决问题以降低交互成本，并且通过利用与用户的交互中的延迟来提出查询预取技术。

语义解析方法。Bordes A.等人通过学习单词和知识库成分的低维嵌入，用于根据候选答案对自然语言问题进行评分，使用成对的问题和答案的结构化表示以及问题释义对训练系统；Jonathan Berant 等人改进了 SEMPRE 系统，在此前基础上提出了基于代理的语义解析方法（Agenda-based parsing），并引入模仿学习（imitation learning）对生成的逻辑表达式打分，这种方法提高了解析效率及性能，是一种由强类型约束引导的逻辑形式驱动的解析算法；Andreas J.描述了一个适用于图像和结构化知识库的问答模型，使用自然语言字符串从可组合模块的集合中自动组装神经网络，提出动态神经模型网络方法，在视觉和结构域中的基准数据集上实现了最先进的结果。

深度学习方法。Dong L.等人引入了多列卷积神经网络（MCCNN）来从三个

不同方面（即答案路径，答案上下文和答案类型）理解问题并训练实体、属性、答案类型的 Embeddings，然后利用训练好的 Embeddings 及神经网络来评估问题和以答案为中心的子图之间的语义相关性，对答案进行打分和排序；Gao W. Y. 在 Semantic Parsing 的基础上尝试运用卷积神经网络来评估问题与逻辑表达式之间的语义相关性，利用 QA 问答对作为训练集，训练问题的词序列的 Embeddings 与逻辑表达式词序列的 Embeddings，然后利用训练好的 Embeddings 及神经网络对问题与逻辑表达式的语义相关性进行打分并排序；Bordes A.采用激进的学习方法将问题映射到矢量特征表示，通过将答案映射到相同的空间，可以独立于其模式查询任何知识库而无需任何语法或词典，采用新的优化程序进行训练，结合随机梯度下降，然后使用通过自动混合和协作生成的资源提供的弱监督进行微调步骤。

11.3.4.2.大图计算平台的支撑

子图的答案匹配和检索技术需要大图计算平台的支撑。大图计算平台的研究工作主要包括大图的划分与索引建立、大图的分布式并行计算和图数据管理系统等方面。

大图的划分与索引方面。2014 年，微软的 Bourse 等提出了关于图平衡的边分区近似算法，并且量化了边分区针对点分区的优势。2016 年，耶鲁大学的 Huang 等针对动态图提出了轻量级的边分区算法，并且证明了在图变化的情况下，可以达到和当前最优的静态分区算法相当的性能。2004 年，伊利诺伊大学香槟分校的 Yan 等提出了基于频繁子图的图索引结构，达到减少索引结构的目的。

大图的分布式并行计算方面。2013 年，国立雅典理工大学的 Afrati 等提出了基于 MapReduce 的超立方体算法，解决了分布式环境下子图罗列问题。2016 年，马里兰大学的 Quamar 等提出了 NScale 系统，使用户能够以子图为单位进行程序设计，实现了对复杂图问题的高效处理。2017 年，华为美国研究院提出了图引擎平台 EYWA，提供了从图存储和管理、高性能计算引擎，到图分析、图查询的一整套解决方案。目前国外关于大图的并行计算的研究多关注图本身，通常忽视在特定应用中的语义信息，用通用方法处理所有类型图数据，容易造成不必要的通信开销和引起负载不均衡等问题。在图的分布式并行计算方面，现有系统多是针对基于遍历运算和迭代操作的算子进行优化，而对于大图的结构操作，如知识图谱中常用的子图匹配计算，并行优化效率不高。

图数据管理系统方面。图数据具有很好的扩展性，图数据的关键词查询方式可以方便地检索结构化数据、半结构化数据和非结构化数据。较之像 SQL、

SPARQL 之类的结构化查询，关键词查询不需要使用者了解查询语言的语法，也不需要用户了解底层数据如何存储，只需要输入待查询内容的关键词，算法会自动查询相关结果。常用的图数据查询语言包括 SPARQL、Cypher、Gremlin 等。其主要用途是针对子图匹配运算（如 SPARQL）、图上的遍历运算（如 Cypher 和 Gremlin）等。在图数据的关键词查询问题中，查询结果的形式多种多样，有最小树、关联连通簇、r-半径斯坦纳图、r-极大团、多中心导出子图等。由于图数据模型在网络空间大搜索应用中具有重要作用，相关搜索引擎公司也开发了相应的图数据系统，如 Google 公司开发的 Pregel。

总的来说，在答案匹配和检索方面，面临着以下几个方面的挑战：（1）基于限定语料的词义扩展匹配无法适应泛在网络开放数据；（2）语义匹配大图索引与搜索方面，已有基于图推理的方法，都只考虑封闭式的小规模知识库，无法支持海量巨规模的开放知识库；（3）基于纯粹的匹配排序无法获取隐藏的知识；（4）现有图计算平台所定义的算子大部分基于图遍历运算和迭代操作，对于图上结构操作（如子图匹配运算）的管理与分析效率不高，同时对于不断更新的动态图，现有系统的查询处理的效率较差。

11.3.5. 大搜索安全可信与隐私保护

大搜索安全可信与隐私保护是指大搜索从用户意图理解、数据获取、知识综合到返回智慧解答结果，整个生命周期过程是可信的、安全的、支持隐私保护和有害信息过滤的，是大搜索的基本保障。其中，可信是指大搜索数据来源正确、权威，并可溯源；安全是指大搜索的结果不会被非授权用户滥用；隐私保护是保证用户隐私（个人、位置等信息）不会在搜索过程中被违规泄露；此外，还支持对暴力、色情等有害信息的精确过滤。大搜索强大的关联分析能力，可以对许多重要的事和重要的人的方方面面进行关联和归纳，形成新的认知和知识；这些内容可能是个人的隐私，乃至国家的机密，因此严格的访问结果内容研判和控制，是大搜索推广使用的前提。涉及的关键技术包括：数据源可信、抵抗关联分析的隐私保护、粒度可控的访问控制和定向过滤等。

大搜索安全可信与隐私保护目标是保证用户搜索的结果是可信赖的，保证合适的搜索结果只返回给合适的用户，而不被滥用，保证用户的隐私不被泄露，并过滤掉暴力、色情等恶意信息。

大搜索安全可信与隐私保护技术主要解决源数据获取、融合分析、结果返回使用等环节中的信息来源可信、数据访问安全和隐私泄漏保护等问题。主要技术包括：

（a）数据源可信与信息溯源技术：研究数据源可信方法，包括数据来源真实性的快速验证、不完整数据快速清洗与恢复、数据质量管理机制与方法；研究数据在演化过程中的纵向溯源演化的理论模型和方法；研究搜索结果的推理过程溯源方法；

（b）细粒度的搜索访问控制技术：研究支持数据复用的访问控制模型及其动态策略调整机制；不同数据源综合结果的所有权动态划分及其访问控制；针对不同隐私保护方案的访问控制模型及其机制的融合、冲突消解等问题。

（c）防关联分析的隐私数据处理方法与技术：研究信息隐私与行为隐私的综合建模与测评；研究面向情景感知的深度融合隐私保护机制，研究面向搜索的高效隐私保护理论；研究设计能够抵御关联分析的隐私保护策略；研究隐私保护方案的动态调整机制，实现对海量用户的高并发隐私保护方案。

大搜索安全可信与隐私保护主要包括大搜索安全可信、隐私保护两个方面。

11.3.5.1.大搜索安全可信

在数据源可信认证方面。泛在网感知设备海量，分属不同的管理域。预计2020 年将有 10 亿的蜂窝 M2M 设备接入。大搜索过程中感知设备的认证和信任管理已经成为最主要的挑战。在特定的应用场景下，对搜索请求者保护感知设备位置等敏感信息也有迫切的需求。在感知设备认证方面，基于蜂窝网连接的感知设备认证可以基于 2008 年瑞士苏黎世联邦理工学院 Frank 等人提出的 3GPP 标准 SIM 的 AKA 协议提供对感知设备的认证。美国马萨诸塞大学的 Yan 等人也针对 M2M 群组认证给出了许多解决方案，然而这些设备管理和认证协议主要基于运营商提供，相关认证能力对应用层用户的开放仍有待于解决。行业用户也可能更倾向于自己管理自己的感知设备，传统 PKI 技术对于海量的感知设备管理而言显得力不从心。对于搜索得到的数据中，可能会包含有用户的隐私，对数据中隐私的甄别、度量和在结果返回前对数据进行隐私化处理是泛在网安全搜索的又一个关键问题，其中数据中隐私及可用性的度量是基础性科学问题。在隐私度量方面，信息论对于搜索结果数据的隐私量化评估提供有力的工具。目前基于信息论的隐私度量方面的研究，主要采用不确定度（Uncertainty）和信息增益（Information Gain）两种指标。

在可信度传播计算方面。如果攻击者在推测用户敏感信息时的不确定度越高，那么用户的可信程度就越低。2017 年英国曼彻斯特大学的 Aberer 等人提出了通过香农信息熵来度量可信度传播的计算方法。条件信息熵可用来衡量当攻击者已获得某一观察量 Y 后，其在推测敏感信息 X 时存在的不确定度。基于信息增益

的度量方法衡量了攻击者基于观察值可获得关于原始值的信息量。2017 年耶鲁大学的 Grosky 等人针对互信息的度量方法衡量了一个可信度传播的量化方法，提出了一种在一定的失真限制条件下来衡量数据库泄露的最少信息的方法。基于最大信息泄露（Maximum information leakage）的度量方法修改了互信息的定义，度量了在攻击者仅观察到一个输出事件 y 时，其能够获得的额外关于隐私事件 X 的最大信息量。通过对数据集隐私的信息论建模，确定隐私-可用性量化度量，对于选择最佳折中的隐私化机制，返回用户可接受的搜索结果同时保护数据关联方的隐私具有重要的意义。目前这方面的研究仅处于起步阶段。

11.3.5.2.大搜索隐私保护

在隐私访问控制研究方面。2011 年阿肯色大学 Yu 等人不仅实现了细粒度的访问控制，同时可以抵御如传感器妥协和用户勾结等攻击。2012 年亚利桑那州立大学 Zhang 等人提出了一种分布式令牌重用检测方案去防止恶意用户对令牌的重用攻击。2009 年，Frias 等人提出了基于行为个性的访问控制机制，利用提取的行为特征进行控制。2006 年加利福尼亚大学 Goyal 等人提出了一种 KP-ABE 方法将访问控制策略嵌入用户的私钥中，实现了细粒度的访问控制。2014 年香港城市大学 Yang 等人提出了一种可撤销的多授权机构的 CP-ABE 结构，有效的解决了属性的撤销问题。2014 年，IBM 的 Jan 等人提出了一种匿名的权限控制方法，在解决数据隐私性的基础上解决了用户身份隐私性问题。2008 年，伊利诺伊大学 Maji 等人提出了一种签名方式称为基于属性的签名（ABS）。2013 年，达姆施塔特工业大学 Bugiel 等人设计了联合控制灵活细粒度的强制访问控制方案。2013 年，波士顿大学 Rohrer 等人提出并实现了基于动态角色的访问控制方案来实施最小特权原则。2013 年，针对处理隐式访问信息的访问控制问题，IBM 的 Kapil 等人提出了适用于混杂移动应用的上下文感知权限控制方法。2015 年，亚利桑那州立大学 Ave 开发了称之为 Auto-FBI 的原型系统，实现了敏感数据的自动隔离。2013 年，特拉华州立大学 Hu 等人提出了一种对多人共享的数据的保护方式，设计并实现了一个多机构访问控制策略。2014 年，佛罗里达大学 Jung 等人对现有的 CRiBAC 模型进行了扩展，并保证了社交网络用户之间合作的安全性[173]。现有的访问控制研究大多围绕着感知层访问控制、基于属性的访问控制、面向身份隐私保护的访问控制、移动操作系统中的访问控制等技术展开。然而，针对泛在网搜索模式的扁平化、搜索用户的开放性与海量性、节点动态性等特征，海量动态用户访问权限实时更新和撤销的问题有待进一步研究与完善。

总的来说，在网络空间大搜索相关的计算体系结构、网络空间知识获取、知

识表示与管理、多知识库协同推理、用户搜索意图理解和可信搜索与隐私保护等方面，目前国内总体落后于国外，个别技术上国内外有较大差距。但近年来国内人工智能发展势头迅猛，迅速缩小了与发达国家的差距，有些技术已经处于国际领先水平。发展和综合运用人工智能 2.0 的新技术，发展新一代网络空间搜索技术是大势所趋。

11.4. 领域产业发展现状及趋势

网络空间大搜索是新一代具有“智慧”的搜索，力求准确洞察理解用户的搜索意图，在海量、多源、异构、多态、不确定的数据中，实现对与人物、物体和内容等相关信息的智慧搜索，为用户提供最贴切的搜索结果，这势必影响我国的社会、经济和生活等各个方面，具有广阔的前程和非凡的意义。

11.4.1. 满足国家安全需要方面

搜索引擎可以通过技术措施操控人们获得信息的范围，谁掌握了搜索引擎，谁就掌握了信息网络空间的入口，掌握了为人们提供信息甚至答案的权利，因而由此产生的政治、经济和社会驱动力日益受到各国重视。搜索引擎通过对用户的搜索问题进行战略性统计和计算，这将对国家、社会和商业具有重要意义。

搜索引擎与国家安全密不可分。美国国家安全局与 Google 公司签订了合作协议，可获得 Google 搜集的来自全球的海量信息。搜索引擎通过对用户的搜索问题进行战略性统计和计算，这将对国家、社会和商业安全构成严重威胁。2010 年初，俄罗斯提出建设全球首个国家搜索引擎，加入安全接入、过滤内容等。俄政府认为，搜索引擎是一种影响公众舆论的手段，将其纳入“国家基础设施建设”符合国家利益。德国宣称“有了自己的搜索引擎，就不用担心在文化、政治上被国外 “任意摆布”。美国军工公司研发“开源引擎”，搜索范围扩展到“暗网”，目的是捕捉到某些潜在危机的苗头。该引擎曾成功掌握了墨西哥头号武装贩毒组织的人员、装备、活动地区等情报，为保障美国的国家安全提供了可靠的数据来源。

大搜索保障国家信息安全。所谓国家信息主权是由经济主权、政治主权和文化主权派生出来并与新型信息网络空间相结合产生的，是当代国家主权的组成部分，是国家主权在信息网络空间的具体体现。保护国家信息主权就是保护国家具有允许或禁止信息在其领域内流通的最高权威，包括通过国内和国际信息传播来发展和巩固本民族文化的权力，以及在国内、国际信息传播中树立维护本国形象的权力，也应当包括平等共享空间信息、传播资源的权利。搜索引擎关系着文化与社会的信息安全，关系着信息处理标准掌握在谁的手上，也关系着一个多民族

多语言国家的文化传承。而且随着人们的生活、工作、学习、娱乐等越来越多地转移到互联网上，对这些海量搜索记录信息的整理和分析，无论是在经济领域、还是文化领域、还是国家信息安全都具有很高的价值。但是，我们也要清醒看到任何搜索引擎都是有立场，有价值取向的，这是外在价值作用的必然结果，不能指望自己国家的主流文化、价值观通过别国的搜索引擎来传播。从这个意义上讲，谁抓住了搜索引擎，谁就抓住了话语权，抓住了互联网上信息传递的主动权，抓住了保护国家信息主权的利器。

大搜索将互联网搜索推广到移动互联网、物联网等领域，这为保障国家安全的情报获取提供了直接手段。例如，网络安全研究者曾经利用 Shodan（网络设备搜索引擎）在网络中找到核电厂的指挥控制系统及一个离子回旋加速器，可以想象，如果不研发自主的搜索引擎尽早发现国家重大信息设施中的安全问题，势必对国家安全造成重大安全威胁。因此，发展我国自主知识产权的大搜索系统，可为国家安全提供信息情报支持。

总体上，虽然我国信息技术与欧美等信息产业发展与技术先进国家存在较大的差距，这种状况在今后相当长的时期内还不能彻底改变。但从国际宏观上看，目前大搜索技术仍然处于起步阶段：大多数欧美发达国家仍在探索网络空间大搜索的理论、方法和技术，没有建立统一的标准和规范。这也意味着：目前我国的技术和欧美发达国家在大搜索上几乎处于一致水平，至少我国与欧美国家在技术上不存在数量级差异。

我国已经失去了掌握互联网搜索引擎核心关键技术的契机，应当把握切入大搜索的机会机遇，努力与发达国家展开技术竞争，抢占网络空间大搜索引擎这一产业的制高点，把握这一大机遇，力争掌握相关自主知识产权，以争取在下一轮的信息革命中占据先机，从而提高社会运转效率，推动国家经济的健康发展。

11.4.2. 提高人民生活质量方面

大搜索将与各种现实生活具有深度关联，将在环保、医疗、教育、交通等各种领域都有深入的应用，在可以预测的将来，大搜索将重新定义我们的生活，服务于更多的大众，全方位提高人民的生活质量。例如：

在教育方面，通过标签和校园智能卡系统的结合，大学思想政治工作者可利用物联网系统对学生学习情况、到课情况进行分析从而有利于学生工作部有针对性地开展思想政治教育工作。同时还可以对学生在校园的行踪进行监控，设立校园安全控制区域，减少不必要的校园安全事故的发生增强学校与学生及家长的联系和沟通便于学校的管理。

在娱乐方面，依据用户的个性化行为习惯，选择最佳的娱乐方法。比如，在冬季，搜索“娱乐”的时候，可依据行为习惯为用户推荐滑雪，并依据雪场的厚度、雪场的温度、湿度、到雪场的交通路线、雪场的人数等推荐最佳滑雪场所等。

提高政府、企业、机构及个人的决策能力：大搜索将为决策者（包括政府决策者）制定有更好的依据，大大提高的决策能力，提升决策透明度。决策者决策需要准确的数据信息，目前一些决策者获得数据往往是不一致、不准确的、甚至有错误的，这为决策者正确决策带来了诸多问题。大搜索利用大数据分析工具对决策数据进行处理，具有强大的污点数据去除能力，决策者可通过大搜索获得准确客观的数据，从而提高正确决策能力。

和任何新技术的出现一样，大搜索将改变我们的生活，为人们提供更为方便、快捷、智能的服务，将在教育、文艺、道德、宗教、价值观念、风俗习惯等文化方面产生积极影响。

11.4.3. 拉动巨大商业价值方面

从全球市场来看，2021 年 2 月 Google 占全球搜索引擎市场的 92.03%，占比最大；Bing 占全球搜索引擎市场的 2.72%；Yahoo!占全球搜索引擎市场的 1.60%；Baidu 占全球搜索引擎市场的 1.18%；YANDEX 占全球搜索引擎市场的 0.64%。从中国市场来看，2021 年 2 月百度占中国搜索引擎市场的 71.10%，占比最大；搜狗占中国搜索引擎市场的 17.47%；好搜占中国搜索引擎市场的 3.40%；google 占中国搜索引擎市场的 2.80%；神马占中国搜索引擎市场的 2.70%。

近年来中国搜索引擎用户规模逐年攀升，截止 2021 年 6 月底中国搜索引擎用户规模达 7.95 亿人，较 2020 年 12 月底增加了 0.26 亿人。智研咨询发布的《2021-2027 年中国搜索引擎行业市场行情监测及市场分析预测报告》数据显示：从 2018-2021 年上半年中国搜索引擎用户规模对比数据可以看出，2021 年上半年中国搜索引擎用户规模平稳增长，2021 年上半年中国搜索引擎用户规模达 79544 万人，较 2020 年同期增加了 2990 万人，同比增长 3.9%。随着智能手机广泛普及以及移动互联网的不断发展，中国手机搜索引擎用户规模快速增长，截止 2020 年 12 月底中国手机搜索引擎用户规模达 76836 万人，较 2020 年 3 月底增加了 2301 万人。自 2016 年起中国手机搜索引擎用户规模占据搜索引擎用户规模九成以上的比例，截止 2020 年 12 月中国手机搜索引擎用户规模占搜索引擎用户规模的 99.82%，较 2020 年 3 月底的 99.36%增长了 0.46%。

搜索引擎用户活跃水平保持增长，一是得益于搜索引擎内容建设和小程序服务的深入发展，用户使用日趋活跃。数据显示，2021 年 3 月，百度 APP 月活跃

用户数达到 5.58 亿，较 2020 年 12 月底增长 2.6%，微信搜一搜月活跃用户数自上线以来一直保持快速增长态势，截至 2021 年 1 月已超过 5 亿。二是随着经济形势好转，围绕搜索产生的收入规模出现回暖趋势。2021 年第一季度,百度网络营销收入同比增长 27%。此外，头条搜索、微信搜一搜等以持续强化连接能力、完善搜索生态建设为发展重点，为商业化提供增长动力，如搜一搜加速连接小程序，推动内容、服务、品牌接入微信小程序，助力交易额快速增长。

可以预测，在未来融合各种数据的大搜索将对全球经济产生直接深远的影响，体现在三个方面：（1）大搜索自身所带来的广告价值，由于大搜索在环保、医疗、教育、交通等方面具有巨大的应用，将会带来庞大的广告市场。目前大搜索将传统互联网、移动互联网和物联网等有机整合，将会吸引更庞大的用户群，带来更为广阔的广告市场。（2）大搜索对企业将产生直接收益，企业利用大搜索进行精准广告展示，这将大大提高企业的销售量、降低广告成本。（3）搜索服务给搜索用户带来的价值，虽然这一部分很难计算，但可以粗略估计其效益是巨大的。假设互联网搜索服务使用用户平均每天节约 3.75 分钟，按每小时 18 元人民币计算，每天节约价值 1.125 亿元人民币。如果按中国有 7 亿劳工每天使用一次计算，使用搜索服务每年节约的时间价值 2874 亿人民币。

11.4.4. 推动 IT 技术的发展方面

物联网、移动互联网、大数据、云计算四大 IT 技术的发展为智慧搜索创造了良好的生态环境，催生了其诞生。反过来智慧搜索作为连接各种应用的桥梁，将上述各项技术紧密地结合在一起，使其相辅相成，促进四大 IT 技术的发展。

随着物联网、社交网络、电子商务、信息系统大规模使用，大量信息设备互连互通，感知识别无处不在，海量信息生成传输，信息量开始了爆发式的增长，世界开始迈入大数据时代。大数据时代的来临使人类第一次有机会和条件，在非常多的领域和非常深入的层次获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律，获得过去不可能获取的知识。物联网数据具有时效性、空间性强、数据量大和动态性高的特点，运用云计算模式可以使物联网中超大规模物品的实时管理与智能分析变得可能。而大搜索的出现，使得通过数据分析获得知识、商机和社会服务的能力，从以往局限于少数象牙塔之中的学术精英圈子扩大到了普通的机构、企业和政府部门。大搜索将物联网感知客观物理世界的浩瀚信息整理分类，可以帮助人们更高效地从中找到所需要的内容和信息，使物联网资源得以高效利用，促进物联网技术的发展。

正如搜索引擎的出现极大地推动了网络技术及 Web 应用的发展、网络技术

和 Web 应用的繁荣也为搜索引擎提供了巨大动力一样，智慧搜索的发展也必将与移动互联网、物联网、云计算、大数据等相互促进，推动信息产业的繁荣。

11.5. 总结及展望

网络空间大搜索是新一代搜索引擎，旨在面向泛在网络空间中的人、物、知识，综合利用大数据分析、自然语言处理和人工智能等技术，针对用户搜索需求返回全面准确的知识解答。发展网络空间大搜索技术，对于我国在信息技术领域抢占 IT 技术高地、保障国家战略需求、促进社会经济发展具有广阔的前程和非凡的意义。

本报告首先分析了泛在网环境下被搜索的人物、信息和物体对象动态演绎的特性，导致了用户“智慧”搜索的需求，带来的多模态、多层次、多粒度知识提取，用户的搜索意图准确理解与表示，海量、分布、异构、演化的知识管理及搜索任务融合推理，搜索结果可信度准确评价及用户隐私保护等方面面临的关键科学问题。

其次，对应上述关键科学问题，从文本知识获取、图片知识获取、视音频知识获取角度总结了泛在网络空间信息获取与发掘的技术进展及趋势。从知识表示管理、知识融合协同推理、基于 MDATA 多维关联模型的知识表示角度总结了超大规模知识图谱构建与管理的技术进展及趋势。从单来源用户、协同式用户真实意图理解角度总结了用户搜索意图准确理解与表示和技术进展及趋势。从关键词倒排索引与匹配、大图计算平台的支撑角度总结了用户意图高效准确匹配与推演的技术进展及趋势。从网络空间大搜索安全可信、隐私保护角度总结了大搜索安全可信与隐私保护的技术进展及趋势。

最后，针对网络空间大搜索对我国社会、经济和人民生活产生的影响进行了全面分析。在国家安全方面，网络空间大搜索为保障国家安全的情报获取提供了直接手段，是保护国家信息主权的利器。在人民生活方面，网络空间大搜索在教育、娱乐、机构决策等方面提供更为方便、快捷、智能的服务，对提高人民生活质量产生积极影响。在商业价值方面，网络空间大搜索将从自身广告价值、基于精准广告的企业直接收益、巨大的用户价值等方面对我国乃至全球经济产生直接深远的影响。在技术发展方面，网络空间大搜索可充分发挥桥梁的作用将物联网、移动互联网、大数据、云计算四大 IT 技术紧密地结合在一起，使其相辅相成发展。

综上所述，从国际宏观上看，目前网络空间大搜索技术仍然处于起步阶段，然而相关技术发展迅速，各个国家都将大力推动网络空间大数据的战略性产业，

以提高网络空间大搜索服务社会和经济决策、保障国家安全、提高人民生活、促进相关领域技术发展的能力。我们应当把握切入网络空间大搜索的机会机遇，抢占相关技术和产业制高点，力争掌握相关自主知识产权，以争取在下一轮的信息革命中占据先机，为社会高效运转、国家经济健康发展保驾护航。

第十二章　隐私计算研究进展、现状及趋势

一个新的理论从创立到得到社会各界认可，往往需要较长的时间，克服各种困难，不断迭代演进，逐步发展完善，隐私计算还需要做大量的理论和技术探索研究。根据大数据安全和隐私计算技术的发展，中国中文信息学会大数据安全和隐私计算专业委员会 2018 年因势而成立，隐私计算是本专委会致力于推动的重要学术工作。经过几年来隐私计算的研究与发展，隐私计算得到学术界和产业界的认同，因此本专委会从 2021 年开始撰写隐私计算研究进展报告。

12.1. 研究背景与意义

随着通信技术、网络技术和计算技术的持续演进和广泛应用，形成了包含因特网、移动互联网、物联网、卫星通信网、卫星互联网、天地一体化网络等异构网络的泛在互联环境。泛在互联环境具有开放性、异构性、移动性、动态性等特性，并与边缘计算、云计算等技术深度融合。在性能越来越强的智能终端支持下，泛在互联环境能够提供不同层次的多样化和个性化的信息服务，实现了“万物互联、智慧互通”，极大地推动人类社会发展，对社会、政治、经济、文化等领域有重要战略意义。

在泛在互联环境下，信息广泛传播，呈爆炸式增长，电商、物流、支付、导航、社交等信息服务新业态不断涌现，大型互联网公司在服务用户的过程中通过采集、存留、交换、衍生等手段积累了海量数据，数据频繁跨境、跨系统、跨生态圈交互在信息服务的推动下成为常态，如图 1 所示。这些加大了隐私信息在不同信息系统中有意或无意留存的可能性，隐私信息保护短板效应、隐私侵权追踪溯源难等问题也随之而来，个人信息保护面临的问题与日俱增。

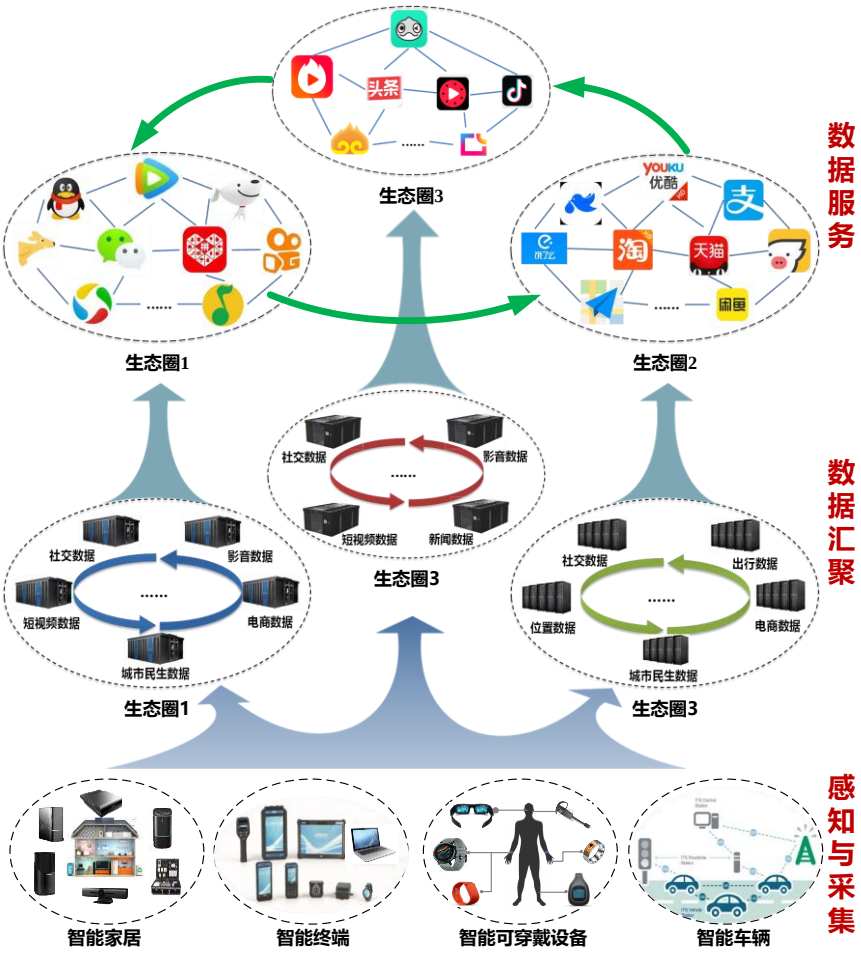


图 1 泛在互联环境下数据跨生态圈泛在共享

针对上述问题，各国政府部门展现出高度重视的姿态。例如，欧盟颁布的《通用数据保护条例》（General Data Protection Regulation，GDPR）强化了对被遗忘权、删除权的要求；我国颁布的《中华人民共和国民法典》将隐私保护纳入法律规定；中央网信办、工业和信息化部、公安部、市场监管总局四部门联合发布《关于开展 App 违法违规收集使用个人信息专项治理的公告》规范个人信息采集；2021 年 11 月 1 日，《中华人民共和国个人信息保护法》生效实施，明确了个人具备对个人信息处理的知情权、删除权等，个人信息的权益保障已成为国家战略。

个人信息保护面临的诸多具体问题包括：缺乏体系化标准规范与指引，APP 过度采集个人信息，后台隐私数据越权使用与个人画像，个人信息过度留存，生态圈之间信息共享缺乏延伸控制和迭代按需脱敏，多副本留存和保护短板效应凸显，删除权无法保障，数据交易和流动缺少有效监管手段，数据利用、脱敏、删除的合规评测缺少技术支撑等等。为解决这些问题，学者们针对某一环节的不同应用场景提出了诸多解决方案，这些方案虽能在特定应用场景、特定假设条件下解决特定的隐私信息泄露问题，但在面对“万物互联”场景下尚未提供体系化的保护能力。

个人信息保护的核心是隐私保护，隐私保护的根本问题是需要体系化的理论和关键技术以实现全生命周期的隐私信息管控，隐私信息管控的核心技术是个人敏感信息的分类分级和延伸控制，并在此基础上实现个人信息使用的知情权、脱敏、删除权/被遗忘权、流转管控和监管五位一体，迫切需要体系化、完善的隐私计算理论。

12.2. 隐私计算内涵与研究范畴

12.2.1. 相关领域的学术内涵

与隐私计算相关领域的概念内涵，目前学者有不同的理解，为了促进隐私计算的健康发展，本报告首先对相关概念内涵进行简要说明。

（1）个人信息与隐私信息：个人信息是指以电子或者其他方式记录的能够单独或者与其他信息结合识别特定自然人的各种信息，包括自然人的姓名、出生日期、身份证件号码、生物识别信息、住址、电话号码、电子邮箱、健康信息、行踪信息等。**隐私信息**是指个人信息中的敏感信息，是不想被非授权人知道的信息，是个人信息记录中的标识符、准标识符和敏感属性的集合。隐私反映了标识符、准标识符和敏感属性的关联关系。

（2）隐私泄露与隐私保护：隐私泄露分为两种情况，一是在有边界信息系统内隐私信息被非授权访问造成的泄露，二是在信息交换过程中未脱敏或脱敏强度没有达到要求而造成的泄露；对应的**隐私保护也分为两种情况**，一是保障信息不受损失前提下隐私不被非授权者获取及处理，我们称之为**隐私防护**；二是在隐私交换与处理过程中信息接收者得到隐私的信息量要小于信息发送方的同一隐私的信息量，使接收方不能完全获知发送方的真实信息，我们称为**隐私脱敏**。例如，去标识化使敏感信息与信息主体失去关联，也是信息量损失的形式之一。

（3）数据安全：主要指保证数据的机密性、完整性、不可否认性等，保证被保护的数据具有可恢复性，即信息的无损性。大多使用密码学、访问控制等方面的技术实施。

（4）密码学：主要研究范畴是保护信息的机密性、完整性和不可否认性的理论及应用技术。机密性的本质是信息没有损失，在共享范围内所有人得到的内容是相同的，主要用于防止在知悉范围之外的人获得被保护的信息；机密性的研究范畴是面向数据安全、传输安全等场景，并不特定针对隐私保护，在特定场景下可用于隐私防护。完整性是防止信息被篡改，其研究内涵与隐私保护没有任何关系。不可否认性可用于确定数据来源、交易等场景的真实性，还可用于隐私全

生命保护过程中的审计取证。

（5）访问控制：主要用于控制信息知悉范围，即确认主体访问客体的权限，不涉及信息内容的变更，但可决定主体访问信息的全部或部分。传统上用于数据保护，在泛在互联环境下也可作为一种知悉范围的控制机制，可在同一授权体系内用于隐私防护，但当信息离开该授权体系时不能提供延伸的访问控制。

（6）可信计算：通过可信基、可信执行环境、信任传递机制等构建可信系统，核心是保障计算环境的可信性和数据在计算过程中不被篡改。可信计算的本质是在可信系统范围内提供数据安全，当数据离开可信环境将无法保证数据安全。从隐私防护的角度，可信计算仅为隐私数据处理提供一个可信赖的计算环境。

（7）机密计算：在受信任的硬件执行环境基础上构建安全区域，所有参与方将需要参与运算的明文数据加密传输至该安全区域内并完成运算，安全区域外部的任何非授权的用户和代码都无法获取或者篡改安全区域内的任何数据。机密计算过程中的元数据不被计算参与方所获取，主要用于云计算场景下计算结果以明文或者机密性保护的方式交换。机密计算可在可信硬件执行环境下实现隐私防护，但当数据离开可信硬件执行环境时无能为力，仅适用于云计算等特定场景下的隐私防护。

（8）密文计算：是指计算过程中的数据不被计算参与方所获取，主要用于外包计算场景。同态加密是密文计算的代表性技术，是在事先确定转换规则的前提下，所有参与运算的明文数据使用该规则转换为密文，在密文空间中进行特定形式的代数运算并得到结果，密文运算的结果再通过相应的转换规则转换为明文运算结果，该结果与明文运算结果一致。本质上密文计算参与运算的明文及明文结果都没有信息损失，因此密文计算仅可用于计算过程中的隐私防护。

（9）安全多方计算：在事先确定参与方数目范围及交互协议的前提下，所有参与方以密文形式交互参与运算的信息并完成预先约定的运算任务，所有参与方都能得到运算结果的明文，但不能得到相互交互参与运算的明文信息。安全多方计算是无中心的计算架构，在有恶意参与者的情况下，诚实参与者仍能得到正确的结果，并且不泄露敏感信息。现阶段参与方的数目一般是两方和三方比较常见。秘密共享和不经意传输协议是构造安全多方计算协议的重要机制。本质上安全多方计算没有信息损失，适合于参与方较少场景下的隐私防护，但不适合于参与方高动态变化场景下的隐私防护。

（10）可算不可识：在 AI 和大数据应用中通常需要使用大量数据，但并不关心某个人的具体信息，可算不可识的目标是去标识化，原始数据不受损失，也不对敏感属性进行脱敏，因此可算不可识是隐私计算的一种应用需求，但并不能代替隐私计算。

（11）可用不可见：指泛在互联环境下用户可以得到数据计算的结果，但不能获取原始数据。可使用机密计算、密文计算、安全多方计算、“数据不动程序动”等技术或机制实现，属于数据安全的应用需求，而原始数据不出域是访问控制的研究范畴，可用于隐私防护。

（12）联邦学习：是多方利用自身拥有数据完成机器学习模型训练的一种分布式架构，合作方之间交换训练中间结果和模型参数，而不交换数据本身，自然而然减少了数据泄露，联邦学习的中间结果也会泄露数据的部分信息。因此，联邦学习是 AI 训练模型的一种模式，对隐私保护而言它仅是一种应用场景。

（13）隐私增强计算（Privacy Enhancing Computation）：Gartner 发布的 2021 年前沿科技战略趋势[1]中提到了隐私增强计算，但我们认为其命名并不妥当，隐私保护的根本目的是不能让隐私本身增强，但“隐私增强计算（Privacy Enhancing Computation）”的中英文词义顾名思义应理解为隐私的增强计算技术，相应地应属于挖掘隐私信息的技术领域，即使隐私特征信息更加凸显出来。我们认为，若要表达用于隐私保护的技术，建议称为“隐私降低计算（Privacy Reducing Computation）”或“隐私保护能力增强（Capability Enhancing for Privacy Preservation）”的计算技术才更为恰当。

综上，我们梳理隐私、个人信息、数据、数据安全、隐私防护和隐私脱敏等概念之间的关系，如图 2 所示。其中，安全多方计算、同态加密、可信计算、密文计算、访问控制等技术是属于数据安全范畴，也可用于隐私防护，仅适用于特定知悉范围内没有信息损失的敏感信息保护。隐私脱敏是面向泛在互联环境下隐私信息泛在共享的隐私保护需求，是按照隐私保护的需求对隐私信息进行适当的损失以保护个人权益。隐私计算是针对泛在互联环境下隐私信息共享的全生命周期隐私保护和管控的理论和方法。

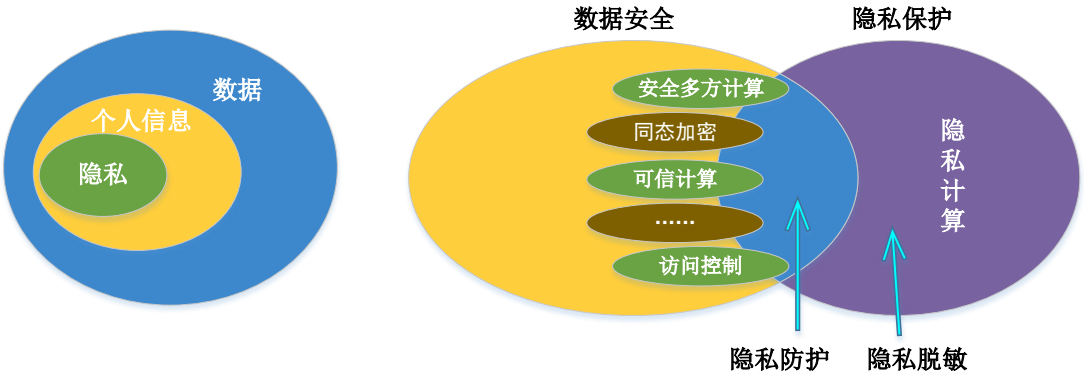


图 2 相关概念之间的关系

12.2.2. 隐私计算内涵

12.2.2.1.隐私计算的定义

隐私计算的核心思想是支撑隐私信息的感知量化，建立隐私信息操作过程中的可计算模型，刻画隐私操作组合时隐私分量的量化演变规则、隐私保护算法能力评估、保护效果量化、隐私传播控制及其相互之间的映射关系，确定不同约束下能达到的最优隐私保护效果以及实现最优效果的隐私保护算法及其组合。隐私计算的最终目标是隐私保护的自动化执行，构建支持海量用户、高并发、高效能隐私保护的系统设计理论与架构，实现不同算法之间的有效组合。

隐私计算的定义为[2]：隐私计算是面向隐私信息全生命周期保护的计算理论和方法，是隐私信息的所有权、管理权和使用权分离时隐私度量、隐私泄露代价、隐私保护与隐私分析复杂性的可计算模型与公理化系统。具体是指在处理视频、音频、图像、图形、文字、数值、泛在网络行为信息流等信息时，对所涉及的隐私信息进行描述、度量、评价和融合等操作，形成一套符号化、公式化且具有量化评价标准的隐私计算理论、算法及应用技术，支持多系统融合的隐私信息保护。隐私计算涵盖了信息搜集者、发布者和使用者在信息产生、感知、发布、传播、存储、处理、使用、销毁等全生命周期过程的所有计算操作，并包含支持海量用户、高并发、高效能隐私保护的系统设计理论与架构。隐私计算是泛在互联环境下隐私信息保护的重要理论基础。

12.2.2.2.隐私信息的形式化描述

信息 M 可以是文本、图像、语音、视频等一种模态数据或者几种模态数据的混合数据。信息 $\boldsymbol{M}$ 中包含的隐私信息 $\boldsymbol{X}$ 用六元组 $\langle \boldsymbol{I}, \boldsymbol{A}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Theta}, \boldsymbol{\Psi} \rangle$ 表示，其中 $\boldsymbol{I}$ 代表隐私信息向量，其分量表示信息 M 中语义上含有信息量的、不可分割的、彼此互不相交的原子隐私信息；$\boldsymbol{A}$ 代表隐私属性向量，其分量表示隐私属性分量，用于量化隐私信息分量及分量组合的敏感度。在实际应用时，不同场景下的不同隐私信息分量可进行加权动态组合，这些组合会产生新的隐私信息，将不同隐私信息分量组合的隐私信息敏感度也作为扩展的隐私属性分量，因此隐私属性分量的数目多于隐私信息分量的数目；$\boldsymbol{\Gamma}$ 代表广义定位信息集合，表示隐私信息分量在信息 M 中的位置信息及属性信息，可对隐私信息分量快速定位；$\boldsymbol{\Omega}$ 代表审计控制信息集合，表示隐私信息分量在传播过程中一个具体的审计控制向量，用于记录隐私信息分量在流转过程中的主客体信息和被执行的操作记录，当发生隐私

信息泄露时，可进行追踪溯源。$\boldsymbol{\Theta}$ 代表约束条件集合，表示隐私信息分量对应的约束条件向量，用于描述在不同场景下实体访问对应隐私信息分量所需的访问权限；$\boldsymbol{\Psi}$ 代表传播控制操作集合，用于描述隐私信息分量及其组合可被执行的操作，例如复制、粘贴、转发、剪切、修改、删除等操作，这些操作不破坏 $\boldsymbol{I}$ 的原子性。

12.2.3. 隐私计算研究范畴

12.2.3.1.隐私计算关键技术环节

为了能够自动地对不同场景、不同类型的隐私信息进行差异化保护，需要构建出清晰的、软硬件高效实现的隐私计算框架，包括隐私信息的感知、隐私化、存储、融合、交换和销毁等关键技术环节。隐私计算所涵盖的 6 个环节的关系如图 3 所示，可指导隐私信息保护系统的实现。

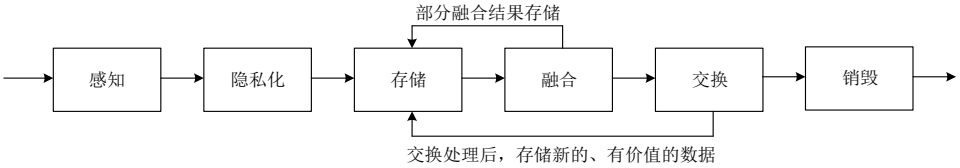


图 3 隐私计算关键技术环节

1. 感知

在感知环节主要关注隐私描述规约、隐私分量判定、分类与分级量化。在隐私描述与规约机制方面，需要解决隐私元数据提取、隐私标记和编码、隐私的描述、隐私信息变化过程、推理规则等；在隐私分量判定、分类分级量化方面，在给定一个或多个数据文档的情况下，判定是否存在隐私，以及隐私分量的量化度量。所设计的隐私计算模型需要具备对主体、时间、空间三维演化的刻画能力。

2. 隐私化

隐私化环节主要关注脱敏机制、算法保护能力的评价理论和方法等问题。在脱敏机制方面，研究如何构造适用于隐私保护、与传统数据加解密不同的脱敏操作，*k*-匿名、混淆、泛化、抑制、解耦、加扰、差分隐私等都可作为大规模隐私保护信息系统的局部组件；在算法评价理论和方法方面，需综合判定和评价所选用的隐私保护算法是否满足相应的保护需求、是否具备对抗关联分析能力等方面要素，并给出相应的评价标准理论和方法。

3. 隐私信息存储

存储环节主要关注同质隐私信息去冗、隐私感知的混合数据分割存储、单副

本的多用户完整性校验等问题，支持远程访问和细粒度访问的新型访问控制机制、局部数据修改和群修改的新型访问控制机制，以支撑隐私保护删除权、被遗忘权的落地实现。

4. 隐私信息融合

融合环节主要关注隐私信息匹配、隐私信息变换和隐私属性衍生、约束条件映射、隐私操作和隐私保护方案的自适应选择等问题。

5. 隐私信息交换

交换环节主要关注延伸访问控制机制、隐私动态调整、隐私侵权行为的判定和溯源取证等问题，通过延伸授权解决二次分发问题。

6. 隐私信息销毁

销毁环节主要关注删除指令通知机制、隐私信息的确定性完备删除等问题。确定性删除需保证隐私化后的信息不能去隐私化，且在接收到用户要求删除指令或者与用户约定信息存储到期后自动删除。建立通知消息机制和一套通知关联系统，通知其他隐私信息控制者和处理者删除隐私信息，释放存储空间。

12.2.3.2.隐私计算框架

隐私计算框架是在隐私信息全生命周期的各个环节中建立应用场景、保护需求与计算模型之间的映射关系。基于场景描述和保护需求，适应性地选择相应环节的计算方法实现相应的计算功能。从全生命周期的角度出发，隐私计算框架如图 4 所示。

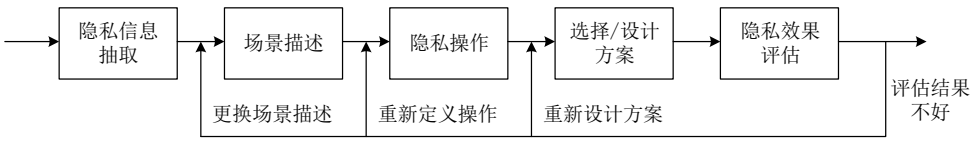


图 4 隐私计算框架

该框架面向任意格式的明文信息 M，具体包括以下 5 个步骤。

（1）隐私信息抽取：根据明文信息 M 的格式、语义等，抽取隐私信息并得到隐私信息向量 $\boldsymbol{I}$ 。

（2）场景抽象：根据 $\boldsymbol{I}$ 中各隐私信息分量的类型、语义等，对应用场景进行定义与抽象。

（3）隐私操作选取：选取各隐私信息分量所支持的隐私操作，并生成传播控制操作集合。

（4）隐私保护方案设计/选取：根据需求选择/设计合适的隐私保护方案。如

有可用且适合的方案及参数，则直接选择；如无，则重新设计。

（5）隐私保护效果评估：根据相关评价准则，使用基于熵或基于失真的隐私度量来评估所选择的隐私保护方案的隐私保护效果。

12.2.3.3.隐私信息系统框架

隐私信息系统框架包括语义提取、场景抽象、隐私信息变换、隐私信息融合、隐私操作选取、隐私保护方案设计/选取、隐私保护效果评估等环节，隐私信息系统框架如图 5 所示。

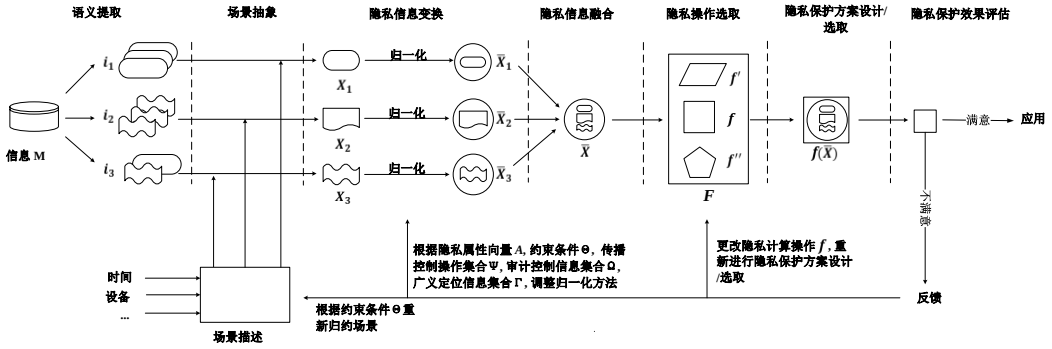


图 5 隐私信息系统框架

12.3. 隐私计算主要研究进展

12.3.1. 隐私计算理论的研究进展

12.3.1.1.隐私计算理论的提出

2015 年 12 月初，在北京首农香山国际会议中心讨论隐私保护相关技术时，中国科学院信息工程研究所李凤华研究员在国内外首次提出将隐私保护相关研究上升到理论体系，强调隐私保护是一种应用需求，而隐私计算才能代表一个理论体系，为了进一步明确隐私计算的内涵，李凤华给出了 2.2.1 节所述的隐私计算定义，并于 2016 年 4 月，联合李晖、贾焰、俞能海、翁健教授[2]在《通信学报》发表“隐私计算研究范畴及发展趋势”，正式发布了隐私计算的概念、学术内涵和研究范畴。同年，该论文被列入由中国密码学会组编的《中国密码学发展报告(2016-2017)》的 4 项年度成果之一。

12.3.1.2.隐私计算理论研究的深入

2019 年 3 月，李凤华、李晖等人[3]在中国工程院院刊《Engineering》上发表了“Privacy Computing: Concept, Computing Framework, and Future Development Trends”，从信息采集、存储、处理、发布（含交换）、销毁等全生命周期的各个环节角度出发，阐明了现有常见应用场景下隐私保护算法的局限性，提出了隐私计算理论及关键技术体系，其核心内容包括：隐私计算框架、隐私计算形式化定义、隐私计算应遵循的四个原则、算法设计准则、隐私保护效果评估、隐私计算语言等内容，并以四个应用场景为示例描述了隐私计算的普适性应用。

2021 年 4 月，李凤华、李晖、牛犇[4]撰写了隐私计算方面的首部学术专著《隐私计算理论与技术》，并由人民邮电出版社正式出版发行。该专著针对泛在互联环境下的体系化隐私保护需求，高度凝练并系统介绍了隐私计算研究范畴、理论及其关键技术，并深入浅出地阐述了为什么要研究隐私计算、什么是真正的隐私计算、如何研究隐私计算、隐私计算成果如何落地，以及隐私计算如何演化发展。

隐私计算得到了学术界的共识和认可，隐私计算研究被列入“十四五”国家重点研发计划“网络空间安全治理”重点专项 2021 年度项目申报指南的基础前沿技术类。

12.3.2. 隐私计算技术的研究进展

12.3.2.1.隐私感知与度量

(1) 隐私信息智能感知

隐私信息的智能感知是针对多模态数据形成隐私信息描述中的隐私信息分量，针对不同类型的数据需要使用相应的方法和工具。例如，针对文本数据，可以使用自然语言处理方法将文本分割为最小粒度；针对图像数据，可以采用图像理解算法识别图像数据中包含的语义。在此基础上，基于隐私智能感知算法，识别其中包含的隐私信息分量。

隐私信息的智能感知可以通过预先构建的隐私识别模板或者隐私知识图谱匹配来实现。要保证隐私信息感知的准确率，则需要重点研究隐私知识图谱。因此隐私信息感知更多的是借鉴自然语言处理、图像理解、知识图谱等方面的研究成果。

(2) 分类分级

在数据分级分类与隐私信息识别方面，2004 年 NIST 发布了《FIPS 199 联邦信息和信息系统的安全分类标准》，从信息的机密性、完整性和可用性三个角度

进行低、中、高三个等级的评定。2015 年，NIST 发布了 SP 1500-2《NIST 大数据互操作性框架：第二卷，大数据分类法》，提出了基于大数据参考架构（NBDRA）的角色样本分类体系，将每个元素分解成多个部分，提供了特定粒度数据对象的描述以及属性、特征和子特征。

在国内，与个人敏感信息相关的分类分级标准包括：GB/T 37964-2019 《信息安全技术个人信息去标识化指南》、GB/T 35273-2020《信息安全技术 个人信息安全规范》、JR/T 0171-2020《个人金融信息保护技术规范》、GB/T 38667-2020《信息技术 大数据 数据分类指南》颁布实施；《信息安全技术 个人信息安全影响评估指南》《信息安全技术 个人信息安全工程指南》展开编制；GB/T 37988-2019 《信息安全技术 数据安全能力成熟度模型》颁布实施，为数据安全能力的评估提供标准。

(3) 敏感信息识别

李凤华等人[5]针对社交网络照片分享场景提出了一种照片隐私感知的方案 SRIM（Social Relation Impression-Management）。照片中含有用户身份、位置、关系等隐私信息，分享照片可能会造成隐私泄露。SRIM 利用关系印象评估算法评估欲展示图片中的社交关系，并根据历史信息将图片接收者划分为推荐和不推荐展示两个类别,该方法不仅可以防止用户社交关系隐私信息的泄露，还可以自动推荐合适的图片分享策略。

基于分类分级标准，国内外已有部分厂商尝试利用自动化方法识别敏感数据。Amazon 公司发布了 Macie 通过机器学习和模式匹配识别 AWS 中的敏感数据。深信服智能数据分类分级平台引入了人工智能与机器学习算法，实现对数据进行多维度元数据特征向量自动提取，对相似字段数据进行聚合归类；华为云数据安全中心支持敏感数据快速识别。

12.3.2.2.隐私脱敏算法

隐私脱敏算法是隐私计算框架中按需脱敏的重要环节。当前隐私脱敏算法理论主要有针对标识符的匿名化技术和差分隐私技术。

12.3.2.2.1. 匿名化脱敏技术

在发布数据时如果不加保护的发布原始数据，会导致严重的隐私信息泄露问题。数据记录的属性一般分为三类：显式标识符属性、准标识符（Quasi-Identifier，QI）属性、敏感属性。显式标识符属性可唯一标识单一自然人的属性，如身份证号码、姓名等；准标识符属性联合起来能唯一标识一个自然人的多个属性，如邮

编、生日、性别等属性联合起来可能构成准标识符；敏感属性包含自然人隐私数据的属性，如健康状况、薪酬、兴趣爱好等。

匿名化脱敏的目标是设法阻止每条记录中的敏感属性与显式标识属性相链接，避免个体的敏感属性值的泄露，同时要保留敏感属性的值，以供数据的使用者对进行数据挖掘和统计分析。典型的匿名化脱敏方法包括：

（1）泛化：将某一属性值用更一般的属性值来替代。聚类是一种特殊的泛化，它将表中的 n 条记录划分至 m 个不同聚类，每个聚类中的点数不少于 k 个。

（2）数据扰动：通过加噪、数据置换、人工数据合成等方法对原始数据进行一定的修改，但保留原始数据的统计信息。加噪用于数值型隐私数据；数据置换是指交换记录的隐私属性值；人工数据合成即依据现有数据构建一个统计模型，然后从模型中抽样来构造合成数据以代替原始数据。

（3）抑制：用特殊符号代替现有属性以使得现有属性值更为模糊的匿名方法，如将手机号码写作 159****9468 以实现匿名。

（4）去耦：其不改变准标识符属性值和隐私属性值，而是将两者分开至两个独立的表中，这样，虽然数据不发生改变，但原有数据挖掘方法将不再适用。

（5）k-匿名[6]**：**由 Latanya Sweeney 和 Pierangela Samarati 在 1998 提出，它通过混淆数据的准标识符属性，可以在保证数据的实际可用性的条件下，保证其中的个体身份不会被恢复出来。因为 k-匿名中不包含任何的随机化属性，其容易遭受背景知识攻击和同质攻击（Homogeneity Attack）。同质攻击指如果一个匿名后等价类的所有个体的敏感属性都相同，如果攻击者知道某个用户在这个等价类中，就能推断出该用户的敏感属性。

（6）l-多样性：针对k-匿名的同质攻击，Machanavajjhala 等人[7]在 2007 提出了一个改进的方案l-多样性，使一个等价类中最少有l个不同的敏感属性值。但是，l-多样性也并不能完全的保护用户隐私不被泄露，因为其只保证了多样性，忽略了属性值上语义相近的情况。例如等价类中不同的敏感属性值为胃炎、胃溃疡、胃癌等，那么至少可以知道数据的主体患有胃病。另外，针对l-多样的偏义攻击也可能引起隐私泄露。比如，一个新冠肺炎疾病信息的数据集中某一等价类内包含阳性和阴性人数各占 50%，从而满足 2-多样性，但我们知道正常数据集整体抗体阳性和阴性比例分别占 1%和 99%。这样若知道某个个体在这个等价类中，其有 50%的概率阳性，事实上已经发生了隐私泄露。

（7）t-邻近性：Li Ninghui 等人[8]在提出的t-邻近方案弥补了l-多样性，t-

邻近指一个等价类中的属性分布和整个表中的属性分布之间的距离不超过门限 t。如果一个数据表中的每个等价类都满足t-邻近，则称这个数据表满足t-邻近。

12.3.2.2.2. 差分隐私脱敏技术

Dwork 等人[9]提出的差分隐私模型来自于密码学中语义安全的概念，即攻击者无法区分出不同明文的加密结果。差分隐私模型不需要依赖于攻击者所拥有多少背景知识，而且对隐私信息提供了更高级别的语义安全。李凤华等人[3]提出了基于差分的通用脱敏算法设计准则，包含以下步骤：

预处理：

在差分隐私保护算法中，记隐私信息为 X，根据 X、约束条件集合$\boldsymbol{\Theta}$和传播控制操作集合$\boldsymbol{\Psi}$，生成对应的隐私信息向量集合$\boldsymbol{I}=i(\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\Psi})$，分析$\boldsymbol{I}$的分布特征$\boldsymbol{\Phi}=\phi(\boldsymbol{I})$，确定$\boldsymbol{I}$的取值空间或者取值集合$\mathrm{Ran}$。根据定义在$\boldsymbol{I}$上的统计查询函数$g(\cdot)$，确定查询次数的期望值$t(\cdot)$和查询结果的社会经验值$v(\cdot)$，得到添加的噪声取值空间或取值集合$S=s(\Phi,\mathrm{Ran},g(\cdot),t(\cdot))$，并计算统计查询函数$g(\cdot)$的敏感度。对于一个定义在$\boldsymbol{I}$的子集$D$上的统计查询函数$g(\cdot)$，其敏感度定义为

$$\Delta g=\max\|g(D_1)-g(D_2)\|_p$$

其中，$D_1,D_2\subseteq\boldsymbol{I}$，$D_1$、$D_2$为任意两个相差最多一个元素的集合，称为相邻集合，$p\geqslant 1$且为整数。

算法框架：

基于预处理结果，充分考虑隐私保护复杂度C、隐私保护效果Q等要素，将差分隐私机制的数学定义表示为

$$\Pr[\mathrm{Alg}(D_1)\in S]\leqslant h(\cdot)\Pr[\mathrm{Alg}(D_2)\in S]+\delta(\cdot)$$

其中，$h(\cdot)=h(\lambda,\varepsilon,\kappa)$表示扩展的隐私预算，其中$\lambda$为常数，与噪声分布相关，$\varepsilon$与查询次数期望值相关，$\kappa$与查询结果社会经验值相关；$\delta(\cdot)=\delta(\varepsilon,\kappa)$为修正参数，用来放宽条件使算法满足差分隐私定义；$D_1$、$D_2$是一对相邻集合；$\mathrm{Alg}$为一

个随机化算法。

差分隐私保护算法框架为

While $\mathrm{Alg}(g) \notin v(\cdot)$

Do $\mathrm{Alg}(g) = g(D) + \mathrm{Noise}(\mu(\cdot), b(\cdot), q(\cdot))$

其中，$\mathrm{Noise}(\cdot)$ 为噪声函数集，产生的噪声满足 $(h(\cdot), \delta(\cdot)) - \mathrm{DP}$ 条件；$\mu(\cdot)$ 为产生噪声的期望；$b(\cdot)$ 为尺度参数函数，控制噪声分布的范围；$q(\cdot)$ 为指数机制中的效用函数，控制数据经过加噪后输出某种结果的概率预期。根据应用场景和信息类别，选择具体的噪声分布和算法参数。可以选择满足拉普拉斯分布$Lap(0, \frac{\Delta f}{\epsilon})$的噪声来实现差分隐私保护，称为拉普拉斯机制。如果噪声选择高斯分布$N(0, \sigma^2)$，则称为高斯机制。针对非数值型数据，可以采用指数机制[10]和网络机制[11]。

算法参数设计：

根据用户对隐私保护强度和可用性的应用需求，并结合隐私信息向量 $\boldsymbol{I}$ 的取值范围 Ran、查询次数的期望值 $t(\cdot)$ 等要素，确定噪声分布的具体参数取值。其中，μ 与输出结果的均值需求有关；$b(\cdot)$ 与 $h(\cdot)$、数据集敏感度 Δg、噪声取值空间或取值集合 S 等有关，即 $b(\cdot) = b(h(\cdot), \Delta g, S)$；$q(\cdot)$ 与 S、查询结果的社会经验值有关，即 $q(\cdot) = q(S, v(\cdot))$。

算法组合：

差分隐私机制具有如下组合特性。

（1）后处理性质（Post-Processing Property）：如果 $\mathrm{Alg}_1(\cdot)$ 满足 ε-DP，则对于任意的算法（可能是随机的）$\mathrm{Alg}_2(\cdot)$，组合后的算法 $\mathrm{Alg}_2(\mathrm{Alg}_1(\cdot))$ 也满足 ε-DP。

（2）顺序组合性质（Sequential Composition）：如果 $\mathrm{Alg}_1(\cdot)$ 满足 ε_1-DP，并且对于任意的 s，$\mathrm{Alg}_2(s, \cdot)$ 满足 ε_2-DP，则 $\mathrm{Alg}(D) = \mathrm{Alg}_2(\mathrm{Alg}_1(D), D)$ 满足 $(\varepsilon_1 + \varepsilon_2)$-DP。

（3）平行组合性质（Parallel Composition）：如果 $\mathrm{Alg}_1(\cdot), \mathrm{Alg}_2(\cdot), \cdots, \mathrm{Alg}_k(\cdot)$ 是 k 个满足 ε_1-DP，ε_2-DP，…，ε_k-DP 的算法，$D_1, D_2, \cdots, D_k$ 是 k 个不相交的数据集，则 $\mathrm{Alg}_1(D_1), \mathrm{Alg}_2(D_2), \cdots, \mathrm{Alg}_k(D_k)$ 满足 $\max(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k)$-DP。

当使用差分隐私保护算法对不同数据集的多种查询统计进行保护时，可以利

用上述 3 种性质对算法的不同步骤进行组合。

算法复杂度和效能分析：

差分隐私保护算法是将噪声与隐私信息相加，因此复杂度主要取决于噪声的生成，隐私保护效果也取决于噪声的大小。这些均与数据集特征、数据集敏感度计算等噪声生成的参数相关，可由算法 Alg 的复杂度 $C(\mathrm{Alg})=c(\Phi,\Delta g,h(\cdot),\delta(\cdot),\mu(\cdot),b(\cdot),q(\cdot))$ 和算法 Alg 的隐私保护效果 $Q(\mathrm{Alg})=\Delta\sigma(h(\cdot),\delta(\cdot),\mu(\cdot),\ b(\cdot),q(\cdot))$ 来刻画。

12.3.2.2.3. 本地化差分隐私机制

本地化差分隐私使得用户可以在上传数据前，先在本地扰动自己的数据，这样就可以保证不可信的服务器无法准确的获得用户的隐私数据。直观上来说，本地化差分隐私提供了一种保证，对于任意一对用户的输入，经过本地化差分隐私算法处理后可以达到一定程度的不可区分。2003 年 Evfimievski 等人[12]给出了本地化差分的概念，2008 年 Kasiviswanathan 等人[13]给出了严格的定义如下：

对于一个随机算法M，如果对于任意的一对用户输入x和x'，算法M满足：

$$\forall t \in Range(M): \Pr[M(x)=t] \le e^{\epsilon} \Pr\left[M\left(x'\right)=t\right],$$

其中$Range(M)$表示算法M可能的输出集合，则称算法M满足ϵ-本地化差分隐私，其中参数ϵ为隐私保护预算。

本地化差分隐私常用于进行特定的统计分析任务。Random Response (RR) 机制是本地化差分隐私典型方法，下面是其在频率估计上的应用示例。该机制的扰动方法如下。

$$\Pr[Perturb(x)=i]=\begin{cases} p=\dfrac{e^{\epsilon}}{e^{\epsilon}+d-1}, & if\ x=i \\ q=\dfrac{1}{e^{\epsilon}+d-1}, & if\ x\neq i \end{cases},$$

其中d是用户所有可能的输入的个数。

经过扰动之后，用户将扰动后的结果上传给不可信的服务器，服务器通过计算统计量

$$c(i)=\frac{\sum_j \mathbb{I}_{Support(y^j)}(i)-nq}{p-q}$$

来得到对第i个项目的频率。其中，y^j表示第j个用户的上传数据，$\mathbb{I}$为指示函数。$Support(y^j)$表示用户用 RR 机制扰动并上传的数据中，可以对计数第i个项

目的频率有贡献的数据。在 RR 机制中，$Support(i) = \{i\}$。统计量$c(i)$为第i个项目频率的无偏估计，同时，该方案的方差为$n\frac{d-2+e^{\epsilon}}{(e^{\epsilon}-1)^2}$，其中，$n$为用户个数。可以看出，该方差随着输入空间大小$d$的增大而增大，因此当输入空间很大时，Random Response 机制的可用性会有明显下降。

12.3.2.3.隐私保护效果评估

隐私保护效果评估是支撑信息发布、统计查询和数据交换的决策依据，也是自动化选择隐私保护算法的基础。在大型隐私保护系统中，算法的保护效果评估可以支撑根据系统要求自适应动态替换算法，同时保持系统框架的相对稳定。

隐私计算所需要的隐私保护效果评估是效果评估与算法保护能力量化、隐私信息感知量化间匹配或映射关系的联动研究。李凤华等人[3]提出了从可逆性、延伸控制性、偏差性、复杂性和信息损失性 5 个维度对隐私保护效果建立综合的评估体系。

（1）可逆性：指隐私保护算法执行前后隐私信息的被还原能力，具体是指攻击者/第三方从所观测到的隐私信息分量i'_k推断出隐私信息分量i_k的能力。若攻击者/第三方能准确推断出i_k，则具备可逆性，否则不具备可逆性。

（2）延伸控制性：指跨系统交换过程中接收方的隐私信息保护效果与发送方的保护要求的匹配程度，具体是指隐私信息$\boldsymbol{X}$从系统Sys_1转到系统Sys_2后，其在系统Sys_1中的隐私属性分量a_k与在系统Sys_2中的隐私属性分量a'_k的偏差。对任意k，在不同系统中，若$a_k = a'_k$，则说明延伸控制性良好，否则延伸控制性有偏差。例如，用户 Alice、Bob、Charles 互为朋友，Alice 在微信朋友圈中发布的一条隐私信息，设置了允许 Bob 看，不允许 Charles 看，但 Bob 将该信息转发至其新浪微博，且未设置访问权限限制，此时 Charles 就会看到。在该情况下，用户 Alice 对该条隐私信息在新浪微博中的访问控制权限与其在微信朋友圈中的访问控制权限就不匹配。

（3）偏差性：指隐私保护算法执行前后隐私信息分量i_k和隐私保护后发布攻击者或第三方可观测到的隐私信息分量i'_k之间的偏差。例如，位置隐私保护中，用户真实所处位置(m,n)与位置隐私保护算法（位置偏移算法）执行后的位置

(m',n')之间的物理距离为$\sqrt{(m-m')^2+(n-n')^2}$。

（4）复杂性：指执行隐私保护算法所需要的代价，即隐私保护复杂性代价。例如，对特定向量进行置换操作（如用*替代特定关键字）所需消耗的计算资源小于进行 k-匿名操作（k=30）所需的计算资源。

（5）信息损失性：指信息被扰乱、混淆等不可逆的隐私保护算法作用后，对信息拥有者来说缺失了一定的可用性。例如，在位置隐私当中，当用户不进行 k-匿名时，用户向服务器发送真实的地址，服务器会返回精确的推送信息；但当用户采取 k-匿名后，服务器会返回对用户来说粗粒度的推送信息，不可用的结果比例增加，造成了一定的信息可用性损失。

12.3.2.4.隐私延伸控制

作为隐私计算的重要内容之一，隐私延伸控制深度影响着当前和未来泛在互联环境下的隐私保护。李凤华等人[14, 15]针对单系统和跨系统图片隐私延伸控制的典型场景提出隐私延伸控制的方案

12.3.2.4.1. 单系统图片隐私延伸控制

人们日常分享的图片中经常会涉及一些朋友和路人的信息，他们可能并不希望自己被展示给未经授权的接收者。现有的图片隐私保护方案大多存在以下问题：一是图片分享中的访问控制方案大多要求图片参与者对每张图片设置策略，导致用户设置策略的时间成本极高；二是图片隐私策略推荐方案大多基于半自动的标签传播算法或图片分类算法，在训练样本过少或增加新的隐私类别时，准确率不高。李凤华等人[14]提出了一种针对图片的用户隐私保护策略生成方法 HideMe，可支撑单一系统中图片分享的延伸控制。用户可以利用丰富的内容要素构建客观场景，再通过一个基于图片场景信息的访问控制模型保护用户的隐私。

12.3.2.4.2. 跨系统交换的图片隐私延伸控制

为实现在社交网络中的朋友互动，用户的隐私图片在多信息系统、多边界之间广泛动态流转已成常态。然而，一旦用户将图片上传到社交网络平台，便失去了对上传图片的控制。传统的访问控制的方法大多关注单一系统，难以应用到跨社交网络的转发场景中；基于加密的图片隐私保护方法较少考虑访问控制策略，访问者能否完全依赖访问者是否拥有密钥；由于图片本身的复杂性和展示问题，传统的策略粘贴方法并不能直接运用到图片分享中。另一方面，追踪溯源方法大

多将隐私信息与溯源记录分开存储，当隐私信息离开信息系统后，无法对隐私侵权行为进行判断。

李凤华等人[15]提出了一种跨系统交互的隐私图片分享框架 PrivacyJPEG。从图片传播的角度出发，分别应用于延伸控制（正向）和追踪溯源（逆向）两个场景中。具体地，该方法将隐私标记和访问控制策略绑定到图片中，并利用加密算法保证图片的隐私区域在传播到其他社交网络时，仍只有拥有权限的用户才能访问。与此同时，通过在隐私标记中增加溯源记录信息，使得在隐私泄露事件发生后，取证人员可以对隐私侵权行为进行追踪溯源。

12.3.2.5. 隐私侵权的判定与取证溯源

在隐私交换过程中，虽然有延伸控制机制，但总存在攻击者试图想办法绕过或者篡改控制机制，或者不完整地按延伸控制要求进行控制操作。任何技术都无法提供绝对万无一失的保护，因此从整个技术发展的历史规律来看，隐私的保护与隐私的滥用是一对此消彼长的矛盾演化过程，所以一个成熟的隐私计算体系应该包含隐私侵权行为的判定与溯源。

在隐私信息系统中，实现隐私侵权行为判定是自动取证的基础，也是阻断隐私侵权行为扩散的重要关键技术，判定技术需要支持在线和离线实现。隐私侵权行为判定是在隐私信息的溯源记录中根据隐私侵权行为的判定标准，判断是否存在违反约束条件和控制策略的行为；溯源是在隐私信息交换过程中将交换的路径、交换过程中的相关操作以不可篡改的方式记录在隐私信息的审计控制记录当中，为判定、取证和追踪提供依据。判定需与追踪溯源联动研究，构建一个有机结合的整体机制，而不是两个割裂开来的不相关的技术。

在隐私计算的框架体系下，隐私侵权行为及取证存在于其各个步骤中。隐私侵权溯源取证框架如图 6 所示。

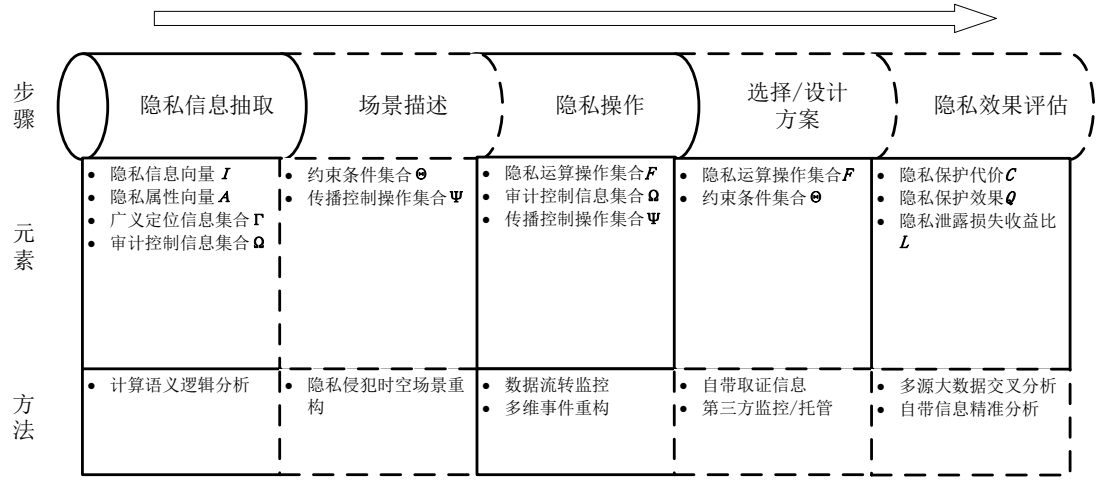


图 6　隐私侵权行为追踪溯源取证框架

（1）隐私信息抽取：当信息 $\boldsymbol{M}$ 产生时，通过语义逻辑的计算分析抽取或标注其隐私信息，得到隐私信息向量 $\boldsymbol{I}$、广义定位信息集合 $\boldsymbol{\Gamma}$ 和审计控制信息集合 $\boldsymbol{\Omega}$，并计算得到隐私属性向量 $\boldsymbol{A}$。此阶段主要用于界定隐私信息。

（2）场景描述：对信息所处场景进行抽象描述，得到约束条件集合 $\boldsymbol{\Theta}$、传播控制操作集合 $\boldsymbol{\Psi}$。该阶段提供了对隐私侵权行为的判定标准，当不满足上述条件时，则判定为隐私侵权行为发生。

（3）隐私操作：依据场景限制给各个隐私信息分量分配可进行的操作，形成隐私运算操作集合 $\boldsymbol{F}$，并在此基础上建立传播控制操作集 $\boldsymbol{\Psi}$；记录信息主体对该信息的隐私操作，生成或更新审计控制信息集合 $\boldsymbol{\Omega}$。超出上述两个集合的操作也会被判定为隐私侵权。

（4）选择/设计方案：在该过程中，分析所选择/设计方案中涉及的运算是否满足隐私运算操作集合，操作的动作、对象、结果等是否超出约束条件集合。防范隐私侵权行为发生，并作为隐私侵权判定标准。

（5）隐私效果评估：该环节包括分析计算隐私保护代价 C、隐私保护效果 Q 和隐私泄露损失收益比 L。当上述因素未达到预定目标时，则需要对隐私信息全生命周期保护进行反馈审核。

当发生隐私侵权时，需对前 4 个步骤中的信息流进行溯源分析，追踪隐私侵权发生的主体。基于隐私信息六元组以及第三方监控或托管，界定隐私信息，判定隐私侵权行为，并通过隐私计算框架中各个步骤的联动，对异常行为进行取证，并找到侵权行为的源头，实现溯源取证。

12.4. 隐私计算发展趋势与展望

12.4.1. 隐私计算的基础理论

从隐私感知与动态度量、隐私保护算法、隐私保护效果评估、隐私信息延伸控制、隐私侵权行为存证和溯源等环节进一步研究并完善隐私计算框架及其数学基础，细化各环节间的关联机制、操作控制及控制信息传递，可借鉴概率论与数理统计、信息论、博弈论、拓扑心理学等学科的思想，提出全流程隐私信息的流转控制模型，持续探索隐私计算的基础理论；研究业务服务与隐私计算深度融合

的高效隐私信息保护系统技术架构，提出典型应用场景的隐私信息保护解决方案，形成不同的隐私保护服务能力，推动隐私计算应用。

12.4.2. 隐私感知与动态度量

从隐私信息的知识表示模型、分类分级、原子抽象建模、特征分析与隐私分量抽取、压缩感知、隐私分量关联关系挖掘等角度入手，研究隐私分量与场景关联模型、隐私分量量化与动态调整、隐私分量组合与重度量等内容，解决时空差异和主体动态下隐私动态交换的精准度量问题，支撑隐私智能保护；提出场景对隐私保护要求的量化指标、隐私动态调整量化指标、隐私组合约束的量化指标，以及这些量化指标的关联关系和动态权值，形成隐私度量的量化指标体系，支撑泛在互联环境下隐私信息交换控制与按需脱敏。

12.4.3. 隐私保护算法

在不同环节研究基于不同数学基础的隐私脱敏原语，及其等价或映射关系，支撑隐私保护算法能力评估、泛在互联环境下隐私信息跨系统交换控制；设计隐私保护算法通用框架与设计准则、脱敏控制模型、算法选择和优化组合设计、算法前后台任务动态调度等内容，支撑隐私信息保护系统的柔性重构和隐私脱敏功能的动态编排；提出算法保护能力与保护效果评估、算法保护能力量化指标之间的等价关系等，形成算法保护能力量化指标体系，支撑隐私保护算法的设计与能力评估。

12.4.4. 隐私保护效果评估

从可逆性、延伸控制性、复杂性、偏差性、信息损失性等维度入手，研究保护算法及其组合的效果评估量化指标，以及量化指标的关联关系和动态权值等内容，形成效果评估指标体系，支撑隐私保护的效果反馈、隐私保护方案的迭代优化；提出效果评估系统的计算模型、自动评估系统的柔性架构等，支撑效果评估高效快捷、隐私保护算法优化选择；研究隐私关联性分析、算法可用性增强、隐私挖掘等内容，支撑隐私保护算法能力评估、隐私发布时脱敏效果评估、隐私信息保护系统能力评估。

12.4.5. 隐私侵权行为判定与溯源

以隐私侵权行为判决规则与约束表示为基础，研究延伸控制策略绑定、全流

程隐私侵权线索存证、侵权行为的场景与内容的存证、侵权事件识别与判定等内容，支撑泛在互联环境下隐私侵权行为精准判定；研究隐私信息流转的主被动协同监管架构、审计信息可信存证、操作控制约束与审计信息描述等内容，支撑隐私侵权的追踪溯源；研究授权控制链构建、传播策略与控制策略动态关联、权限动态调整、策略可验证执行与可信审计、延伸授权、协同溯源、侵权场景构建与行为重构等，支撑隐私信息受控共享。

12.4.6. 隐私信息的完备删除

从删除通知、通知确认、远程验证机制、传播路径发现、通知与确认拓扑生成、删除方案选择、删除操作行为可验证等方面，支撑多副本完备可验证删除；提出自动/指定删除机制、删除粒度协商机制、信息多副本检索、删除粒度控制、自主/自动删除触发、密钥自动删除、删除目标与密钥管理、最小域可信删除、最小覆写删除等，实现个人信息到期自动/按需删除；研究删除效果远程验证机制、存证推送机制、多副本全删除确认、删除不可恢复性评估、删除操作行为审计、违规留存取证、合规评测、删除流程与验证的可视化等内容，支撑删除可信验证。

12.5. 结束语

新的研究领域需要持续深入地开展研究，隐私计算也是如此。我们认为，学者应切切实实地区分数据安全和隐私保护研究范畴的异同，不应热衷于“旧酒换新瓶”。本报告仅列出隐私计算重点研究进展，并展望了重要研究方向，当然研究范畴还可以合理地扩展；隐私计算并不排斥传统数据安全的数学基础，也不排斥在某个局部环节采用数据安全的传统方法，比如加密、签名等。为了促进隐私计算理论与技术体系的不断发展和完善，更好地服务于泛在互联环境下的隐私保护，还需要围绕隐私计算的基础理论和各个环节开展更多的针对性深入研究。

但当前社会上存在借用“隐私计算”热度的现象，一些公开发布的学术观点混淆了隐私计算的概念和研究范畴。基于对个人信息保护的使命感、责任感，大数据安全与隐私计算专委会主动承担起从学术角度服务社会、促进学术研究的职责，我们深感有必要编纂并发布隐私计算研究进展报告，借此机会给出隐私计算与其他相关领域的学术内涵差异，希望能引导和促进隐私计算的理论研究与应用。

总之，作为隐私计算领域的第一份报告，立意定位于促进隐私计算的研究与发展，而不是一本白皮书。因此，限于篇幅仅介绍隐私计算的主要工作。

12.6. 参考文献

[1] Gartner Top Strategic Technology Trends for 2021 [EB/OL] https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021

[2] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.

[3] LI F H, LI H, NIU B, et al. Privacy computing: concept, computing framework, and future development trends[J]. ELSEVIER Engineering, 2019, 5(6):1179-1192.

[4] 李凤华、李晖、牛犇,《隐私计算理论与技术》,人民邮电出版社，2021.4

[5] LI F H, SUN Z, NIU B, et al. SRIM scheme: an impression-management scheme for privacy-aware photo-sharing users[J]. ELSEVIER Engineering, 2018, 4(1): 85-93.

[6] SWEENEY L. k-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[7] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. L-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, ACM 2007: 1(1) 3.

[8] LI N H, LI T C, VENKATASUBRAMANIAN S. t-closeness: privacy beyond k-anonymity and l-diversity[C]//2007 IEEE 23rd International Conference on Data Engineering. IEEE 2007: 106-115.

[9] DWORK C. Differential privacy: a survey of results[C]//International Conference on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1-19.

[10] MCSHERRY F, TALWAR K. Mechanism Design via Differential Privacy[C]// IEEE Symposium on Foundations of Computer Science. IEEE 2007: 94-103.

[11] BLUM, AVRIM, K. Ligett, and A. Roth. "A Learning Theory Approach to Non-Interactive Database Privacy." Journal of the ACM, ACM 2011: 1-25.

[12] EVFIMIEVSKI A V., GEHRKE J, SRIKANT R. Limiting privacy breaches in privacy preserving data mining[C]. Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM 2003: 211–222.

[13] KASIVISWANATHAN S P, LEE H K, NISSIM K, et al. What Can We Learn Privately?. the 49th Annual IEEE Symposium on Foundations of Computer Science. IEEE 2008: 531–540.

[14] LI F H, SUN Z, LI A, et al. HideMe: privacy-preserving photo sharing on social networks[C]//IEEE International Conference on Computer Communications. IEEE 2019: 154-162.

[15] 李凤华, 孙哲, 牛犇, 等. 跨社交网络的隐私图片分享框架[J]. 通信学报, 2019, 40(7): 1-13.

第十三章　开源情报技术研究进展、现状及趋势

13.1. 研究背景与意义

情报的重要子课题——“开源情报”（Open Source Intelligence, OSINT）源于“来源公开、方法公开、手段公开、内容公开”等“公开”理念的融入，包括对公开领域合法可获数据、信息或其他资源进行规划、搜集、处理、分析、分发、以及形成产品等情报活动。随着大数据时代的到来，公开信息源的信息量极为庞大和复杂，大量公开数据的搜索和分析都需要基于网络爬虫、大数据、数据挖掘、云计算、区块链等技术手段，也使得开源情报技术的研究在开源情报的发展中愈发重要。

开源情报技术是交叉学科研究的聚焦方向，主要是为了搭建政、产、学、研、用交流合作平台，解决国家、领域关键问题，在政治、经济、国防、人民生活等各领域都发挥了重要的作用。在国防开源情报中，利用开源情报技术可以提供宏观态势性和微观精准性情报分析，及时掌握最新动态，从而为维护国防安全提供有效的战略支援。在军事开源情报中，雷达探测范围有限，制约了海上大中型目标的检测与识别，而借助开源情报技术，可以在网络公开信息搜集的基础上，建立目标身份信息库，同时建立网络爬虫工具，利用实体识别工具等，实现目标检测、识别与验证。而在公安侦查方面，公安机关侦查人员则可以通过对互联网信息资源的搜集、整理、分析形成开源情报，在案件侦办过程中，开源情报技术有助于拓宽侦查线索，丰富证据来源，提升情报搜集的效率，在互联网背景下，开源情报技术逐渐成为各类侦查手段的有力补充。在应对突发公共卫生事件时，开源情报技术扩大了信息的收集范围，可以夯实政府信息公开的基础，同样起到了不可替代的作用。开源情报技术分析的多样性更是推动了疫情防控工作，其全面性可以帮助修复危机中的社会信任，同时循环处理流程促进了情报融合，从而有助于维护疫情条件下的社会稳定。此外，在经济方面，开源情报技术以大数据为核心，可以更大范围、更深层次地实现在“发现、跟踪、追溯、研判、预警”等方面的功能作用，从而更好地助力经济发展并为企业提质增效。开源情报技术以大数据、云计算等新兴技术为手段，可以更加全面综合的挖掘信息的变化趋势和规律，提高舆情预警和跟踪监测的有效性和时效性，实现对国际局势、社会态势、经济趋势的精准把握。

在当前信息化环境下，开源情报技术可以影响到社会生活的方方面面以及各个领域，在应对我国国家安全、社会稳定、经济发展等领域的挑战时，可以发挥

十分重要的作用。在此背景下，中文信息学会开源情报技术专业委员会旨在充分运用多种综合技术手段，采集、传输、存储/处理、抽取、分析横跨多个领域的公开来源数据，在鉴别、筛选、综合研判的基础上生产出满足各领域切实应用需求的情报产品，为国家安全和社会和谐发展提供重要保障。

13.2. 领域发展现状与挑战问题

开源情报是指“从收集、利用的公开信息中产生，并及时传递给相应用户以满足特定任务的需求”的知识激活过程。开源情报技术虽然是在最近几年成为研究热点，但实际上，开源情报技术也是伴随着互联网的发展而发展。随着互联网的信息网络（Web）技术的演变，开源情报也经历 3 个阶段的发展。

第一代开源情报阶段对应 Web1.0 阶段。该阶段的开源情报主要是从传统网页文本中获取信息，并进行相应的分析和处理。但侧重于原始信息的收集，分析手段比较落后；第二代开源情报分析阶段对应 Web2.0 阶段。此阶段的主要特征是互联网中包含动态网页和用户生成内容，任何人在任何时候和任何地点都可以在互联网发表信息，此时互联网成为人类有史以来最大的信息集散地，网络用户成为客观世界的最佳感知器。开源情报可以从更广泛来源、更多样形式的数据来源中获取数据；另外，此阶段的数据挖掘、和信息分析技术也得到飞速发展，促进了开源情报分析技术进步。第三代开源情报分析阶段对应 Web3.0 阶段，即语义网阶段。随着语义网络，特别是社交网络的发展，社交网络中除了包含文本、图像和音视频多模态媒体信息之外，还包含大量元数据、连接信息、关联信息等语义信息，网络信息类型、内容更加丰富，实时性和关联性更强。另外，此时人工智能、机器学习、数据挖掘和知识工程等开源情报的相关分析技术也取得长足进展和广泛应用。第 3 阶段开源情报将通过机器学习、数据挖掘和知识获取等先进技术从语义网中获取更加大量、实时、含义丰富的信息，实现更隐蔽传递和高效知识激活。

现阶段，开源情报正处于第二代向第三代演变阶段。越来越多的具有情报价值的信息来自开源的互联网、移动互联网和物联网等公共开放网络。第二代开源情报在单个技术的研究成果比较显著，但存在最大问题还没有形成开源情报技术体系，开源情报技术研究的系统性不突出。因此，也使得向第三代演变过程中进展比较缓慢。另外，第 2 代开源情报数据主要包括网络公开的文献、新闻等数据，情报处理主要是进行多语言的翻译、多源数据的整合，而第 3 代开源情报数据将主要采集社交网络等信息，而且相应的智能化技术手段和工具需要更高。开源情报的情报周期包括收集、处理、分析、生产、利用和传播等环节，但在第 2 代向

第 3 代演变过程中，需要重点变革是四个关键步骤：收集、处理、利用和传递。这 4 个阶段是第 3 代开源情报技术挑战所在。

（1）第 3 代开源情报的收集内容包括传统新闻媒体内容、文献内容等互联网信息外，将更加关注各种社交网络、移动网络和物联网等信息，甚至还包括电商平台数据和车联网数据等。收集阶段的挑战主要包括：社交媒体的多模态媒体数据的完整获取、不断演变社交媒体主题时序的数据完整获取，热点主题数据的实时获取等。另外，由于社交媒体的使用人口分布和阶层分布不均匀，对收集数据的分析会受到更多因素的影响；对同一信息及时追踪和早期发现也是数据获取难点。

（2）在开源情报处理阶段的主要挑战是如何实现对非结构化数据智能处理和语义理解。以社交媒体内容为代表的多模态数据往往都是以非结构化格式呈现，但却拥有大量可用的信息。这些社交网络数据构成复杂、多语言、非正规表达，还可能包含排版错误，给开源情报的处理工作带来了挑战。另外，在开源情报处理和分析阶段还面临的挑战是多源异构数据融合问题；异构社交媒体内容的完整聚合和精确融合是情报加工的基础。

（3）开源情报的利用阶段挑战主要体现为如何有效地确认情报的真实价值，具体包括验证真实性、评估可信度和知识情境化。验证指验证真实性，评估可信度指验证是否可信，知识情境化是指根据语境和情景来激活知识。尤其是对社交媒体内容的信息真实性和可信度的评估是开源情报的利用阶段最大难点。

（4）开源情报的传递阶段，需要解决的问题是如何将情报产品和以安全的形式提供给消费者。目前的挑战主要情报高效的表示形式，以及高安全性和高隐蔽性的传递机制和技术。

随着开源情报领域的不断发展，面向开源情报处理的智能技术也有相应的发展和变化。在 Web3.0 阶段，机器学习、知识图谱和数据挖掘等技术使得情报加工的词汇分析、网络分析、地理空间分析变得更加高效；自然语言处理、计算算法、自动推理为主导的计算机技术提升了开源情报领域获取和生产的效率，扩展了情报界处理信息和发现信息的情报价值的能力。另外，社交媒体分析方法是第 3 代开源情报领域核心技术。社交媒体分析方法包括词法分析、关键词分析、情感分析、立场分析、地理空间分析等。词法分析解析语言背后的含义、推断出人的信息，包括年龄、社会阶层、经济背景和教育程度等人口统计特征等；关键词分析关注词在给定的句子或文章中出现的频率；情感分析将个人或群体的观点归类，发现群体的思想状态；立场分析挖掘观点背后隐含的内容。另外，社交网络分析将个人之间的关系解释为一系列的连接，目的是了解社会的二元、三元以至于多元关系，支持对关键目标的跟踪，挖掘目标之间的隐形关联。

总之，随着 Web3.0 即语义网的发展，开源情报正在转向第三代开源情报，开源情报技术还面临很多挑战和机遇。开源情报领域需要寻求新方法来分析和利用不断增加的新的信息来源，应对不断变化的新的情报需求；借助机器学习和人工智能、知识工程和数据挖掘等先进技术和工具，对海量的数据进行实时处理、分析、利用、传递；另外，开源情报应用也正在从国家安全和社会反恐等领域向科技、商业、金融等领域扩展。

13.3. 领域关键技术进展及趋势

13.3.1. 采集

开源信息情报是指从公开或半公开的网络信息资源和可获得的综合应用大数据资源中通过运用相关技术手段收集和挖掘分析出的情报信息相关内容。随着网络技术的发展，网上信息资源日益丰富，网络开源情报信息的传播形式也逐渐多样化。网络开源情报的信息除了在一些传统媒介和其相关的网络化产品之外，社交媒体、网络社区和智能搜索引擎，也成了获取情报的新型服务媒介，开源网络情报的收集渐渐成为某些机要部门获取相关特定情报的一种主要方式。

数据采集技术可分为公开数据采集、半公开数据采集、匿名数据采集等多种技术路线。面向开源数据情报主要为公开数据或半公开采集方式，采取的主要手段包括网页爬虫、数据包还原（dpi）、用户端模拟采集等技术去获取相关数据。

爬虫技术又称网络蜘蛛、网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。网络爬虫按照系统结构和实现技术，大致可分为以下类型：（1）通用网络爬虫：爬虫覆盖尽可能多的网络，如搜索引擎。（2）聚焦网络爬虫：有目标性，选择性地访问万维网来爬取信息。（3）增量式网络爬虫：只爬取新产生的或者已经更新的页面信息。（4）深层网络爬虫：通过提交一些关键字才能获取的 Web 页面，如登录或注册后访问的页面。通过各类爬取方式结合，高效获取到网页中承载的文字、音频、视频等相关内容。

DPI(深度包检测)信息采集是通过专用设备对网络的关键点处的流量和报文内容进行检测分析，可以根据事先定义的策略对检测流量进行过滤控制及内容解析，能完成所在链路的内容精细化识别提取、业务流量流向分析、业务流量占比统计、业务占比整形等功能且可实现平台及用户侧均透明无感，目前应用场景较为广泛，为大型信息系统广泛应用的数据采集方式。

用户端模拟采集主要是针对部分 app 类或者有爬取防御机制的网页中内容数据采集，此类平台界面多采用特殊框架构建，且页面的层级链接方式多变，较

多内容需要用户登录或者进行操作才能访问，针对这一类信息数据，多采用虚拟机方式模拟用户终端使用方式，采用模拟器模仿用户点击、登录、浏览等各种行为的方式进行相关数据的采集。因为此类采集方式可实现用户的各类行为，除信息采集功能外，还可通过用户行为进行信息的传播干预等动作，可实现诸多场景的应用。

13.3.2. 传输

随着我国综合实力的不断增强，国家利益在全球范围内快速扩展，对安全情报传输的现实需求日益迫切，急需建立起适应新对抗环境的安全、隐蔽、高效的情报信息隐蔽传输技术体系。

现有的情报传输系统主要依赖加密技术保护秘密信息的内容。但是这类系统对所传递的密文以及传递信息的行为并没有进行专门地隐藏，因此在非受控的网络环境下，密码通信的通信行为和所传递的密文容易被敌方获知。一旦通信源或通信人被确定，敌方一方面可以通过监听获取大量的密文数据，利用当前大数据关联分析、高性能计算与定点打击技术对密码通信行为和信息内容进行数据挖掘与分析；另一方面，敌方可以通过分析“元数据”（包括通话号码、通话时间、通信位置以及手机 IMEI 号等背景数据）推测目标对象的决策习惯、社交关系、行为特点，进而获得敏感或重要信息。因此，作为密码通信急需且重要的补充手段，开展复杂环境下隐蔽通信技术的研究成为当前情报传输领域的前沿热点。

隐写术（steganography）是隐蔽通信系统中的核心技术，其主要研究如何将秘密信息高效且安全地嵌入到其他信息载体中，掩盖信息的存在性从而保障其安全，它是加密技术的必要补充。隐写技术及其应用曾出现在许多古代东西方的文字记载中，最早可以追溯到 Herodotus（公元前 486-公元前 425）所著的 Histories。近代也有很多隐写应用的例子，如隐形墨水和伪装物品等被广泛应用于 20 世纪的两次世界大战和间谍活动中。20 世纪 90 年代以来，随着数字多媒体和互联网的逐渐普及，古老的隐写技术获得了巨大的发展。近年来，基于数字多媒体数据的隐写技术因其在个人和商业隐私保护、情报与军事、国家安全领域不可替代的作用，正日益引起国内外学术界和相关部门的高度重视，并成为当前多媒体内容安全研究领域的热点之一。隐写技术的相关研究也受到政府和国际组织的资助，包括美国国防高级研究计划局(DARPA)、空军科研处（AFOSR），自然科学基金委（NSF），俄罗斯科学基金委(RSF)，欧盟（EU）等，单项资助金额高达 850 万美元。

现在主流的隐写术通常是对单一数字载体采用单一隐写算法实现隐蔽信息

嵌入。但随着近期新兴的深度学习和人工智能等技术的飞速发展，基于统计分析的信息隐藏分析技术随之也产生了跨越式的进步，分析者有能力设计出准确率接近 100%的单一模态隐藏统计分析模型，从而使这类信息隐藏方法失效。为了抵抗单一载体上的统计隐写分析方法，伴随着当前网络模式多样化的共享和发展，逐渐发展起基于网络富媒体的多维信息隐藏方法，甚至更进一步发展成以网络大数据环境下的多维信息隐藏新架构。清华大学的黄永峰教授从载体形式、机密信息形式以及隐写技术手段等方面，将隐写术的历史演变划分为三个时代：第一代是早期的物理载体隐写，其主要采用物理载体进行信息隐藏，秘密信息主要是语义表示形式，采用的隐藏方法主要是一些工艺技巧。第二代是当前主要研究的数字载体隐写，其主要采用数字载体进行信息隐藏，秘密信息转化为比特表示形式，采用的方法主要是信号处理技术。第二代隐蔽通信技术面临的关键科学问题是：载体的隐藏容量和隐蔽性之间的调控机理。第三代是随着技术的发展即将跨入的网络空间隐写，其以整个网络空间为隐藏载体，机密信息由第二代的比特表示转向语义表示，技术方式也由信号处理技术转向数据挖掘技术。黄永峰教授早在 2013 年就系统阐述了网络多维信息隐藏和矢量隐写等核心概念，并指出第三代隐蔽通信技术面临的关键科学问题是：单位组合载体的隐藏效率和隐蔽性之间的调控机理。目前，关于隐写术的“三代论”分类体系已经获得相关领域研究人员的广泛认同，实现隐写术从第二代向第三代的跨越发展也成为当前情报隐蔽传输领域的研究热点和前沿。

13.3.3. 存储

目前开源情报已经成为一种国家竞争战略，开源数据日益成为一种新的自然资源。从开源信息获取、存贮、搜索、分享、挖掘到展现，开源情报呈现了前所未有的复杂性：首先开源数据的规模很大；第二是数据具有多样性，不仅有结构化数据，还有大量的非结构化信息；第三是数据的价值密度低，形象地说就是要从稻草堆中找到针；最后是开源情报具有实时性，情报线索的价值有时间价值。然而，利用传统的数据存储和计算技术已经很难实现高效的处理，针对目前这种现象，对于开源数据存储需具有高可靠的架构设计，完全分布式的、多副本机制的、对等的、不共享的系统，没有单点故障或瓶颈。这种架构系统能随着数据的增长而线性增长，每新增加一个节点能同时增加系统的性能和存储容量。

开源情报存储采用分布式集群存储。分布式集群存储将海量开源情报数据压力分散到多个并发存储节点，数据和元数据均匀分布于各个节点上，避免资源争用，系统性能（吞吐量）按照比例扩展，并且各个存储节点之间负载均衡，有效

避免单节点性能瓶颈，可根据业务增长需求进行平滑扩容，使得分布式存储架构系统具有更好的扩展性；同时，对于开源情报多模态数据（结构化数据、半结构化数据、非结构化数据）采用柔性多引擎技术，对于不同的应用需求采用不同的引擎来对外服务，包括全文检索引擎、知识图谱引擎等。全文检索引擎基于ElasticSearch 构建全文检索库，全文检索可以具备多线程设计机制、分布式检索和负载均衡调度、无单点故障、高可靠性、可扩展性好等特点。

引入大数据的预处理机制，基于知识图谱构建的知识库，对开源情报数据进行知识图谱化处理，丰富数据管理模式，加强数据处理能力，加深数据分析，提高数据价值，实现从数据到知识的转化。整合非结构化、结构化开源情报数据，以知识图谱技术为核心，构建复杂知识网络并实现对复杂知识网络结构的索引、存取、检索和关联匹配等底层支撑性服务；利用知识挖掘抽取高质量的预测模型、推断模型等战法成果输出，知识推理引擎关注在已有的挖掘结果上，支撑实时化、海量稀疏的实体关系上的关联关系定位。

知识图谱引擎采用相关动态本体论技术、自然语言处理技术，构建全局对象-属性-关系的知识图谱库。基于深度学习的自然语言处理，利用语义分析、深度学习的方式，对各类结构化信息和非结构化信息数据关系挖掘、数据分析、文本语义分析等，抽取出目标人、物、地、组织机构、虚拟标识号等实体，并根据实体的属性联系、时空联系、语义联系、特征联系等建立相互的关系，构建一张具有业务特性多维多层的实体与实体、实体与事件等的关系图谱网络。综合提高开源情报深化应用能力，实现从数据到知识到智能的升级转变，有效支撑情报业务信息化向智能化应用。

综上，随着大数据、云计算、物联网、5G 等新技术的发展，开源数据会呈爆发式增长，这些数据所需的存储空间也非常大，目前基本上采用分布式的方式进行简单存储，而正是由于这种存储方式，存储的路径视图相对清晰，而数据量过大，导致数据保护，相对简单，黑客较为轻易利用相关漏洞，实施不法操作，造成安全问题。因此，在考虑核心开源情报数据存储时，应考虑按数据密级分级分类进行存储和管理。

13.3.4. 处理

开源情报数据处理是指对已采集并存储在数据库中的原始情报数据进行清洗、分类、标注、元数据抽取、知识抽取等工作，将原始数据加工为可直接调用的内容，为下一步情报分析提供经过处理的高质量情报知识。用形象的比喻可以理解为“摘菜”、“洗菜”、“切菜”等过程，为厨师“炒菜”（情报分析）提供可直接

用的材料，当然实际过程比这更为复杂。与开源情报数据处理紧密相关的概念和技术包括“数据治理”、“数据科学”、“知识管理”、“自然语言处理”、“语音识别”、“图像识别”、“视频识别”等，这些概念和技术之间有大量交叉。

从需求角度，对海量数据的处理非常关键，否则这些数据将成为“死数据”，难以有效利用。如美国 21 世纪前 10 年建立了大量新的情报收集系统，包括采集了大量互联网文本、音频、视频数据，也包括大量无人机采集的图像和视频数据，这些资料的数据量异常庞大，TB、PB 成为储存量常用词，往往难以得到处理或难以及时处理，使得很多情报数据失去了大部分价值。为了解决这些问题，美国在 21 世纪的第二个 10 年通过多个项目投入大量资金，在文本、图像、视频等数据处理方面进行了广泛、深入的研究。从全世界技术发展来看，近 10 年以来，随着人工智能技术的发展、相关需求的牵引和大量研发工作的开展，开源情报数据处理技术得到了长足的发展，为情报分析处理提供了更好的基础。

数据清洗主要是将重复、多余的数据筛选清除，将缺失的数据补充完整，将错误的数据纠正或者删除，并提供数据一致性，还包括重整数据格式，使之成为机器可读的格式，或与目标数据模型一致，整理成为我们可以进一步加工、使用的数据。数据清洗从名字上也看的出就是把“脏”的“洗掉”，指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。数据清洗是与问卷审核不同，录入后的数据清理一般是由计算机而不是人工完成。

在分类和标注方面，近年主要发展和应用的技术是机器学习、深度学习和知识图谱等。构建像人类一样能较准确地对数据进行分类和标注的模型，需要大量训练数据，训练数据必须针对特定用例予以适当分类和标注。数据的主要类型包括：文本、音频、图像和视频。数据分类根据数据类型和特点有多种方法和技术，如 K 均值聚类等。数据标注包括情绪标注、意图标注、语义标注、命名实体标注等；音频标注即包括语音识别，也包括语音情绪识别、人群类型识别等；图像标注在广泛的应用中至关重要，包括计算机视觉、机器人视觉、面部识别以及依赖机器学习来解释图像的解决方案；视频标注相对处理难度大一些，简单的标注技术包括识别视频里的语音转换为文本进行标注，以及对视频中的图像进行识别等，

难度大的是识别视频中发生的事件，将视频核心内容转为文本，以便实现大量视频内容的快速检索和进一步分析。

元数据抽取指抽取图像或视频等文件中的元数据信息，如照片文件中隐含的拍摄地点经纬度坐标。知识抽取的范围较广，既包括用自然语言处理技术对文本中的命名实体、事件、时间、地名等进行抽取，也包括对图像和视频信息进行抽取，是比标注要求更高的数据处理方法。

13.3.5. 分析

社交媒体情报是开源情报的一个重要方面。以 Facebook、Twitter、微博、微信为代表的新型社交媒体包含大量针对新闻时事、政策法规、消费产品等话题的主观评论文本，反映了用户个体的观点、情感、态度、情绪等。对社交媒体中的文本进行自动情感分析、挖掘和管理，是社交媒体情报的重要组成部分，对于国家、政府、企业和个人，都具有及其重要的实际意义。对于国家安全机构，需要及时了解网络信息内容的安全，识别是否存在反动、诈骗、不良信息传播的可能性，以便及时防范、引导和管理；对于政府管理部门，网络信息可为了解民众意向、制定和改进政策提供重要依据；对于企业单位，基于网络信息了解用户对产品的意见和建议，进行精准营销，改进产品性能和售后服务；对于网民个体，在选购产品和服务之前，可以了解人们对于产品的综合评价、优缺点介绍以及注意事项等等，为生活带来极大便利。

情感分析，也称作观点挖掘，是对文本中的主观信息（比如观点、情感、评价、态度、情绪等）进行提取、分析、处理、归纳和推理的技术。早期的文本情感分析方法主要分为两类，即基于情感字典的规则化方法和基于情感特征的统计机器学习方法。随后的情感分类研究工作也相应分为两类。一类基于情感词典的方法，根据情感词典所提供的词的情感倾向性，结合语言知识和统计信息，进行不同粒度下的文本情感分析。另一类基于统计机器学习的方法，大量的研究集中在如何在文本表示层面寻找更加有效的情感特征，以及如何在机器学习模型中合理使用这些特征上，这些特征包含：词序及组合、词类、高阶 N 元语法、句法信息等。

早期的情感分析主要针对 BBS、电商网站中的评论文本，自 2009 年以来，以微博为代表的社交媒体逐步兴起，极大增加了网络主观信息内容的规模。社交媒体情感分析的研究始于针对 Twitter 的情感分析与应用，经过几年已经发展成为文本情感分析新一轮的研究热点。从任务的角度，覆盖了 1) 社交媒体文本预处理与规范化；2) 社交媒体文本表示与特征工程；3) 社交媒体情感分析话题与

领域适应；4) 社交媒体自然标注与半/无监督情感分类；5) 社交媒体情感词典自动构建；6) 社交媒体情感分析应用等多个方面。从方法的角度，随着深度学习的深入发展，大量的神经网络模型被引入到社交媒体情感分析任务中，如卷积神经网络、循环神经网络、递归神经网络、注意力机制网络等。近年来，随着预训练语言模型的兴起，以 BERT 和 GPT 为代表的预训练语言模型在不同的社交媒体情感分析任务中均取得了较大的成功。

社交媒体情感分析的应用主要涵盖舆情分析和社会治理、社交媒体心理健康监测、电商评论分析以及辅助各类商业决策等方面。

开源情报中除了对情报对象的主观情感进行分析外，还有一部分很重要的内容是对客观实时发生的事件进行分析。开源情报，在某种程度上可以看作是一个巨大的信息空间，瞬间变化的国际态势、复杂多元的政治外交等要素、实时回传的传感器数据，共同构成了开源情报的“数据汪洋”。开源情报从某种角度看可以认为是由一系列的事件组成，通过发现已有的知识库与新近发生国际事件之间的关联，通过已知事件推导预测未知事件，对夺取情报感知优势和辅助决策优势至关重要。在复杂纷繁的海量情报事件数据中发现其内在规律，快速有效地感知态势、提供决策所需数据资源，可以牢牢把握住未来情报战场的主动权。

事理图谱技术是近年来对开源情报事件数据进行分析的有力手段。事理图谱技术的出现，可帮助决策人员从海量情报事件数据中分析获取有价值信息，进而为指挥员决策和筹划等提供有力支撑。在拥有基于海量情报数据构建起来的事理图谱的情况下，可通过设计科学的算法，将人和智能系统模块角色进行统一描述、计算、调度、协同及优化，对过往情报事件进行自动分析并发现其内在规律，对于研究情报对象方重点任务决策特点、习惯偏好、行动策略提供有力支撑；在已构建好情报事理图谱的基础上，根据当前国际态势，还可借助事件推理技术，对已有事理知识进行建模进而推理出事件的演化方向，实现“人谋”与“机谋”的深度融合，对国际形势的未来走向进行预判，为指挥人员决策指挥提供辅助支持。伴随事理图谱技术在情报领域的广泛应用，人们有望从情报数据中“读懂”并“感知”未来。

事理图谱的数据来源可以是文本、图像、语音、视频等多模态信息，同时，语言也不局限于中文，可以覆盖多种语言。事理图谱的节点是事件，边是事件之间的关系，包括但不限于因果关系、顺承关系、子事件关系。事件需要预先定义好事件类型体系以及事件元素，例如：“政变”属于“冲突事件”类型下的“政治冲突”子事件，事件元素包括：地点、需求、发起者、受事者等，而事件元素如果是实体的话又可以跟知识图谱当中的实体相关联上，将实体关系及属性关联到事理图谱当中，帮助更好的理解事件。事件间的因果关系表明了某一情报事件发生后

可能引起的后续情报事件是什么，并且边上还会标注因果强度，表明因果发生的概率。顺承关系主要是情报事件按照时间偏序关系依次发生，重在记录事件的演化规律和模式。子事件主要是从更加细致的角度不断理解事件的更多发生细节。

事理图谱主要涉及到事件抽取、事件关系抽取、事件表示学习以及事件推理等相关技术。事件抽取技术主要涉及的方法包括基于规则的方法、基于统计学习的方法和基于深度学习的方法，目前国际主流技术发展趋势是采用基于深度学习框架进行事件抽取，尤其是随着以 BERT 为代表的预训练语言模型发布后，事件抽取的方法较多采用基于预训练语言模型的深度学习框架。事件关系研究热点主要集中在因果、顺承、子事件关系的识别和抽取上。尤其是事件因果关系的研究被认为是对人工智能可解释性的重要支撑。在已经构建好的事理图谱基础上进行基于图的事件表示学习，并进行事件推理工作是人工智能领域的一个难点工作。国内外均有团队在进行科研攻关，并取得了一定进展。

开源情报中的事件分析还可以有力地支持我国军事智能的发展，自动感知战场环境、理解战场事件数据信息，把数据优势转化为制胜优势的有力抓手。有效分析并成功利用海量战场事件数据，可能会成为决定战争胜负的关键因素。

13.3.6. 应用

开源情报来源广泛，包括因特网、报刊、广播、电视、无线电、地图、数据库、灰色文献、电话、通信和商业活动等，因其低成本、低风险和高效益而日益成为情报的主体，同时受到包括各国军方、国家机构、学术团体、企业乃至个人等广大用户的高度重视。

浩如烟海的开源信息需要通过一定的技术加以采集、提炼和分析研究，才能形成用户所需要的情报。因此，开源情报技术的应用非常广泛。随着各行各业对开源情报的需求不断增长，开源情报技术的应用领域也不断拓展，主要包括以下几个方面:

一是国家安全领域。各国国防和安全等国家强力机构高度重视研究和应用开源情报及其相关技术。美国军方和中央情报局等一直高度重视开源情报技术来获取关键情报。在击毙基地组织头目本·拉登和斩首伊朗伊斯兰革命卫队司令苏莱曼尼的过程中，美国都依靠了大量的开源情报。2021 年 5 月，以色列对加沙地带的哈马斯分子实施斩首作战。行动之前，以方事先致电加沙地区的无辜居民要求其撤离。此种作战方式在全世界引起了强烈反响，凸显了开源情报的重要性。通过网上和其它公开渠道获得对象国国防军事合同信息，可掌握对手的军事技术研发动态。此外，在反恐作战中，由于恐怖主义势力往往借助互联网和社交媒体

平台进行通讯联络、宣传煽动和组织实施暴恐活动，可以利用开源情报技术从各类公开渠道获取有关恐怖组织、恐怖分子的关键信息和数据。以社会网络分析方法绘制恐怖活动组织和人员关系网络，为反恐决策与反恐战术行动提供重要的情报支撑。从开源途径获取的情报，可以大大提高反恐行动的效能。

二是经济发展领域。开源情报在经济发展领域也正发挥着越来越大的作用。特别是在当今世界产业链和供应链重组的情况下，各种物资和原料的来源、产出、销售等都已经成为至关重要的数据，影响着经济产业的布局、调整和发展，通过开源情报技术可以实现有效监控。另外，随着互联网的逐步发展，以互联网公开信息源为依托的开源网络情报目前已广泛应用于不良资产清收处置全过程，涵盖了至少四大类，包括企业主体相关查询、涉诉信息查询、资产信息查询、投融资查询等。开源网络情报调查可以有效解决信息不对称的障碍，从而实现深度人物画像、联系方式重建、隐匿关联资产、资产转移线索等的调查，提高不良资产的清收效率。

三是社会发展领域。社会的发展同样离不开开源情报技术。在城市管理、农村治理、婚姻家庭、法律诉讼、人口调查、老龄化、乃至幼儿和中小学教育及应急处理等社会发展各方面，开源情报技术都发挥着越来越大的作用。通过广泛采集以人为中心的数据和遍布城乡各处的监控摄像数据等开源信息，根据一定的算法，可以分析研究出人员活动的规律、社会需求和发展趋势以及目前存在的问题等；掌握地域内的文化、宗教和种族的构成成分，以及社会成员的信仰、价值观、习惯与行为等，并据此提出有针对性的、合理性对策建议。

在上述各领域，沃德舆情大数据系统依靠独立的大数据采集能力和独特的算法，提供了很好的开源情报平台，为党政军和有关企业机构作出决策提供了有力支撑。

13.4. 领域产业发展现状及趋势

开源情报的主要生产者包括政府关联机构和私营机构，主要用户包括政府公共部门、军事安全部门、国际组织、商业组织、科技组织等。从技术上讲，任何知道如何使用工具和技术来访问信息的人都被称为使用了开源情报。例如美国情报局、军队和执法机构情报员，IT 安全专业人员，私营企业和私人侦查人员都在使用该方式[2]。开源情报产业，主要集中在技术、工具、应用三个维度，一份关于开源情报市场的新报告预计到2027年，OSINT的市场估值将超过200亿美元，开源数据的增加、加上网络威胁、恐怖主义和其他非法行为的增加，会客观上推动该行业的增长。

13.4.1. 国外产业情况

近几年,欧美等发达国家愈发重视开源情报工作,逐步建立起比较完整的开源情报工作体系。

首先，政策支持开源情报。1941 年 2 月，美国成立对外广播监测处，成为最早的开源情报研究机构，这标志着美国对情报来源与内容的认识开始发生转变。1992 年颁布的美国国家安全法开启了美国情报界的改革，奠定了开源情报在情报界中的地位，该法最早提出系统性发展开源情报的必要性并建议成立开源项目办公室。1995 年 2 月，该办公室发布《公开资源战略计划》，正式确立了美国开源情报的制度规范，包括开源信息系统、开源需求、信息收集、信息加工与转换、情报共享等政策。911 事件之后，美国出台了《美国情报与打击恐怖主义改革法》，明确了开源情报的规范概念,成立国家情报总监开源中心。据美国中央情报局统计，2007 当年情报总量中超过 80%来自开源情报[4]。

2001 年北约组织在《开源情报手册》中定义了开源情报，相较于美国国会的定义，北约组织将公开源情报的地位提升到情报科目的水准，并将其看作其他情报科目的基础[5]。

澳大利亚在 2001 年建立了国家开源情报中心，为各政府部门和商业机构提供社会安全、跨国犯罪、恐怖主义、激进主义等领域的开源情报检测、研究及分析支持。

其次，开源情报市场规模逐渐扩大。据美国国土资源安全研究公司（HSRC）发布的《OSINT 市场与技术 2017-2022》调查结果显示，开源情报市场规模在 2020 年超过 50 亿美元，并有望在 2021 年至 2027 年间以超过 25%的复合年增长率增长。从公开来源收集数据以获取关键业务洞察力的需求不断增加将推动行业增长[6]。OSINT 使组织能够更好地了解竞争对手采用的策略，并采取对策以扩大在不同客户群中的市场占有率。

德国数据分析部门在 2020 年创造了约 4266 万美元的收入。开源数据的数据分析已成为清理、组织和分析大量数据的主要技术。数据分析确保企业部署的关键资源用于挖掘和提取有价值情报数据，而忽略大量难以理解的无用数据。德国安全机构通过人工智能技术对未分类数据进行量化建模，以提高国家安全情报质量。

日本文本分析市场收入在 2020 年超过 2857 万美元，由于企业越来越多地采用商业智能（BI）工具来提高业务流程的效率，文本分析可帮助公司通过集成的 BI 工具分析非结构化数据，并补充了 BI 工具的功能，使其可以从早期丢弃的文本数据中深入挖掘。大量非结构化数据的激增使得企业更多地采用文本分析解决

方案和 BI 软件来解决问题。提供基于文本分析的开源情报解决方案的企业正在通过附加服务提高市场竞争力，例如实时语音到文本 API 和光学字符识别 API，这些技术补充了从文本分析中获得的结果。

韩国的军事和国防领域的市场份额预计在未来 5 年将达到 25%以上。由于利用人工智能和大数据技术收集安全情报的需求日益增长，韩国武装部队越来越多地采用开源情报解决方案。融合开源数据和机密信息，可以有效地优化作战情报、识别正确的战术战争策略、最大限度地利用军事资源。韩国国防部宣布计划在印度举行的 2020 年国防博览会上展示支持人工智能的军事技术，其中的开源情报技术是对客户的最有吸引力的部分。预计到 2027 年，亚太开源情报市场的复合年增长率将超过 28%。亚太地区行业的特点是大型企业在网络安全、网络智能和大数据分析方面的投资不断增加，以获得市场影响力。

国外开源情报市场仍然适度分散，科技巨头如泰雷兹集团、达索系统、NICE Ltd.和 Verint Systems，占据了大部分市场份额。市场上的知名参与者越来越重视提供先进的开源情报解决方案，以应对不断变化的威胁动态，尤其是在新冠疫情流行期间。例如，2020 年 1 月，Palantir Technologies 在日本推出了用于安全、健康和福祉的真实数据平台。该平台旨在改善日本的医疗保健，简化跨行业的供应链，并提高安全性和弹性。

另一个主要市场发展包括澳大利亚中小企业 Fivecast 于 2020 年 11 月发布的有针对性的海量数据收集工具，该工具可以跨开源平台收集结构化和非结构化数据，包括文本、图像和视频等。OSINT 市场还见证了主要参与者之间的几个战略联盟，以推出具有附加功能的新产品并保持收入份额和盈利能力，例如 Thales 与 Fivecast 的合作以提高 Fivecast 的产品能力。

一些关键的 OSINT 市场参与者是 Context Information Security Limited、CYBELANGEL SAS、Dassault Systems SE、Dataiku Inc.、Digimind SA、Expert System SpA、INTELLEXIA Srls、Intrinsec Security, Inc.、IPS SpA、KB Crawl SAS、Maltego Technologies GmbH , NetSentries Technologies FZCO, NICE Limited, Octogence Tech Solutions Pvt. Ltd.、Palantir Technologies, Inc.、Recorded Future Inc.、SAIL LABS Technology GmbH (HENSOLDT)、Social Links Software BV、Thales S.A.和 Verint Systems, Inc.等。

最后，技术成为开源情报的发展关键因素。在开源情报发展过程中，技术成为瓶颈。情报收集过程主要面临以下技术挑战：信息过载问题，有价值信息容易被海量无效信息淹没，此外，庞大的数据量成为情报采集和分析效率的一大挑战；精准采集问题，开源数据来源、数据形式多样性强，给数据采集工作带来了困难；文本分析问题，文本形式数据需要通过语义分析，得到上下文、主题等信息，过

滤得到有效信息，在此过程中还存在跨越多语言的问题；计算能力过度依赖，人工智能、大数据等技术需要强大的算力支持。

美国中央情报局和国防部等机构已开始投资人工智能公司，开展项目合作，成立国防创新实验单元，通过研究人工智能算法优化开源情报解决方案。美国国防部加强云计算技术建设，在图形处理单元和图形处理器的计算能力上做了大量投资，并开展量子计算研究，旨在为人工智能技术发展提供算力支持。到 2022 年，美国国家地理空间情报局希望尽可能多的情报、监视和侦察部门实现自动化，而仅将最重要的决策任务留给人类分析师。

13.4.2. 我国开源情报发展现状

在互联网+和大数据的时代背景下，我国开源情报政策环境积极向好，行业多领域开展业务，政府、军队、公安行业需求明显，落地模式相对成熟，但商业行业由于政策限制，开源情报有待推广应用。

首先，政策环境利好，传统安全领域向商业领域扩展。开源情报的研究对象是公开信息源，向用户提供能够决策的情报依据。开源情报依据用户需求或潜在情报需求，在法律法规允许、伦理认可的搜集手段从公开渠道获取数据，并对其进行处理、提炼与分析产生的有价值的可信情报。《数据安全法》和《网络安全法》对数据隐私和数据安全使用进行限定和说明，《网络数据安全使用法》也即将出台，相关的数据合规使用规则正在不断完善，这也为开源情报的数据合规获取提供政策保障。

开源情报的前期需求在传统安全领域，如军事领域、公安情报领域、政府部门。典型的应用有通过开源情报实施战区人文地理和安全风险态势评估、开源反情报工作、防大规模杀伤性武器扩散、跨国人道主义救援、反恐、反诈骗、网络安全与打击网络犯罪。在军事领域普遍认为公开资料是获取军事情报的一条经济、安全、迅速渠道，为解决开源情报的有效分析和利用问题面向军事领域的 Web 开源情报主题挖掘方法并通过实际 Web 数据对军事开源情报主题生成效果进行评估。在公安行业，可以将互联网、社交媒体、研究报告等信息与公安内网信息进行关联、聚类与协同分析。在反恐方向，由于恐怖组织的地域性和宣传目的，在互联网公开途径往往能获能够取恐怖组织内部情报，互联网开源情报应用于反恐工作中具有重要的现实价值和良好的适用性。在政府管理方面，开源情报在情报风险管理中的应用也是一项重要的应用路径，能够利用互联网上的开源信息通过开源情报分析选择更优的干预策略，实现更好的干预效果，如 2020 年初国内新型冠状病毒肺炎疫情爆发，当年 2 月 14 日我国将生物安全纳入国家安全体系，

至此在政府管控安全威胁的预警与防范上增加了一定的挑战。

在非传统安全领域，近年来国内主要有企业竞争情报,以及反病毒及网络安全、商业零售、保险等行业、针对客户群、产品和市场的精准推送分析和应用。尤其在商业竞争领域，相比于企业机密，企业间的竞争和情报分析更多依靠公开信息进行，商业经济领域亟须让公开的数据、信息和知识发挥出情报的价值。

其次，技术成熟度高，各行业加速部署应用。开源情报是近期提出的学科名词，是以需求为导向的技术手段。我国在互联网和大数据的环境下，已经长期进行开源情报的技术研究和技术应用。在开源情报的数据层面，针对网络信息数据、网络流量数据、网络行为数据等互联网公开数据，在进行合规采集，政策允许下的公开信息收集后，在舆情、情报推送、案件侦办、反恐、商业情报咨询、医疗、客户精准推送等需求方向，都有长期的研究和技术应用开发，并在行业实践中取得良好效果。我国的开源情报在大量的实践应用中，技术手段不断趋于成熟。

国内大数据分析公司比如阿里、腾讯、百度等互联网大厂，有较好的数据分析处理技术，并且又有先天的数据存储优势，所以在开源情报工作中，有较好的技术和数据基础。市场上也出现了以开源情报为标签的一些科研公司，如北京知道创宇信息技术股份有限公司开发了开源网络情报工具 ZoomEye，福韵数据服务，乐思软件，拓尔思、天眼查、帆软等。

行业需求急切、技术相对成熟的环境下，开源情报在各行业加快部署应用。在公安行业，互联网犯罪的增加，网络空间安全威胁和社会维稳问题增加，对开源情报工作提出新的要求；在军队中，军事安全和国家安全在新的国际形势下愈发重要；政府管控手段急需优化，开源情报能为管控手段找到发力点，提高工作质量和效率；在医疗行业，新冠疫情的大流行带来很多健康问题和社会问题，开源情报的应用有助于提供更好的解决方案；此外，在教育、科技生产中，也都急需开源情报部署应用并发挥作用。

13.4.3. 现状问题和趋势

当前开源情报技术已经步入成熟稳定期，为了服务不同行业的应用需求，技术发展不断向着数据技术管理更高效、网络规模更广泛、技术服务更细致和政策保证更完善的方向发展。

开源情报在技术产业上主要有两个技术工作，一是数据源的获取，二是情报生成。需要开发互联网主流媒体和社交网站的数据采集工具，和部分基于网站和平台漏洞研发的数据采集工具，包括：元数据搜索、音视频采集、代码搜索、身份搜索、虚拟身份收缩、无线网络监测、数据包分析等。目前国际上进行相关技

术工具有 searx 元搜索引擎，支持匿名同时检索 70 多个搜索服务，能够根据目标集中收集汇聚相关数据，也可以实现 Tor 网络在线匿名采集。Twint 抓取工具，是针对 Twitter 的一种抓取工具，无需注册 Twitter 和 API 密钥，不需要身份验证，可以直接搜索，能够按地理位置和时间范围进行搜索。其它的还有一些常用的数据采集工具如：metagoofil 数据提取工具，可以基于谷歌搜索对列表中的 pdf，表格、word 进行元数据搜索，并自动汇聚数据。另外还有 spyonweb 网站注册信息收集器、crunchbase 商业公司信息收集平台、domainlq 域名采集器、dnsdumpster 侦查 dns、第三方子域爆破工具、ASN Lookup 地址采集器、built with 软件技术检测、waybackmachine 站点参数收集器、八爪鱼、contentgrabber、mozenda 网页抓取等数据采集工具。

情报生产技术产业，开源情报的目的性工作，也是体现数据价值的重要步骤。开源情报的情报分析工作在实际中和传统的大数据分析有一定的差别，更突出目标性和可用性，但却也继承着大数据分析的技术体系。市面上比较成熟的情报分析工具有 knime 分析平台支持数据挖掘和机器学习的外部扩展，openrefine 数据清理工具支持数据清洗格式转变能够自动形成标准化版本，r-pogramming 图形和计算统计编译器，weka 数据挖掘和机器学习集合工具，semantria 社交媒体情感分析工具，sas sentiment analysis 数据聚类和信息抽取工具等。

基于国际上认可的开源情报主要研究结构（开源信息搜集、开源信息整合与分析、开源情报应用），国内常用的开源情报公认的技术流程是：目标规划、信息搜集、信息处理、情报分析、情报分发。在目标规划上，主要以业务需求为导向，设定工作内容的实施步骤，在战略上制定规划，设定情报产出方向和目的。信息收集，是开源情报工作的重要步骤，开源数据的应用效果依赖数据全面性和可信性，主要的采集类别有：互联网可采集的公开信息（言论信息、网页信息、地址信息、流量信息、网络行为信息、技术信息、境外互联网信息等）、深网和暗网信息、报纸杂志、政府公告信息等能够合法合规获取的公开信息。政策上，在《数据安全法》出台后，需要行业部门注意信息的合规采集和使用，公开信息的获取也需要技术手段进行支撑，尤其是暗网数据和境外数据，需要特殊手段进行采集，如多语言网络爬虫、元数据采集、Tor 网络跟踪、ＵＲＬ解析、推特(Tweet)数据的监视算法等。信息获取并安全存储后进行信息处理，需要对信息进行加工和分析，针对目标需求，对信息进行分析和可信评估产出可信情报，主要的技术手段有：非结构化音视频数据、文本和语料数据融合处理分析，语言自动翻译技术、动态和静态目标发现技术，目标模型建立，可信度系数分析、探索性因素分析等成熟技术。情报分析，针对信息产出情报，进行分类聚合再分析，形成可用、可展示、可信的，能够提供决策的开源情报。在商业应用中，情报定向分发也需

要有技术支撑，商业目标客户的广告定向推送，医疗行业的情报分析，咨询服务行业的情报定制等方向，需要对情报进行差别和精准推送。

开源情报技术正在趋于成熟，但同时，开源情报技术也存在一些急需解决的问题。信息的可信性评估方面，信息量大、信息过载、真实性难以保证等固有缺陷，可能成为对手故意散播虚假信息和错误信息的源头。情报反馈方面，目前的开源情报遵循线性的步骤解决问题，各任务之间过于分明，忽略了各部分之间的交流和反馈过程，缺乏对目标情报的可用性评估。情报融合机制方面，中国各领域的情报工作各自为战，情报信息封闭，缺乏融合共享，存在资源浪费和重复劳动。个人隐私问题，网络数据的安全合规使用，需要一套基础网络技术架构和政策规范支持，怎样实现数据的可用不可见，分级分类管理，数据权益实现等关键技术急需出台相关建设标准和支撑技术体系。

开源情报处于行业应用增长迅速阶段，要真正发挥其技术优势和数据潜在价值。开源情报的重要作用和地位已被世界各国家所认知。国外正在将开源情报设立为一门独立科目，大数据和人工智能技术、数据安全防护技术在开源情报工作中，已经形成一种技术趋势，如构建基于知识库的主动式专题搜索引擎系统模型，设计基于互联网的开源情报挖掘系统，多源跨域数据融合分析技术，开源情报和基础情报的融合分析技术，人物画像与精准推送技术，数据隐私防护和权益归属技术，数据安全处理技术等多方面的技术，已经在开源情报的工作中起到重要的技术支撑作用。国内的社交媒体比西方起步晚，但是反信息摄取能力却比西方强，要使开源情报工作既要广泛应用技术手段，又不能完全依赖技术手段，需要技术与政策相结合，才能为开源情报工作提供有力支撑，发挥重要作用。比如在法律隐私保护问题、数据的权益问题、多部门的融合处理问题都需要法律法规、政策进行工作方案的制定。在开源情报的技术方面，未来还可能加入人类行为学和心理学等社会学科研究，辅助开源情报的行业应用。

13.5. 总结及展望

开源情报行业是一个集社会科学、数据科学、计算机科学和军事科学等多学科为一体的交叉行业。开源情报的核心特点是数据来源丰富但杂乱、情报范围覆盖广但密度低。随着互联网的兴起、社交媒体的发展与人工智能技术的飞速发展，开源情报工作在近年来得到了极大地推动，目前已进入了高速发展期。互联网上的大规模易获取数据为情报的来源丰富度、信息覆盖度提供了支撑。人工智能和相关的情报分析工具则可以帮助优化情报信息流，增强情报信息提取的质量和效率，从而在短时间内以更少的数据中获得更多的洞察。此外，开源情报中的开源

不仅体现在数据层面，目前还有许多能够被用于开源情报收集、处理、分析的工具已经广泛存在于开源社区之中，如何更好地整合利用这些开源工具，服务于开源情报业务,也是开源情报体系构建的重点方向。

在展望方面，开源情报将在数据源、处理技术和应用层面呈现巨大的机遇与挑战。

首先，在数据源方面，开源情报的数据来源将更加泛在、新颖、实时、海量。一方面，过去几十年来社交媒体、大数据、泛在网络的兴起，给开源情报带来了革命，新型的收集和利用活动给开源情报带来更多有价值的、新颖的数据源。另一方面，随着移动终端和 5G 网络的发展，数据的来源将从媒体机构转向实时发布消息的海量个人，这必将导致开源情报数据源的实时、海量和高噪音特性，为未来开源情报技术带来挑战。

其次，在处理技术方面，人工智能技术将重塑整个开源情报循环。通常，一个开源情报循环由收集、处理、利用和生产等环节组成，其中收集是指开源信息的获取，处理是指验证这些信息的方法，利用是识别信息的情报价值，生产是将价值送达情报用户。虽然在当前阶段，人工智能技术仍然无法替代情报分析人员的认知、背景知识以及思维方式。但是作为一种重要的辅助手段，人工智能技术在数据呈现、数据管理、假设生成与验证等方面已经逐步展示出无法替代的显著优势。通过利用人工智能对成倍增长的情报数据进行优化和感知，并结合上传统闭源情报领域的经验知识，能够为情报分析人员快速提供多维度、多角度、多层次的情报信息，并最终重新塑造整个情报行业。

最后，在应用层面，开源情报将逐渐从传统的安全、军事、政治等领域逐渐扩展到开放泛在领域。越来越多的组织寻求使用情报信息来改善客户关系、运营效率和研发过程。例如，许多药物公司使用海量信息资源来决策其药物的研发对象、目标和过程。同时，由于海量数据的积累，许多组织的高级管理人员越来越多地采用不同类型的分析来解决其业务需求和高效配置资源。例如，越来越多的科研管理机构使用科技数据情报来指导其资源的配置过程。总的来说，开源情报将被用于越来越多的领域，面向各种各样的任务，并无缝衔接到各种各样的需求解决过程中。

13.6. 附录 A：开源情报技术应用工具

目前，开源情报技术应用工具主要包括：

（一）电子邮件泄露查询

have i been pwned?

网站网址：https://haveibeenpwned.com/

（二）事实检查网站

1、Hoaxy

网站网址：https://hoaxy.iuni.iu.edu/faq.html

Hoaxy，可在线可视化文章的传播。

2、MediaBugs（媒体错误）

网站网址：http://mediabugs.org/

该站点可用于识别虚假或不正确的新闻，还可以查找正确的版本。

3、PolitiFact

网站网址：https://www.politifact.com/

Politifact 专注于事实检查新闻。Politifact 使用有用的评级量表对记者、政治人物和其他人的言论进行评级。是了解谁在讲真话和谁在说谎的有用方法。

4、SciCheck

网站网址：https://www.factcheck.org/scicheck/

SciCheck 是 FactCheck.org 的一项功能，致力于评估对公共政策有影响的虚假和误导性科学主张。

5、Snopes

网站网址：https://www.snopes.com/

该网站内容包括城市传说、谣言、神话、可疑的照片和视频，文章和公众人物提出的主张。

6、Verification Junkie

网站网址：https://verificationjunkie.com/

Verification Junkie 是一组工具，旨在帮助验证和事实检查信息以及评估目击者报告的有效性。

（三）黑客与威胁评估

Norse

网站网址：https://norsecorp.com/

Norse 拥有全球最大的专用威胁情报网络。拥有超过 800 万个传感器，可模拟 6000 多个应用程序。主页显示了实时攻击图以及有关攻击的实时信息。

（四）OSINT 图像搜索

Image Identification Project

网站网址：https://www.imageidentify.com/

Wolfman 图像识别项目使用算法来识别图像。

（五）公共记录（财产）

1、梅利莎数据属性查看器

网站网址：https://www.melissa.com/v2/lookups/propertyviewer/zipcode/

此工具可查看几乎所有属性的属性信息。输入一个邮政编码，然后使用地图或卫星视图一直放大到特定属性。单击特定属性以获取公共记录信息，例如完整地址，所有者名称，居民名称，价值，建造年份，建筑物和地段面积等。

2、Emporis 建筑物搜索

网站网址：https://www.emporis.com/buildings

Emporis 可搜索世界各地的建筑物，公司和设计/建筑图像。

（六）OSINT 搜索引擎

1、Google trends

网站网址：https://trends.google.com/trends/?geo=US

Google Correlate 可确定与现实趋势相关的搜索模式。它可以用来识别彼此相似的搜索模式。

2、millionshort

网站网址：https://millionshort.com/

millionshort 通过多种方式对搜索结果进行排序和过滤。示例包括受欢迎程度，电子商务，实时聊天，日期，位置等。此外，它会自动拉出通常在任何搜索结果中都占据首位的顶部站点（例如 Amazon.com，eBay，YouTube 等）

3、Shodan

网站网址：https://www.shodan.io/

Shodan 是物联网的搜索引擎和网络安全工具。它可以找到互联网上的设备，例如 Web 服务器，网络摄像头，设备，交通信号灯，甚至是发电厂。

4、TalkWalkerAlerts

网站网址：https://www.talkwalker.com/alerts

TalkWalkerAlerts 除了监视网络中的某些关键字外，它还监视社交媒体，博客和论坛。

（七）OSINT 社交媒体搜索工具

1、Facebook 搜索工具

2、TweetBeaver

（八）OSINT 工具网站

1、英特尔技术

网站网址：https://inteltechniques.com/

2、Hunchly

网站网址：https://www.hunch.ly/

Hunchly 是针对调查专业人员的在线证据收集工具。该软件会记录所有在线活动，以加快研究和发现过程。

3、Maltego

网站网址：https://www.maltego.com/

Maltego 是 Paterva 开发的软件工具。执法人员，法医调查人员和安全专业人员都使用它来分析开源信息资源。它可以在 Windows，Linux 和 OSX 上运行。研究人员使用该软件从各种来源收集数据和信息，并以图形方式显示它们。这有助于减少分析时间，建立连接并发现潜在客户。

（九）监视摄像头

opentopia

网站网址：http://www.opentopia.com/

汇总全球公共实时流网络摄像头和监视摄像机。

（十）运输–车辆，飞机，轮船

1、Flight Radar 24

网站网址：https://www.flightradar24.com/

该网站可查看全球实时飞行跟踪信息。每天跟踪超过 180,000 个航班。

2、海上交通

网站网址：

https://www.marinetraffic.com/en/ais/home/centerx:5.4/centery:50.8/zoom:2

该网站提供全球船舶跟踪情报。

3、飞机登记册

网站网址：https://registry.faa.gov/aircraftinquiry/Aircraft_Inquiry.aspx

搜索在美国联邦航空管理局（FAA）注册的所有飞机的登记册。

4、VINCheck

网站网址：https://www.nicb.org/vincheck

VINCheck 是美国国家保险犯罪局提供的在线工具。该工具有助于确定车辆是否被报告为被盗但未被追回。

（十一）用户名检查

1、CheckUserNames

网站网址：https://checkusernames.com/

CheckUserNames 可检查 500 多个社交网络上用户名的可用性。

2、Namech_k

网站网址：https://namechk.com/

此网站可搜索域名以查看可用的内容，然后进行注册或报价的过程。此外，它将检查数百个社交媒体网站上是否有用户名。甚至可以为他们注册。

（十二）病毒扫描仪

VirusTotal

网站网址：https://www.virustotal.com/gui/home/upload

VirusTotal 可扫描文件或 URL 以查看其是否具有恶意软件。

（十三）视觉/集群搜索引擎

这些搜索引擎通过对结果进行分类和组织，可以帮助用户缩小特定区域的范围。而且，其中一些允许用户使用可视化工具以不同方式分析数据和信息。

1、all-io

网站网址：https://all-io.net/

all-io 搜索多个源并提供一组结果，类似于任何搜索引擎。

2、alltheinternet

https://www.alltheinternet.com/

（十四）网站分析

BuiltWith.com

网站网址：https://builtwith.com/

可搜索收集技术的详细信息。

（十五）沃德舆情大数据平台

网站网址：www.wodeyuqing.com

该网站是国内外舆情搜索引擎，可对网上热点事件进行追根溯源和图表文字等多种形式的详细分析。

13.7. 参考文献

[1] 丁波涛. 国外开源情报工作的发展与我国的对策研究[J]. 情报资料工作, 2011(6):4.

[2] 董尹, 赵小康. 公开源情报研究综述[J]. 2021(2013-1):119-123.

[3] 范昊,郑小川.国内外开源情报研究综述[J].情报理论与实践,2021,44(10):185-

192+201.

[4] 邓胜利,王子叶,杨璐伊.美国开源情报的产生与发展[J].保密工作,2020(04):50-51.

[5] Steele R D. Open source intelligence[M]//Handbook of intelligence studies. Routledge, 2007: 147-165.

[6] OSINT Market & Technologies - 2017-2026[R], Homeland Security Research Corporation, 2017.9

第十四章　自然语言生成与智能写作研究进展、现状及趋势

14.1. 研究背景与意义

自然语言生成(Natural Language Generation, NLG)是自然语言处理领域的重要分支，它是指从给定输入信息（或者没有输入信息）生成满足特定约束条件的人类可读的自然语言文字的过程。传统的自然语言生成一般分为六个独立的模块：内容确定(Content Determination)、文本结构规划(Text Structuring)、句子聚合(Sentence Aggregation)、词汇化(Lexicalisation)、参考表达式生成(Referring Expression Generation)和语言实现(Linguistic Realisation)[1]。近年来由于数据的积累和算力的进步，基于编码器-解码器框架的神经网络语言生成模型得以快速发展。神经网络以数据驱动的方式快速建立输入与输出之间的联系，比传统生成方法具有更简单的结构和更好的泛化性，在性能上显著地超越了传统生成模型。基于现代深度学习框架的自然语言生成模型在许多应用中取得了长足的进步，如机器翻译、对话系统、文本摘要、故事生成等。

自然语言生成的关键问题是可控性，指模型在给定输入条件下应该生成符合预期的文本，这些文本在语法、用词、结构等方面应符合人类语言的规范或者给定的约束条件。在传统模块化生成框架中，基于规则的方法往往能生成稳定可靠的文本，但是缺乏多样性、特异性和泛化性。基于神经网络的端到端方法，从模型估计的概率分布中取词或采样难以精确预测和控制，因此生成的文本常会出现语法错误、语句重复、连贯性差、前后矛盾、违背常识、对输入或约束条件的忠实度差等问题。

目前，自然语言生成的主流框架包括以循环神经网络、Transformer、变分自编码器、生成式对抗网络为代表的自回归生成模型，以及在机器翻译领域崭露头角的非自回归语言生成模型。近年来大规模预训练模型快速发展，GPT2[7]、GPT3[8]、BART[9]等预训练模型在文本摘要等多种生成任务上显著优于非预训练模型。但是对于开放端文本生成任务，如对话生成、故事生成等，输入信息十分有限，预训

练模型的生成结果依然会面临逻辑性和连贯性较差、缺乏常识等问题。为了实现更好的可控性，语言生成需要通过内容规划(planning)来达到多种约束条件下的的从语义到文本的构建过程。因此内容规划在语言生成，尤其是长文本生成中有广泛的应用。其次，由于语言具有高维和稀疏的特性，引入外部常识库或事实知识库等可以进一步指导模型生成，融合知识的语言模型也有广泛的研究。最后，语言生成性能的提高也十分依赖于高质量的数据和可靠的评价指标，能够自动评价语言生成的连贯性、逻辑性、事实性等属性能够为生成模型提供更加显式和直接的指导。这些方向值得研究者继续挖掘和探索。

14.2. 领域发展现状与关键科学问题

自然语言生成领域涵盖众多文本生成任务，其中绝大多数任务为有条件文本生成，即根据特定的输入信息（例如数据表格、文本、图像等）生成相应的自然语言文本，要求生成结果自然流畅、与输入信息相关且一致。近几年自然语言生成领域得到显著的发展和广泛的关注，从一个小众领域变为一个热门领域，NLP 重要国际会议上文本生成领域的投稿量已经名列前茅。这主要得益于两方面的推动：一是深度学习技术的发展；二是文本生成相关产业需求的爆发。

深度学习技术给文本生成带来了极大的便利，我们可将有条件文本生成任务看作是从输入到输出的转换问题（大多数情况下是一个序列到序列的转换问题），并通过端到端的深度学习模型加以解决，这类模型通常基于编码器-解码器框架，可采用 RNN、LSTM、Transformer 等模型加以实现。编码器用于理解输入信息，并得到输入信息的表征，解码器则用于根据输入信息的表征以及已生成的文本进行后续文本的预测。在具有充足的任务相关训练数据的条件下，上述模型能够取得不错的生成效果。为了进一步增强编码器、解码器的能力，业界推出多种大规模预训练语言模型，例如 BERT、GPT、BART、T5 等，这些模型通过自监督任务在大规模生语料上进行预训练，能够学习获得通用高效的编码能力和解码能力，将这些模型在任务相关训练数据上进行重新训练之后，通常能取得任务相关文本生成性能的显著提升。这种预训练+微调的方式已经成为当前各类文本生成任务的最佳解决方案。此外，在文本生成过程中引入各类知识，包括语言学知识、常识、知识图谱等，也能进一步提升文本生成效果。各任务相关的发展现状以及技术介

绍将在后文进行展开，此处不再一一赘述。

尽管近年来自然语言生成领域进步显著，但仍面临如下关键难题：

一是评价问题：无论是否给定参考答案，文本的质量评价一直是一个难题，尤其是自动评价。如果我们能找到一个简单有效的自动评价指标或方法，那么文本生成任务就有可能跟下棋一样被 AI 彻底攻克。但遗憾的是，自然语言的多样性、歧义性以及进化性等特点使得我们难以找到这样一种自动评价指标。目前业界所广泛使用的 ROUGE、BLEU、PPL 等指标都只能近似反映文本质量的某个方面（例如内容覆盖性、流畅性等），很多情况下指标数值的提高并不意味着文本生成性能的有效提升。只能作为性能参考。尽管业界设计了更为复杂的模型驱动的评估方法，例如 BERTScore、MoverScore、BARTScore 等，这些方法在不同数据上的表现各有千秋，同时不具有可解释性。

二是质量可控性问题：尽管文本生成效果显著进步，但是深度学习模型所生成的文本质量仍存在一些重要缺陷，这些缺陷会极大阻碍文本生成技术的应用。概括来说，所生成文本在覆盖性、连贯性、一致性、多样性等方面会存在缺陷。覆盖性指所生成文本对输入信息中重要内容是否完全覆盖，尤其针对自动文摘、Data2Text 等任务。连贯性指所生成文本中多个语句之间的连贯程度，尤其是针对长文本生成任务。一致性指所生成文本的语义信息、事实信息是否与输入信息保持一致，以及是否与世界知识、常识保持一致，这个问题在几乎所有文本生成任务中都会发生，业界针对不同任务提出多类方法试图解决该问题，但总体效果并不理想。多样性则指所生成文本在语言表达上的多样性，即所生成文本是否与输入重复、是否与其他生成文本重复，尤其是针对文本复述、对话生成等任务。文本多样性可以简单通过在解码过程中引入随机性（比如 top-k，top-p 等）加以提升，但是这类方法通常会损害文本的其他方面的质量，因此难点在于如何在保持文本基本质量的前提下提升文本的多样性。

三是偏见问题：近期一些研究工作发现深度学习模型所生成的文本具有不同类型的偏见（bias），例如年龄偏见、性别偏见、种族偏见等。以性别偏见为例，模型倾向于生成女性在家带娃、男性外出工作的文本结果，这是对男女性别的偏见。尽管这种偏见的产生是由于训练数据所造成的，但是业界还是期望通过一些

技术手段消除偏见。

四是隐私安全问题：大规模预训练生成模型通常能取得比较好的生成效果，例如 T5 模型有百亿级参数，而 GPT-3 则有千亿级参数。GPT-3 在 570 G 语料上训练，具有“强大”的语言生成能力，包括写新闻、故事、对话、代码等。业界有一种观点认为，预训练模型基于大规模数据训练并且具有超大规模参数，所以本质上体现一种泛化的记忆能力。有研究工作表明，如果我们输入特定的上文，模型会输出一个文本片段，包括完整的用户隐私数据，例如电话、住址、邮箱等，这段文本是可以从训练数据中找到的。从这方面来说，模型记住了训练数据中的很多内容，容易导致隐私安全问题。

NLG 领域所面临的其他难题则包括深度学习模型所面临的通用性问题，比如：模型如何小型化，降低对存储和计算资源的需求；如何提高模型输出结果的可解释性；如何基于小样本进行模型训练；等等。

14.3. 领域关键技术进展及趋势

14.3.1. 生成式摘要

自动文本摘要，是指将长文档转化为简洁，包含原文重要信息，流畅的文本摘要的方法。文本摘要总体分为抽取式和生成式摘要，相较于抽取式文本摘要，生成式文本摘要具有生成新语句的特点，形成原文本重要信息的复述形式。

生成式文本摘要包含两个核心要素：重要信息抽取和语言生成。模型需要在理解原文语义的基础上，生成一段包含重要信息、且可读性强的摘要文本。高质量的摘要在重要性、简洁性、流畅性，与原文的事实一致性上都有良好的表现。

早期的生成式摘要发展缓慢，多依据词法及句法人为定义规则，再根据重要程度组成新文本摘要形式[66]。目前，基于神经网络的序列到序列模型被广泛应用，它将条件性输入序列转化为语言模型的输出序列，用编码器提取重要信息，用解码器进行语言建模，生成文本，同时也可适用于文本摘要领域。朴素的序列到序列模型经历了由 RNN[1]到 Transformer[2]的发展。近年来，大家将预

训练语言模型（PreSumm、BART、任务导向的 PEGASUS 等[3][6][7]）引入摘要任务，大幅度提升了生成式摘要的性能。

然而，这种数据驱动的文摘模型需要大量人工标注数据，很难在不同场景中推广。因此，我们需要探索如何进行少样本生成，乃至无监督条件下的摘要生成。Fabbri 等人采用数据增强的方法来获取和目标数据集性质相近的伪数据，显著提升了模型的 zero-shot 和 few-shot 表现[4]；Laban 等人则直接抛弃了标注语料，对文摘模型的输出进行评价来作为激励，采用强化学习的方法控制模型的行为[5]。Yu 等人提出了三种第二阶段预训练的方案，提升了模型在特定低资源领域的摘要能力[63]。Fu 等人设计了两种自监督的任务辅助模型完成 zero-shot 下的摘要[64]。Liu 等人利用去噪自编码器，通过两种指示符指导模型分别完成原文和摘要的重建，控制模型完成了 zero-shot 场景下的摘要任务[65]。近期，Li 等人则把基于提示微调（prompt-tuning）的方法应用到生成式摘要中来，在低资源情景下取得了超越 model-tuning 的结果[8]。

针对生成摘要的评估是另一个难点，它涉及摘要的信息度、重要性、简洁性、流畅性以及与原文的事实一致性。传统的评价指标以 N 元匹配方法为主（ROUGE, METEOR, BERTScore 等[9][10][11]）。由于语言的多样性以及摘要内容的主观性等特点，这些方法都无法准确地评价上述指标。所以，通常我们需要人工评测来弥补其缺陷。但是，这样也存在着人工成本、主观偏差等局限性。因此，现阶段仍然不断有新的自动评价方法被提出。比如，在衡量流畅性时，常用 GPT[13][14]的困惑度作为评价指标；为了评价事实一致性，Wang 等人提出使用自动问答系统对原文档和摘要文本同时评测，若他们针对同一个问题给出的答案一致，则结果具有较好的事实一致性[34][35]。另外，Zhou 等人提出“幻觉”检测机制，也可以被用来衡量文本摘要的事实一致性[15]。

“幻觉”（hallucination）问题是基于序列到序列文本生成范式存在的普遍问题，同样也存在于文本摘要任务中[33]。即模型会生成与原文事实不一致的文本，这是由于模型过于维护语言模型的流畅度，而忽略了原文的事实。研究者们提出了诸如增强数据方法、对比学习、裁剪损失函数等方法[36][37][38]，来提高摘要模型的事实一致性。除此之外，有些研究者提出了利用原文本中的结

构信息引导摘要生成提高事实一致性的工作。Jiang 等人提出使用结构性的张量积表示将依存句法等信息加入到模型中[29]；Dou 等人则提出了利用源文本中的关键词等信息引导摘要生成，并提出了统一的基于引导的生成式框架[30]。除了可以增加事实一致性，在对话摘要任务中，结构信息常被显示地建模以提升模型对数据的理解[18][19][20]。

由于深度学习框架下的文本生成过程多为“黑盒”行为，导致通过主观需求（如：具体的查询词、属性、事件等）调整摘要输出变得十分困难，降低了生成式文本摘要的应用价值。因此，文本摘要的可解释和可控性成为了重要的研究课题。为此，Wang 等人提出通过可解释性矩阵对摘要生成过程中不同属性(信息度、相关度、新颖度等)建模，通过该矩阵控制生成摘要内容[16]。摘要的抽象程度也是一个重要指标，它衡量模型重新组织原文信息的能力。实践中，可以通过控制摘要长度的方法，对摘要的抽象程度进行建模和控制[25][26][27][28]。另外，Amplayo 等人提出根据不同方面控制对同一段文本进行摘要的方法[62]。

与此同时，根据不同应用场景摘要系统衍生出了许多子任务。跨语言摘要是指将源文本转换为另一种语言的摘要技术[39]。同时，它也继承了机器翻译和单语言文本摘要的难点。现有方法集中于解决平行数据的匮乏问题，大多数方法通过构造伪平行数据辅助模型训练[39][42]，Ladhak 等人通过爬取 WikiHow 多语数据获得了较好的平行语料[44]。基于这些数据，更多研究则重点关注挖掘摘要和翻译之间的关系，从而提升了跨语言摘要任务的性能[40-48]。相较而言，多语 言摘要则聚焦于用一个模型同时实现多个语言的单语言摘要，该任务的难点在于如何共享多个语言共有的摘要知识，当前该任务已有多个不同的数据集被提出[52][53]。针对共享多语言知识的问题，目前研究者采用共享编码器、解码器的思路对其进行缓解[54]。

对话摘要技术旨在从复杂的对话中提取出重要的信息，从而帮助用户快速获取复杂对话信息。与普通摘要数据不同，对话产生于多位参与者，因此，存在大量的主题变化、频繁的指代以及多种不同的交互信号。而在多轮对话中，重要信息通常分布在不同的位置，数据具有较低的信息密度。这些特点为实现对话摘要带来了极大的挑战。Chen 和 Yang 提出了从对话中抽取主题和对话阶

段构造多视角摘要模型[17]。为了建模对话数据的结构，Zhao 等人利用细粒度的主题词构造对话图结构，并以此指导对话摘要生成[18]。类似地，Chen 和 Yang[19]和 Feng 等人[20]以构造对话图结构的方式提高摘要的质量[20]。为解决对话数据中存在大量指代的问题，Lei 等人[21]采用对话者自注意力构建了对话者与人称指代之间的复杂关系；Liu 等人[22] 利用指代消解进行后处理减少指代错误问题；Narayan 等人和 Wu 等人[23][24]选择利用草图和实体链的方法从粗到细的生成。

多模态摘要旨在处理文本、语言、图像、视频等多种模态信息，生成综合考虑多种模态信息后的核心内容。该技术的主要在于多个模态信息的融合和重组，Zhu 等人提出输入和输出均为多模态信息的任务形式，并提出使用多模态注意力机制、多模态目标优化等方法优化模型[31][32]。此外，多文档摘要[57][58][59]、时间线摘要[55][56]、邮件摘要[49][50][51]等也受到了不同程度的关注。

目前针对生成式摘要的数据集较为丰富，涵盖了新闻、专利、社交媒体、会议等多个领域，同时包括中文、英文等多种语言，并且抽象程度不同，具体如表 3-1 所示。

表 3-1 生成式摘要数据集

数据集名称	train/dev/test	平均输入长度	平均输出长度	领域	语种	说明
XSum	204,045/11,332/11,334	431.1	23.3	新闻	英语	
CNN/DailyMail	287,113/13,368/11,490	781	56	新闻	英语	
NY Times	589,284/32,736/32,739	800.0	45.5	新闻	英语	
NEWSROOM	995,041/105,760/105,759	658.6	26.7	新闻	英语	
Multi-News	44,972/5,622/5,622	2103.5	263.7	新闻	英语	多文档摘要
Gigaword	3.8m/189k/1,951	31.4	8.3	新闻	英语	
WikiHow	157,252/5,599/5,577			WikiHow 答案	英语	
WikiLingua	168k/ /	391	39	Wiki	18 种多语言	跨语言摘要/多语言摘要
LCSTS	2,400,591/10,666/1,106	106	18	社交媒体/新闻	中文	PART I/II/III

CNewSum	275,596/14,356/14,355	730.4	35.1	新闻	中文	
Reddit TIFU	79949 42984	342.4 432.6	9.3 23.0	社交媒体	英文	短摘要/长摘要
BIGPATENT	1,207,222/67,068/67,072	3573.2	116.5	专利	英文	
arXiv	215,913	6029.9	272.7	科技文档	英文	
PubMed	133,215	3049.0	202.4			
AESLC	14,436/1,960/1,906			电子邮件	英文	
BillSum	23,455	1813.0	207.7	法案	英文	
GOVREPORT	19,466	9409.4	553.4	政府报告	原文	
Webis-TLDR-17 Corpus	1.6m/2.4m (total: 4m)	202.99/382.7	22.21/33.6	社交媒体	英语	分为 comments 和 submissions 两部分，only training set
Webis-Snippet-20 Corpus	10,758,392/3,842/3,894 3,581,965/3,842/3,894	841	190	web page snippets	英语	分为普通 summarization 和 webpage triples（query-based）
Zh2EnSum	1,693,713/3,000/3,000	103.6	13.7	社交媒体/新闻	中文/英文	跨语言摘要
En2ZhSum	364,687/3,000/3,000	755.1	96.0	新闻	英文/英文	跨语言摘要
Yelp-Businesses	38,913/4,324	-	-	opinion	英文	
Yelp-Products	1,016,347/113,886	-	-	opinion	英文	
Amazon-Businesees	182,932/9,629	-	-	opinion	英文	
Amazon-Products	3,889,782/205,992	-	-	opinion	英文	
MSMO	293,965/10,355/10,261	720.9	70.1	新闻	英文	多模态摘要

14.3.2. 数据到文本生成

数据到文本的生成技术旨在根据给定的结构化数据生成相关文本，例如基于数值数据生成天气预报文本、体育新闻、财经报道、医疗报告等。数据到文本的生成技术具有较好的应用前景，目前该领域已经取得了较大的研究进展，业界已经研制出面向不同领域和应用的多个生成系统，例如“小南”写稿机器人[53]和 Arria NLG[53]。

利用自动化文本生成技术，对结构化数据进行分析，生成有价值的文本是自然语言生成领域的一项重要研究内容。该项任务面临的核心问题是如何挖掘数据集中的核心内容以及背后所蕴含的知识，从中选择出重要的信息，进而撰写出与输入事实相符的报道。

面向结构化数据的文本生成根据模型结构可划分为如下两大类：（1）流水线模型和（2）端到端模型。早期工作主要围绕流水线模型展开，将面向结构化数据的文本生成任务划分为四个子任务：信号分析子任务、数据理解子任务、文档规划子任务和微观规划和表层实现子任务,采用流水线形式按顺序逐个完成每个任务，前一任务的输出是下一任务的输入，这类模型通过流水线的方式主要解决以下两个问题：(1)写什么，即如何选取数据中的核心内容进行描述；（2）怎么写，如何进行表层实现(Surface Realization)，将选取的内容转化为符合预期的文本。这类方法普遍存在级联错误的问题。随着具有强大表示能力的深度学习的普及和大规模语料的出现，端到端模型成为近期面向结构化数据的文本生成工作采用的主流方法，该类方法通过数据驱动的联合训练方式，缓解了流水线模型存在的级联错误问题。

基于端到端的文本生成模型能生成高质量的文本的关键之一是具有规模较大的高质量数据集。近年来多个相关语料库的出现极大促进了数据到文本的生成领域的发展，表 3-2 列出了该领域最热门的语料统计情况。常用的数据集包括天气预报生成（WeatherGov[66]）、机器人足球比赛报道生成（ROBOCUP[67]）、维基百科人物描述生成（Wikibio[68]）、餐馆描述生成（E2E[69]）、生物领域知识库描述生成（KBGen[70]）和开放域知识库描述生成（WebNLG[71]）等。近年来，出现了 RotoWire[72]、MLB[73]和 WIKITABLET[74]三个长文本生成数据集，它们的平均

文本长度分别为337.1、542.1和115.9，显著长于前述的数据集，同时输入的结构化数据数量也显著多于前述数据集，对端到端模型的内容选择和表层实现能力提出了较大的挑战。ToTTo[75]和DART[76]专注于端到端模型的生成文本正确性，通过严格的人工质量控制方式，改写数据集中存在的一些错误文本，减少数据集的噪声。TWT[77]数据集提供了一种可控数据到文本生成任务，给定开头，要求模型根据表格中的内容进行续写。LogicNLG[78]和Logic2text[79]专注于探索需要进行逻辑推理的场景，探索预训练语言模型进行逻辑推理的能力，研究的难度逐步增大。现有工作主要从模型的内容选择（写什么？）和表层实现（怎么写？）进行优化。近期，有较多工作关注在低资源的场景下，如何训练端到端的数据到文本生成模型。还有一类工作进一步探索如何在端到端模型中融入逻辑推理的能力。接下来将从以上四个方面对近几年的工作进行介绍。

表3-2 数据到文本生成的语料库统计概况

数据集	词表	实例	表格大小	平均长度
WeatherGov	394	22.1K	191	28.7
ROBOCUP	409	1.9K	2.2	5.7
WIKIBIO	400.0K	728.0K	19.7	26.1
E2E	-	50.6K	5.4	20.1
WebNLG	-	13.3K	2.6	24.4
ROTOWIRE	11.3K	4.9K	628.0	337.1
MLB	38.9K	26.3K	565.0	542.1
WIKITABLET	1.9M	1.5M	51.9	115.9
ToTTo	136.8K	136.0K	32.7	17.4

DART	33.2K	82.2K	-	21.6
TWT	-	27.0K	32.8	15.8
LogicNLG	122.0K	37.0K	13.5	14.2
Logic2Text	14.0K	10.7K	-	16.8

许多数据到文本生成任务输入大量的结构化数据，要求模型从中定位关键的信息进行报道，这就需要端到端模型具有识别重要信息的能力。Mei 等人[80]针对天气预报领域，提出了一个结构化数据预选择器，在编码器端判断各项结构化数据的重要性，显著提升了模型的内容选择能力。Puduppully 等人[81]提出了一套两步生成框架，第一步是内容选择和规划，生成一个仅包含关键信息的结构化数据序列，第二步根据该序列生成对应文本。Gong[82]等人从数值数据理解和数据验证两个方面提升模型的内容选择能力。Su[83]等人采用两步生成的策略，首先对输入数据中的属性信息进行规划，然后利用预训练语言模型生成文本。Bai[84]等人在规划的时候引入对于属性信息间组合关系的树状建模方式。

在提升模型的表层实现能力，提升生成文本正确性方面，一系列工作从不同的角度进行改进。在如何更好的建模表格数据编码器方面，Liu 等人[85]提出了一个引入属性信息的编码器，在建模结构化数据值的信息的同时考虑了对应的属性信息。Gong 等人[86]提出了针对表格行、列和时间的多维度信息进行结构化数据建模的方法。Distiawan 等人[87]针对开放域知识库描述领域，修改循环神经网络内部连接结构，提出一个基于图的三元组编码器，同时建模三元组内部元素和三元组之间的关系。Zhao 等人[88]利用图神经网络建模三元组之间的关系。Xing 等人[89]针对表格的结构化信息，引入多个预训练目标，提升预训练模型生成文本的质量。Li 等人[90]在学习编码器表示的时候引入多个辅助训练任务，提升文本质量。Nie 等人[91]引入了预先计算好的操作符和对应计算的结果，帮助提升文本的保真度。

在如何让解码器更好的生成文本方面，Jain 等人[92]提出了一个静态和动态注意力机制相结合的混合注意力机制。Qin 等人[93]探索采用隐半马尔可夫模型

建模结构化数据和文本之间的语义关系。Sha 等人[94]针对维基百科描述领域，提出了混合注意力机制，在生成文本的时候考虑了结构化数据之间的相对顺序。Bao 等人[95]引入了考虑全局信息和局部信息的混合注意力机制。Shen 等人[96]在解码器段提出分块生成策略，提升文本的保真度。Song 等人[97]利用多任务学习框架，引入多个自编码损失函数，引导模型生成高保真度的文本。Li 等人[98]提出两步生成框架，首先自动生成模版，然后进行填槽，通过拷贝机制填充关键信息。Li 等人[99]通过检索的方式引入原型文本，允许模型参考原型文本的内容生成，提高文本流畅性和保真度。Wang[100]等人通过先生成梗概，再生成文本的方式，结合非自回归生成模型提升文本的正确性。Liu[101]等人在先规划后生成的两步生成模型中融入实体信息。Puduppully 等人[102]提出在生成过程中对实体信息进行跟踪的模型。Ghosh[103]等人探索逆向强化学习在提升文本正确性方面的效果。

有一类工作尝试控制生成文本的语言风格，Iso 等人[104]在生成的时候融入了作者的信息。Feng 等人[105]探索根据指定结构化数据，遵循用户指定的文本的语言风格生成对应文本。

在评价指标方面，Dhingra[106]等人提出了 PARENT 评价指标，在评价生成文本的时候不仅与参考文本进行对比，同时还与输入的结构化数据中的信息进行对比。Rebuffel[107]等人通过自动问答的形式提出了一个不需要参考文本的评价指标 Data-QuestEval。Faille[108]等人围绕表格中的实体信息，提出了一种衡量文本正确性的指标。

一系列工作探索在缺少大规模训练数据的情况下，如何训练高质量的文本生成模型。Ma 等人[109]探索低资源场景下采用数据增强的方法缓解训练数据不足导致的欠训练问题。Chen 等人[110]探索低资源场景下利用拷贝机制和预训练语言模型在预训练阶段学习到的语言知识缓解训练数据不足导致的问题。Gong 等人[111]通过多任务学习的方式，提升低资源场景下预训练语言模型的性能。Su 等人[112]根据输入的表格检索相关的原型文本，在生成的过程中提供参考。Zhao 等人[113]通过引入记忆机制帮助模型在低资源场景下提升生成文本的正确性。Perez-Beltrachini 等人[114]利用多示例学习方法探索结构化数据和文本分别

来自不同领域的情况下没有高质量对齐语料的问题。Fu 等人[115]进一步探索在只有文本的情况下，如何自动的构造数据到文本生成数据集。

为了提升端到端模型的推理能力，Suadaa 等人[116]通过引入针对推理设计的模版，结合拷贝机制帮助预训练模型提升推理能力。Chen 等人[117]引入变分推断，帮助提升模型的推理能力。

14.3.3. 复述生成

复述是使用不同词汇和语法结构来表达基本语义相同的文本[118][119][120]。文本复述生成技术指根据给定文本生成其对应的复述文本。复述生成技术是自然语言生成领域的一项重要的问题，并且具有极强的应用场景[118][120][121][122]。文本复述生成技术已经被广泛使用于提升问答系统[123][124]，机器翻译[125][126]和语义结构提取[127][128]等模型的性能，并且也是自然语言处理领域一种很好的数据增强的方法[129][130]。

早期的文本复述生成通常有基于规则的复述生成[131]、基于词典的复述生成[132][133]，基于语法的复述生成[134]和基于统计机器翻译的复述生成[135][136]等方法，也有通过回翻技术来生成文本复述的研究[137][138]。近年来随着深度学习技术的发展和多个文本复述语料库的出现，文本复述生成技术逐渐向深度学习模型过渡，对文本复述模型的研究的重点也逐渐从复述的流畅性、相关性转向了多样性复述生成，词汇级复述生成，语法级复述生成和多粒度复述生成等方面。表 3-3 列出了该领域的语料统计情况。

表 3-3 文本复述语料统计数据

数据集	题材	大小	平均长度
PPDB	短语、单词	220,000,000	2.85
WikiAnswer	问题	18,000,000	11.43
MSCOCO	图片描述	493,186	10.48

Quora	问题	404,289	11.14
Twitter URL	Twitter	2,869,657	14.80
ParaNMT	小说、法律	51,409,585	12.94

多样性复述生成的研究主要关注对于给定的文本生成多样化的复述。近年来有许多关于多样性复述生成的研究，显著地提升了生成复述的多样性。复述多样性的研究主要集中在两个方面，一是生成多条复述之间的多样性，代表性工作包括：Gupta 等人提出一种基于变分自动编码机的文本复述模型[122]，该模型通过对隐空间多次采样来生成同一文本的不同复述；Qian 等人在对抗生成网络中加入了两个判别器和多个生成器来生成多条不同的复述[139]；Cao 等人在训练生成对抗网络时最大化生成文本和其对应的隐空间编码的距离，来使得生成器更多的关注隐空间编码，并生成多条不同的复述[140]。二是生成复述和原文本之间的多样性，代表工作包括：Xu 等人通过在解码器中引入多个复述模式向量来指导解码器生成多样化的文本复述[141]；Kumar 等人对复述任务设计了新的子模块，并最小化子模块的目标函数以平衡生成复述的多样性和语义相关性[142]；Chen 等人指出变分自动编码机的隐变量会被语义相关信息污染，提出利用对抗学习来确保的隐变量的语义无关性，并引入判别器来提高词汇级和句子级的语义相关性[143]；Zhou 等人定义了关于词汇变化和多样性的奖励函数，并使用强化学习来指导复述模型的训练[144]；Lin 等人提出采用多轮生成的方法来提升复述的多样性，并在每一轮生成中采用回翻来保证复述文本和原文本的语义相似度[145]。

词汇级复述生成的研究主要关注于通过词汇的替换来生成复述，词汇级复述生成的代表工作包括：Cao 等人通过 IBM 模型获取同义词映射，并提出一种从原文本中复制并从限定同义词库中生成的受限解码器模型，来生成单词粒度替换的复述文本[146]；Lin 等人利用 WordNet 检索同义词来指导生成复述中的同义词替换[147]；Ma 等人提出一种通过学习获取同义词映射的新的模型结构[148]；Fu 等人提出将词袋作为隐变量的变分自动编码机模型，在复述生成过程中提供了语义相近的候选单词[119]。

词汇级复述的多样性和质量会受到限制，因此也有工作关注于语法级复述生成的研究。其中 Iyyer、Chen、Goyal 和 Kumar 等人的工作[149][150][151][152]通过引入额外的句法信息来显式地控制复述文本的语法结构，具有很好的可解释性。而 Chen 等人的提出的模型通过隐空间自动学习语法表示，在生成过程中对隐空间进行采样也能够隐式地控制生成复述的语法结构[143]。

近期，也有不少工作同时关注词汇级和语法级的复述，在多粒度上对原文本进行改写。Li 等人将原文本分为语法级模式和词汇级模式，并将其分别编码使得生成的复述有更好的可解释性和可控性[120]；Huang 等人提出通过外部词典引导编辑网络对原文本进行词汇替换和语法改写来生成复述[153]；Kazemnejad 等人提出一种基于编辑的复述生成模型，该模型从大规模的外部语料库中抽取和原文本相近的语料指导编辑器对原文本进行修改来生成复述[154]。

除此以外，还有许多关于文本复述生成的研究工作，如：Lin 等人分析了句子级复述和篇章级复述的区别，并提出了一个基于句子改写和重排序的篇章级文本复述模型[155]；Liu 等人提出一种基于模拟退火算法训练的无监督的文本复述模型[156]；Mallinson 等人利用神经翻译模型重新研究了通过回翻来生成复述的方法[157]；Cai 等人发现回翻模型也适用于 Text-AMR 模式，即原文本先解析成 AMR，再通过 AMR 生成原文本的复述[158]。

近年来随着文本复述生成技术的发展，复述文本的质量也越来越高，越来越多自然语言处理的下游任务通过结合文本复述使得模型性能有了显著的提升，复述生成技术的实用价值也在不断地提高。目前文本复述领域仍然存在许多亟待解决的问题，包括篇章级复述的生成、平衡多样性和相关性的复述评价指标的建立和零样本复述生成等。总之，文本复述仍然是一个具有很高研究价值的领域。

14.3.4. 对话生成

人机对话是人工智能领域的一项前沿性研究，打造可以和用户进行自由对话的聊天机器人是人工智能和自然语言处理领域之中的一个长远目标[159]。而生成式对话是解决这一问题的一条潜在路径。其核心方法是跟据已知的上下文信息，对对话中的语义信息进行建模，获得一个或者多个稠密向量表示，进而通过循环

神经网络等方式将向量表示转化为生成结果。本文首先概括生成式对话的模型框架，然后总结近一两年的发展现状，并探讨潜在的研究主题。

近些年来，人机对话的广泛的应用场景吸引了众多的研究者，伴随着他们的努力，产生了众多的模型和方法。现有的对话模型可以大体上归类于两大范式：检索式对话和生成式对话。尽管检索式对话可以保证回复的语法正确性，但是限于候选集的大小，检索式对话所产生的答复是可枚举的，缺乏必要的丰富性，也无法产生一个新的、不在候选集之中的答复。相反，生成式对话的假设空间是无穷大的，因此有可能产生丰富和多样化的回复，但是回复的语法质量是难以得到保证的。本文主要探讨生成式对话的一般模型和框架，并对起近几年的重点发展和突破进行系统性的梳理。

形式化地来说，生成式对话的任务可以表述为训练得到一个概率模型$P(R|C,S)$,跟据已知的上下文对话语境C，以及一些额外的外部信息S，来生成一个对于此对话的答复R。借鉴于早期机器翻译领域的方法和工作，众多生成式对话模型采用的是经典的 encoder-decoder 模型方法和框架。外部化信息的形式非常多样，可能是结构化的知识图谱，可能是非结构化的文档，或者是图片、视频等视觉和音频信息，这些外部资源给模型提供了必要的外部知识，方式模型生成的答复过于一般和乏味。

对于给定的上下文对话语境C以及其他的外部信息和知识S，编码器首先将其转化为一个隐状态向量表示，此后，在解码端，解码器依赖初始的隐状态条件s_0，在解码的过程中的每一个时间步，通过注意力机制不断地更新当前的隐变量状态，并将隐变量映射为一个维数为$|V|$的高维向量y_t，这种方式称为自回归（auto-aggressive），也就是说，每一个词的生成都依赖于之前生成的词。尽管也存在非自回归的生成方法，但是他们就生成效果而言，相比于自回归的方式有所逊色。

基于对比学习和连续学习的方法

最初的对话生成是通常是基于模板的匹配[160]，后来随着神经网络和深度学习技术的发展，encoder-decoder 框架逐渐开始盛行。在这个时期，许多模型通过循环神经网络来实现编码器和解码器[161][162][163]。

近些年来，除了使用传统的神经网络进行 encoder-decoder 结构，许多方法也在探讨如何将新的学习方式和学习方法迁移并应用到对话生成的场景中去。其中一个对于对话生成有借鉴意义的学习范式就是对比学习[164][165]。比如，Gunel[166]等人提出将有监督的对比学习融入到训练目标之中，以此来解决常规的交叉熵损失函数泛化能力不足的问题。进一步，Lee[167]等人提出，简单地通过随机采样作为对比样例并不合适，并提出了通过对正样例和负样例分别施加不同程度的数据干扰来来获得困难的学习材料。

除了对比学习之外，连续学习也是一个重要的学习方法。由于对话系统应该随着用户的使用而不断优化更新升级，而不是在训练完成之后在用户的使用过程之中始终保持恒定。鉴于此，Mi[168]等人提出将连续学习的概念应用到对话生成之中，通过合理地选择具有代表性的样例进行反复训练，可以防止灾难遗忘现象的发生。

基于大规模预训练模型的方法

近些年来，随着 transformer 结构的提出以及大规模预训练模型的风靡，BERT[169]，GPT-2[170]，BART[171]等大规模预训练模型在许多下游的文本生成任务上取得了令人瞩目的成绩，越来越多的方法模型已经预训练好的大规模模型直接在目标任务上进行微调，而不是从头开始训练。

由于已有的预训练模型一般都只单独适用于自然语言理解或者是自然语言生成，但却没有一个统一的模型能够将这两者结合起来。为此，Dong[172]等人提出了使用 UniLM，将自然语言理解和自然语言生成任务融入到同一个框架之中，使用不同的注意力机制进行多任务训练。而 DialoGPT[173]则提出使用专门的来自于 reddit 的对话语料进行预训练，以此来提高对于对话生成任务的效果。Plato[174]借鉴了 UniLM 的思想，同时使用回复生成、词袋模型等多种训练目标进行多阶段优化，同时使用隐变量来标识不同的对话行为，以此来模拟对话中的一对多现象。

但是，对于大规模预训练模型而言，总是存在一个模型所能容纳的最长的对话序列和外部知识的上界，因此，如何裁剪和选择外部资源以及如何进一步优化预训练模型，成为一个重要的问题。为了，Kim[175]等人提出了 SKT，将知识选择

的过程进行序列化，提高知识选择的准确率；dukenet[176]则进一步将机器翻译领域之中的 dual learning 概念应该用到知识选择之中，以此优化序列化的知识选择和建模过程。而 Zhao 等人[177]提出了一种通过序列化的选择模型来有效裁剪外部知识的方法，并使用强化学习将知识选择与回复生成放在同一个框架之中进行联合训练。

基于前缀提示的方法

然而，大规模预训练模型的缺点也是显而易见的。由于模型规模过大，在具体的下游任务上进行微调时需要大量的数据，否则会导致模型过拟合。巨大的参数量也给模型的部署带来了很大的困难。基于此，近几年来，基于前缀的生成方法开始受到学术界和工业界的普遍关注[178]。

Zheng[179]提出了如何使用前缀进行基于外部知识的对话生成，使用特定设计的前缀来抽取模型之中的相关知识，并在低资源的情况下有良好的效果。但是，prompt 的设计往往是一个难点，prompt 的细小差别可能会导致模型效果的巨大变化。为此，Li[180]等人提出了 auto-prompt，也就是使用由模型计算出来的 prompt 来代替经验设计的 prompt，使得 prompt 具有更强的解释性。此外，Gu[181]等人还提出，prompt 的词向量嵌入不应当保持恒定，而应该收到对话的上下文语境的影响而发生改变。

由于对话本身就存在着一对多的不确定性，因此单纯依靠上下文语境无法完全确定将来的答复，如何利用其他的有效信息生成可信度更高的回答是一个长久的研究问题。从另一个角度来说，对话一对多特性的存在使得对话的自动评估成为一个重要的议题。此外，随着大规模预训练模型的普及和应用，如何减小微调阶段的代价，增加模型在低资源甚至是零资源场景下的表现，或许会是将来一个重要的研究方向。

14.3.5. 故事生成

故事生成任务是指：给定一个约束条件，要求生成一个连贯的满足约束的故事。其中约束条件可以是故事的题目、开头、结局等。尽管基于自监督学习的大规模预训练模型已经能够生成具有良好的局部连贯性的故事，但是这些模型仍然

难以在整个故事中规划全局连贯的事件序列，同时也常常倾向于生成简单通用的情节。故事生成是典型的开放端生成任务，输入仅仅提供了非常有限的信息。对于这类任务，GPT-2 等现有的模型在生成时仍然会出现很多严重的问题，包括语句重复、逻辑冲突、缺少上下文一致性和相关性等。尽管 GPT-2 也能够生成一些相关的概念，但从生成的故事全局来看缺乏连贯性和常识。目前提升故事连贯性和逻辑性的方法主要有三种：规划、高层次表示、知识增强。

规划：指由粗到细逐步完善故事的层次化生成方法，即从给定输入出发，规划出故事的某种中间表示，如故事的提示、（抽象的）事件序列、故事的骨架、摘要等，然后根据这种中间表示生成故事。例如，Fan 等人[185]提出了一个三阶段模型：第一阶段根据给定的主题规划事件序列，事件表示中的实体均为占位符；第二阶段根据事件序列生成故事；第三阶段为故事中的实体生成指称表达。三个阶段均使用序列到序列模型进行建模，并利用交叉熵进行监督训练，其中第一阶段的监督信号可以利用语义角色标注等工具从故事中自动抽取。然而，这类基于规划的方法面临着暴露偏差的问题，即模型在训练和生成时中间表示的分布存在差异，这种差异会严重影响模型性能。

高层次表示：现有的模型通常基于自回归的生成方法，通过优化词级别的最大似然来学习故事生成，这种方法能够很好地建模词级别的共现关系，因此能够生成局部连贯的文本，但是却难以实现在句子级和篇章级的连贯性。因此 Guan 等人[184]提出在语言生成模型中引入高层次的表示来提高故事生成的连贯性。他们在故事中每句话后插入两个特殊符号，并通过句子相似度预测和句子顺序判别两个预训练任务来分别学习句子级别和篇章级别的表示，实验表明该模型生成的故事在上下文相关性和时序关系上能够取得更好的表现。

知识增强：指通过使用外部的常识知识作为指导，根据有限的信息联想相关的常识知识，并处理上下文实体、事件之间的因果和时序关系，从而增强故事生成的连贯性。Guan 等人[186]提出使用生成式后训练的方法在故事生成模型中引入常识知识，该方法将外部知识通过模板等方法转化为自然语言语句以构成训练语料，然后使用预训练语言模型 GPT2 在语料上后训练，进而在故事语料上对模型进行微调，以增强模型对知识的理解和生成能力、生成更合理的故事。Xu 等人[187]

提出在故事生成过程中，从常识知识库中检索相关的知识三元组，将其编码如故事生成模型，来增强模型生成性能。然而，如何有效地利用常识知识在故事生成过程中动态地进行长距离的常识推理仍然是极具挑战的问题。

数据集对于故事生成模型的发展也至关重要，目前故事生成的相关工作大部分均依赖于 ROCStories[191]和 WritingPrompts[192]两个数据集。ROCStories 由众包构造的五句话日常故事组成，包含了丰富的因果时序关系和常识知识。WritingPrompts 包含百万个爬取的“提示-故事”对。最近有工作为建模更长的事件序列收集了几千字至几万字的故事数据，如 roleplayerguild[193]、PG-19[194]以及 storium[195]。Guan 等人[196]也为中文故事理解和生成的发展提供了一个评价基准 LOT，其中包含了几千个长度几百字的中文故事，然而更长的高质量中文故事语料仍然十分稀少。

随着大规模预训练模型的发展，故事生成的研究已经取得了显著的突破。但是实际应用如辅助人类写作还面临着一系列问题，如生成连贯性依然较差、缺乏常识知识、创造性较差等，这些问题一方面依赖于收集更多高质量的故事数据，另一方面更依赖于生成算法和模型的进一步发展。

14.3.6. 多模态生成

多模态生成是指利用文本、图像、视频、音频等多个模态的信息进行融合，学习不同模态之间的关联，生成富含多源信息的文本，例如根据图像进行诗歌故事的创作或者对话生成、通过音视频信息生成更有表现力的文本等。多模态学习通过模拟人在处理信息的时候利用自然语言、听觉和视觉等多个模态的信息，通常会以一个模态为主，其他模态为辅进行信息的补充和增强，因此具有广泛的应用。

传统的文本生成技术往往是利用数据到文本或者文本到文本的生成，即利用结构或者非结构化的数据对其内容进行提取，再通过既定的模板或者构造的方法进行文本生成，或是通过对源文本的改写和再创作转化为预期文本。近年来，随着深度学习的发展[198][199][200]，多数的文本生成方式通过神经网络对大量文本数据进行基于统计分布的学习，或者通过映射到高维隐式空间，添加适当的变换再投

影回词表空间进行生成。然而，上述的单模态生成方式仅仅局限于在文本空间进行变换和重组，缺乏其他模态可能包含的丰富的信息。通过对不同模态之间的信息进行表征、翻译、对齐、融合和共同学习[201]，机器能够在最大程度上模拟人整合不同感官信息和组织自然语言的方式，从而生成更真实的文本。

在自然语言生成中，多模态生成主要包含基于图像或视频的文本生成、基于音频的文本生成、以及多种模态混合的文本生成方法。根据任务的不同，常见的多模态文本生成任务又包括图像或视频描述生成、视觉诗歌或故事生成、多模态翻译、音频描述生成等。

对于描述生成任务，早期以检索为主流的方法被以卷积神经网络（CNN）为架构的编码器-以循环神经网络（RNN）为架构的解码器框架所替代后，取得了一定的成功。但基于 CNN-RNN 框架的方法容易出现目标缺失和错误预测的问题，Li 等人[202]提出了一种全局-局部注意力（GLA）方法将对象级的局部表示与图像级的全局表示结合起来，能够生成更多相关的句子。近年来，受 Transformer 模型在机器翻译中的启发，Yu 等人[203]引入多视角视觉特征并将其扩展为用于图片描述生成的多模态 Transformer（MT）模型，在当时 MSCOCO 图像标题挑战的实时排行榜上取得第一。Wang[206]等人 2018 年提出重构网络（RecNet），设计了两种类型的重构器，利用正向（视频到句子）和反向（句子到视频）流进行视频描述的生成。Aafaq[207]等人通过应用短傅立叶变换的视觉特征编码技术，在视觉特征中嵌入丰富的时间动态，并提供给一个语言模型，在 MSVD 和 MSR-VTT 数据集上取得了最好的结果。此外，一些数据集[204][205][208][209]的提出也丰富了相关领域的评价基准并促进了其发展。

另一类典型多模态生成任务是视觉诗歌或故事生成。比较早进行视觉诗歌创作的工作来自微软小冰[210]，类似工作多是通过图像进行关键词的提取并扩展[211][212]或利用图像的隐式信息[213]输入到 RNN 进行诗歌的生成。在 Liu[214]等人 2018 年的工作中，一种利用策略梯度的多对抗训练方式为多模态诗歌生成任务提供了新的思路。除了单张图片作为局限的输入信息，进一步，视觉诗歌生成任务又可以发展到多张图片[215]的图片流作为输入，通过利用设计的选择和注意力机制进行不同图像间信息的获取并利用到生成中。此外，一些工作研究了这类比

较自由的生成任务中的评估问题[216]，同时一些系统的上线也让这部分多模态生成更贴近工业和生活化[1][2]。视觉故事生成提出于 2016 年[217]，不同于诗歌生成，这类任务会利用更多的图像信息，并且往往在生成文本的连续性和逻辑上具有更高的要求。大致可以分为一类通过模型架构的优化来更好地生成故事[218][219]，或者是通过对抗训练[220][221][222]、知识图谱[223][224]等方法，来结合多模态信息进行故事生成，或来模拟用户进行故事情节的推导。

其他的生成任务如多模态翻译（MMT），是使用语言以及其内容相关的图片进行翻译，补充单用文字模态信息可能会产生的歧义[225]。自 Multi30K 数据集[226]发布以来，MMT 建模机制可以分为基于 RNN 序列模型或基于注意力的模型，近年来又有工作[227]通过 Masking 任务把翻译实体和图片中的实体联系起来增强效果。除此之外，跨模态音频视频到文本的生成任务也备受关注[228][229][230]。Lin 等人[231]今年提出了一个从包括视频加文本、语音或音频等输入的多模态生成文本的框架 VX2TEXT，在多个任务上取得最好的效果。

多模态生成通过利用其他模态的信息，能够对生成文本进行一定的补充，并通过模态间的交互和融合达到很好的水平，也符合人类在生成或者创作的常理。尽管这样，多模态生成的许多任务仍然存在数据稀缺和收集上的昂贵、模态间交互融合的解释性和调优方法等值得深入思考和讨论的地方。随着近年来多模态预训练模型的相继出现[232][233][234][235][236]，多模态生成具有非常合理的实际价值和令人期待的结果，一些数据集、模型、评估方法和对比讨论将会在很大程度上促进本领域的发展。

14.3.7. 代码生成

在本文中，代码（code）特指使用编程语言（programming language）撰写的、机器能够理解并执行的结构化内容。近年来，随着计算机软件领域的蓬勃发展，全世界代码开发者的数量呈不断上升趋势。如何提升代码开发者在开发周期中各个环节的效率和生产力，已经成为软件开发领域中的一个重要研究课题。另一方面，基于自监督学习（self-supervised learning）的预训练模型（pre-

[1] http://poem.xiaoice.com/
[2] http://jiuge.thunlp.org/

trained model）已经成为自然语言处理的主流方法和新范式。由于预训练技术本身所具备的通用性和领域无关性，该类方法同样能够训练基于代码数据的预训练模型，用以支持包括代码生成在内的各种下游任务。基于该背景，本小节将简要介绍基于预训练模型的两个代码生成典型场景：（1）代码-代码生成和（2）文本-代码生成。

代码-代码生成（code-to-code generation）是指基于已有代码，通过补全、翻译或精化的方式，生成新的代码。

- 代码补全（code completion）是指基于开发者已经输入的代码来预测并自动补全剩余代码的过程。该任务通常使用基于语言模型的预训练模型技术。例如，微软提出的 CodeGPT[237]模型和 OpenAI 提出的 Codex[238]模型都是基于语言模型预训练技术。该类模型采用自回归模型（auto-regressive model）预训练任务，基于已给定或已生成的上文去预测接下来每个位置的单词。这样训练得到的代码预训练模型可以直接用于代码补全任务。工作[239]在 CodeGPT 的基础上，进一步引入更多的全局上下文。实验证明，这种同时兼顾全局上下文和局部上文的方式能够有效缓解模型输入长度有限的问题，并在代码补全任务上取得更好的效果。为了缓解自回归生成机制所导致的语法错误问题以及代码补全中常见的歧义问题，Grammformer[246]提出一种自顶向下（top-down）的代码补全机制。通过不断实例化每一轮补全结果中的非终结符为终结符序列或同时包含终结符和非终结符的序列，该方法能够在尽可能保证输入代码语法正确性的同时，将模型觉得不确定的区域留给开发者做进一步补全，从而缓解模型在歧义位置常犯的过度预测错误。

- 代码翻译（code translation）和代码精化（code refinement）是指将一段已有代码在保持其功能不变的前提下，翻译到另外一种编程语言或进行（局部）改写和调优的过程。这两个任务通常使用基于编码器-解码器（encoder-decoder）的模型完成。例如，CodeBERT[240]和 GraphCodeBERT[241]两个工作首先基于单模态代码语料和双模态文本-代码语料训练预训练编码器。其中，CodeBERT 采用掩码语言模型的预训练

任务，GraphCodeBERT 在 CodeBERT 的基础上进一步引入基于代码 AST（abstract syntax tree）和数据流（data flow）的预训练任务。在得到代码预训练编码器后，通过在代码翻译和代码精化任务上精调该编码器和一个解码器，能够取得很好的效果。CodeT5[242]和 PLBART[243]则是使用若干生成任务直接预训练一个编码器-解码器模型。该模型同样能够在代码翻译和代码精化任务上取得很好的代码生成效果。

文本-代码生成（text-to-code generation）是指基于文本生成语义对应的代码。

- 基于文本的 SQL 生成（text-to-SQL generation）是指基于给定的结构化数据库或表格，将一段自然语言描述转化为语义对应 SQL 代码的过程。该类任务同样基于编码器-解码器预训练模型，但能够用于预训练或精调的文本-SQL 标注数据的数量依然非常有限。针对这一问题，研究者提出各种数据增强方法，用来加强预训练模型在 SQL 生成任务上的性能。例如，工作[244]基于表格自动生成 SQL 语句，并通过一个生成器生成 SQL 对应的自然语言描述。通过使用自动构造的自然语言描述-SQL 对作为额外的预训练预料，进一步提升了模型在该任务上的生成效果和泛化能力。
- 基于文本的代码生成（text-to-code generation）是指基于一段给定自然语言功能描述，生成具有该功能的代码的过程。和 SQL 生成相比，该类任务并不需要基于任何结构化数据库或表格。和代码翻译和精化任务类似，由于这个任务有明确的输入，因此可以基于编码器-解码器代码预训练模型[242][243]直接进行下游任务精调。

和很多其他研究领域相同，基准数据集对代码生成领域的研究同样起着至关重要的促进和推动作用。CodeXGLUE[237]是近年来提出的一个专门针对代码智能研究的通用数据集。该数据集由 10 个代码相关下游任务数据集组成，其中包括了代码完型填空、代码补全、代码精化、代码翻译和文本-代码生成等一系列生成类任务。目前，CodeXGLUE 已经成为代码智能研究领域中最受欢迎的一个基准数据集和评测平台。

受大规模预训练技术快速发展的影响，代码生成研究在近年来取得了长足的进步。在一些场景中，基于深度学习模型的代码生成系统已经落地于开发者日常

使用的编程环境中（例如 Visual Studio 和 VSCode 中的代码自动补全功能）。不过，该领域依然处在快速发展的新阶段，并面临一系列亟需解决的前沿研究课题，例如：（1）如何利用代码数据所特有的语法、知识和结构信息学习更好的代码预训练模型？（2）如何保证生成代码的语法正确性和可执行性？（3）如何解决不同编程语言之间的代码翻译以及文本和代码之间的相互转化问题？（4）如何对代码理解和生成任务中的超长输入和输出进行建模？（5）如何对训练数据中的代码进行知识产权和隐私保护？这些问题都有待研究者和开发者通过协作共同解决。

14.3.8. 资源和评价

自然语言生成数据资源

与一些较为简单的自然语言处理任务相比，自然语言生成任务对系统在语言的逻辑及推理方面有着更高的要求。如何让机器生成人类无法区分的自然语言文本，被认为是实现真正人工智能的一个重要问题。

近年来，随着深度学习的不断发展，自然语言生成及数对计算资源和数据资源的需求大幅增加。数据资源建设成为本领域系统性能提升的关键所在。近年来本领域数据资源建设发展迅速，很多开源数据集得到广泛使用，这不仅可以节约人力、物力，还可以基于此组织技术评测，对不同模型、方法的性能进行客观比较，有效推动技术的发展。

对于自然语言生成所包含的几个主要任务（摘要生成、故事生成、数据到文本生成、图片到文本生成、对话生成、问题回答），目前均已有相关的数据集发布。其中，摘要生成任务历史悠久，资源建设也较为成熟。目前，该任务的许多数据集基于新闻构建，因为新闻的标题往往可以视为整篇新闻的摘要。一个代表性资源是 IBM 构建的 CNN/DM 数据集。对于摘要生成任务而言，文章的长短也是影响系统性能的一个重要因素，长文本往往要比短文本更难处理，因此也有专门针对长文本构建的数据集如 WikiHow。同时，也有一些数据集针对特定类型的文本进行构建，如基于专利数据构建的 BIGPATENT。

故事生成任务要求在给定故事开头的情况下，系统需要对故事进行续写，该

任务对生成语言的逻辑要求更高。该任务只需要使用合理的故事对模型进行训练，因此可以使用其他任务的数据集，但应注意篇幅长度和语料领域。如 ROCStories（本用于选择故事的正确结尾）。但目前的数据集大多通过对写作网站进行爬虫获得资源，如 Writing Prompts，其单篇质量差异较大。

数据到文本生成任务是根据某些离散的数据生成自然语言的任务，如根据维基百科的元数据生成对该词条描述语句（数据集 WikiTableT）、根据 NBA 比赛的比赛队伍和得分等信息生成比赛简报（数据集 ROTOWIRE）、根据商品属性生成广告语（数据集 Advertising Text Generation）等。图片到文本生成任务可以进一步细分为看图说话、根据图片回答问题等任务。Flickr30k 是经典的看图说话数据集，而 Flickr30k Entities 在前者的基础上，添加了图片和句子中对应实体的关系标记，使得信息更加丰富。VQA 是根据图片回答问题任务的数据集，系统需要在输入图片和自然语言问题的前提下，用自然语言回答提出的问题。在构建的过程中，VQA 中的问题更注重逻辑推理，也因此有着较高的难度。

对话生成任务是对话机器人的核心，也是工业界目前较为重视的领域，在各类智能终端上有广泛应用。因而该任务也已发布了较为丰富的资源：多语言的 chatterbot、中文的 PTT、英文的 Cornell Movie Dialogs Corpus 等，可以满足不同语言的需求。问题回答任务与对话生成任务有一定的相似性，但输入被限定为一个有意义的问题，回答也被限定为特定答案或者无法解答。该任务可以分为限定域问题回答和开放域问题回答两种，前者需要在给定文章的前提下对问题进行回答（如数据集 DMQA），而后者往往需要预先构建知识库，利用知识库中的信息对任何可能的问题进行解答（如数据集 XQA）。

从整体上看，目前的主要自然语言生成任务都有多个数据集，且数据集之间存在一定差异，研究人员可以根据自身研究任务的特点进行选择。除此之外，由于一些自然语言生成任务的特点，它们可以使用其他任务的数据集进行训练。但与此同时，针对特定领域的数据集相对较少，现有数据集的质量也参差不齐。如何构建更加多元化、高质量的数据集是未来自然语言生成任务研究的一个重要方向。表 3-4 为目前主流开源数据集基本情况和获取方式概览。

表 3-4 主流开源数据集概览

名称	任务	规模	语言	作者/单位	时间	描述	网址	论文出处
CNN/DM	摘要生成	约 30 万篇文章	英语	IBM	2016	修改自 DMQA，使用新闻文章构建	https://s3.amazonaws.com/opennmt-models/Summary/cnndm.tar.gz	Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond
WikiHow	摘要生成	约 23 万篇文章	英语	Mahnaz Koupaee 等	2018	用于长文本摘要生成	https://ucsb.app.box.com/s/ap23l8gafpezf4tq3wapr6u8241zz358	WikiHow: A Large Scale Text Summarization Dataset
BIGPATENT	摘要生成	约 130 万篇	英语	Eva Sharma 等	2019	使用专利文献构建	https://evasharma.github.io/bigpatent/	BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization
ROCStories	故事生成	约 10 万篇	英语	Nasrin Mostafazadeh 等	2016	每一篇都是由五个句子构成的短文	https://cs.rochester.edu/nlp/rocstories/	A corpus and cloze evaluation for deeper understanding of commonsense stories
Writing	故事	约 30	英语	Facebook	2018	从 Reddit 的	https://www.kaggle.co	Hierarchical Neural Story

Prompts	生成	万篇				WritingPrompts 论坛上收集的创作故事	m/ratthachat/writing-prompts	Generation
WikiTableT	数据到文本生成	约 150 万个实例	英文	Mingda Chen 等	2021	基于维基百科，利用元数据生成描述文本	https://www.kaggle.com/mathurinache/wikitablet	WIKITABLET: A Large-Scale Data-to-Text Dataset for Generating Wikipedia Article Sections
ROTOWIRE	数据到文本生成	4853 个实例	英语	Sam Wiseman	2017	基于 NBA 篮球比赛播报，给出参赛队伍、比分等内容，生成比赛简报	https://github.com/harvardnlp/boxscore-data	Challenges in Data-to-Document Generation
Advertising Text Generation	数据到文本生成	约 11.9 万个实例	汉语	Zhihong Shao 等	2019	基于淘宝，给定商品的属性值，生成广告	https://drive.google.com/file/d/1vB0fT1ex2Tsid-i5s-jqdz9QUFbCh0CO/edit	Long and Diverse Text Generation with Planning-based Hierarchical Variational Model
Flickr30k	图片到文本生成	3 万张图片	英语	Peter Young 等	2014	经典的描述图片内容任务语料库	https://github.com/BryanPlummer/flickr30k_entities	From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions

Flickr30k Entities	图片到文本生成	3 万张图片	英语	Bryan A. Plummer 等	2015	基于 Flickr30k，图片及句子中对应实体使用标记进行关联	https://github.com/BryanPlummer/flickr30k_entities	Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models
VQA	图片到文本生成	约 20 万张图片	英语	Stanislaw Antol 等	2015	根据图片内容回答自然语言问题，更注重需要推理的问题	https://visualqa.org/	VQA: Visual Question Answering
chatterbot	对话生成	560 组	多语种	开源项目，作者未知	2019	单轮对话数据，按照不同类别划分对话领域，整体质量不错，但是语料相对较少	https://github.com/gunthercox/chatterbot-corpus/tree/master/chatterbot_corpus/data	-
PTT	对话生成	约 122 万组	中文	开源项目，作者未知	2019	单轮对话数据，PTT 网站上用户的提问与回答构成一组对话，比较生活化	https://github.com/zake7749/Gossiping-Chinese-Corpus	-
douban	对话生成	约 100	中文	北航&微软	2017	多轮对话数据，豆瓣用户的提问	https://github.com/MarkWuNLP/MultiTurnRe	Sequential Matching Network: A New Architechture for Multi-

		万组				与回答，噪音比较少	sponseSelection	turn Response Selection in Retrieval-based Chatbots.
tieba	对话生成	232 万组	中文	开源项目，作者未知	-	多轮对话数据，百度贴吧用户对话数据集	https://pan.baidu.com/s/1mUknfwy1nhSM7XzH8xi7gQ 密码:i4si	-
Cornell Movie Dialogs Corpus	对话生成	约 30 万组	英文	Cornell University	-	多轮对话数据，一个丰富的电影角色对话数据集，从电影台词中抽取出的若干组人物对话	http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html	-
DMQA	问题回答	约 30 万篇文章	英语	Hermann 等	2015	使用新闻文章构建，每篇文章大约对应 4 个问题	https://cs.nyu.edu/~kcho/DMQA/	Teaching Machines to Read and Comprehend
SQuAD	问题回答	536 篇文章,约 10 万个问题-答案对	英语	Stanford University	2016	基于维基百科，其中的问题可能是未解答的	https://rajpurkar.github.io/SQuAD-explorer/	SQuAD: 100,000+ Questions for Machine Comprehension of Text

NewsQA	问题回答	约 1 万篇文章，10 万个问题-答案对	英语	Adam Trischler 等	2016	基于新闻进行构建，设计问题时更关注需要进行推理解答的部分	https://github.com/Maluuba/newsqa	NewsQA: A Machine Comprehension Dataset
SimpleQuestions	问题回答	约 10 万个问答对	英语	Facebook	2015	问题较为简单，重点在于使用该语料研究多任务及迁移学习的影响	https://www.dropbox.com/s/tohrsllcfy7rch4/SimpleQuestions_v2.tgz	Large-scale Simple Question Answering with Memory Networks
WikiQA	问题回答	约 3050 个问答对	英语	Yi Yang 等	2015	基于维基百科构建	https://www.microsoft.com/en-us/download/details.aspx?id=52355	WikiQA: A Challenge Dataset for Open-Domain Question Answering
cMedQA	问题回答	约 5 万个问题，对应 10 万个回	汉语	Sheng Zhang 等	2016	医学问题的问答	https://github.com/zhangsheng93/cMedQA	Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs

		答						
XQA	问题回答	约 9 万个问答对	多语言	Jiahua Liu 等	2019	针对开放式问答构建的数据集，训练集是英文，而开发集、测试集是多种语言，旨在研究跨语言的 QA 问题	https://github.com/thunlp/XQA	XQA: A Cross-lingual Open-domain Question Answering Dataset

自然语言生成的评价

随着自然语言生成任务的迅速发展，对 NLG 系统性能的评价就成了当务之急。与分类任务不同，NLG 任务的判断结果更加主观，因而难度更大。对此显然存在人工评价和机器（自动）评价两种技术方案。

人工评测较为直观且易实施。向多位人类评测者这展示两个系统输出，并要求评测者对进行打分最后进行汇总排名。显然此方案对评测者有较高要求，且会导致评测周期较长，评价成本高昂，一致性亦无法保障。对于机器（自动）评价，实施较为简单的 BLEU[247]、METEOR[248]、ROUGE[249]等自动评价指标得到广泛使用。但是这些指标十分通用，并未考虑到不同 NLG 任务在语义上的不同要求[250]。

我们将分别介绍人工评价指标和机器自动评价指标的研究情况，并重点分析现有的自动评价指标的缺点，并简要分析针对 NLG 具体任务进行自动评价指标选择的方式。

人工评价指标

在进行人工评价时，通常应关注以下几个方面：

1. 评价者类型：评价者可能是专家、众包、最终用户。

2. 评价尺度：即对结果进行打分时的角度和评分等级。

3. 提供参考集和上下文：应将上下文和一些参考结果提供给评价者。这对提高评价一致性具有良好作用。

4. 绝对评价/相对评价：可以对目前系统结果进行单独的评价（绝对评价），也可以同时对多个竞争系统的结果进行评分或排序（相对平价）。

5. 给出理由：在评价时可要求评价人员提供打分或评级的理由，据此进一步改进系统。

无论使用何种评价方法，通常会为多个评价者提供相同的系统输出，对多人评价进行汇总后得出最终分数。可使用平均值或加权平均值来对结果进行汇总[251]。除此之外，需要评价者之间有着较高的一致性。

自动评价指标

自动评价指标分为上下文无关指标和上下文相关指标两大类。上下文无关度量在判

断结果性能时不考虑上下文，而只检查结果和给定参考集之间的相似性。这使得它们与任务无关，更容易广泛地用于多种 NLG 任务。上下文相关度量通常是针对特定任务提出的，用来判断结果的适当性，跨任务性能较差。

除了是否考虑上下文这个维度，文末还可以根据指标使用的技术分为两类：需要使用人工标注数据进行训练的训练指标(trained metrics)和不需要任何训练，只需使用一组公式直接进行计算的未训练指标(untrained metrics)。更细得，还可以根据它们是对单词、字符还是向量嵌入进行操作来进一步分类。

上下文无关指标

（1）未训练指标

基于单词的指标通常将系统结果和参考集视为 N 元文法的集合，然后根据系统结果与参考集之间的 N 元文法的重叠情况进行打分。还有一些指标根据使系统结果与参考集相似所需修改的单词数量来打分。大多数早期评价指标，如 BLEU、NIST、ROUGE、METEOR、GTM 等都是基于单词的指标。这些指标已被广泛用于许多 NLG 任务。

基于字符的计算指标主要有 characTER、EED、chrF 等。它们在使用时通常不需要对句子进行分词，而是直接将参考集与系统输出结果进行字符串对比。这类指标在评价形态更加丰富的语言时有优势。以 chrF[252]为例，它比较参考集和系统输出结果中的字符级 n-gram，而不是像 BLEU 那样匹配单词级的 n-gram。其计算公式如下所示：

$$chrF_\beta = (1+\beta^2)\frac{chrP.chrR}{\beta^2.chrP + chrR}$$

chrP 和表示系统结果中匹配到参考集的字符级 n-gram 的百分比，而 chrR 表示参考集中也存在于系统结果中的字符级 n-gram 的百分比。β为可调节的权值。

上述指标依赖于字符串的匹配，因此会忽略单词之间的语义相似性。例如，“猫”和“狗”这两个词虽然不是同义词，但比“狗”和“船”更接近（都是宠物）。这种相似性可以通过 Word2Vec、GloVe 等词嵌入方法捕获。这类指标通过比较系统结果和参考集中单词嵌入之间的相似性对系统性能进行评价。基于嵌入的指标有 WMD、MEANT、YiSi、BERTscore 等。

（2）训练指标

这类指标通常利用机器学习或深度学习模型将评价任务建模为从给定输入到评分的回归任务，这就要求使用的数据集需要包含一些人工评分结果作为拟合的目标。根据输入不同，训练指标又可以进一步分为两类：使用预先计算的启发式特征（n-gram 精确率、召回率等）的基于特征的方法，以及使用系统输出结果和参考句子作为直接输入进行训练的端到端方法。前者的代表有 BEER、BLEND 等，后者的代表有 BERT for MTE、SIMILE、ESIM、RUSE 等。

上下文相关指标

（1）未训练指标

基于单词的上下文相关指标通过使用系统结果和上下文的 n-gram 特征来对结果进行评价。以 ROUGE-C[253]为例，该指标针对摘要生成任务设计，计算公式如下：

$$\text{ROUGE-C-N} = \frac{\sum_{s_h \in \text{hypothesis}} \sum_{n\text{-gram} \in s_h} Count_{match}(n\text{-gram})}{\sum_{s_c \in \text{Source Document}} \sum_{n\text{-gram} \in s_c} Count(n\text{-gram})}$$

其中，s_h 和 s_c 分别是属于摘要和文档的句子。ROUGE-C 特别适合没有参考摘要的情况。

使用嵌入的上下文相关指标相对较少，YiSi-2[254]是一种使用跨语言嵌入来计算机器翻译系统输出结果与源语言输入相似性来对系统进行评价的指标。该方法无需参考集，使用多语言 BERT 提取的上下文嵌入来评价输入和输出之间的跨语言词汇语义相似度。

（2）训练指标

大多数上下文相关训练指标都是针对对话评价任务的。以 RoBERTa-eval[255]为例，该指标利用 RoBERTa 模型产生的上下文文本嵌入将对话上下文和系统结果编码为单个向量，然后使用以 sigmoid 为激活函数的多层感知机进行回归，最终输出范围为 1 到 5 的打分结果。这种方法可以在没有参考集的情况下对系统进行评价，但与之对应的，本指标需要对负采样及现有生成系统生成结果进行人工标注来训练回归模型。

自动评价指标的不足

目前的一些研究针对 BLEU、ROUGE、NIST 等广泛使用的自动评价指标进行分析，

发现自动评价指标有着以下不足：

与人类判断的相关性较差：有研究发现 BLEU、NIST 等几种方法的评分结果与人类对流畅度的判断呈负相关、与充分性分数相关性中等至较低，表现出了自动评价指标与人类判断的相关性较差[256][257]。

缺乏可解释性：如果自动评价指标打出低分，那么这个低分是对应于较差的流畅性、较差的信息量还是较差的连贯性？自动评价指标无法给出回答[258][259]。

指标倾向性强：指标在设计时本身会对某些情况给予“优待”。如果指标本身有所倾向，那么系统之间的性能对比结果很可能是不可靠的[260][261]。

跨任务适应性差：由于缺乏对应的语境语义信息，对任务的适应性存在问题[262]。

无法捕捉语言中的细微差别：例如，Kryscinski 等人[263]批评用于摘要生成的自动评价方法，认为这些方法没有检查出摘要中的事实不一致问题。

基于此在选择自动评价指标时，需要对自动评价指标有效性进行评价。最广泛使用的方法是计算自动评价指标给出的分数与人类判断之间的相关性。计算相关性可以使用皮尔逊相关系数、斯皮尔曼相关系数以及肯德尔的 τ 系数进行计算。对于相关性上的差异，通常还需要通过 William 检验来判断两个指标之间的差异是否显著。

综上所述，未来在 NLG 评价方面还有诸多工作需要进一步开展，包括且不限于以下内容：

（1）为自动评价指标开发公共代码库；

（2）构建包含人类判断的各类较大规模的数据集；

（3）开发用于特定任务的上下文相关指标以及相关数据集；

（4）开发更多可解释的指标，以便从多个角度给出自动评分的解释；

（5）创建用于评价自动评价指标的基准，如基本标准及其含义定义。

14.4. 领域产业发展现状及趋势

自然语言生成常见的应用场景包括内容创作、智能对话等，应用行业主要涉及媒体、金融、营销等。2017 到 2021 这五年期间，是自然语言生成技术逐渐成熟的阶段，在技术进展和国家政策的推动下，自然语言生成相关的产业规模迅速提升，大型互联网公司和创业公司持续加强投入，新技术加速落地。未来，随着生成技术在可控、可靠、低资源学习、预测效率等方面的突破，自然语言生成与智能写作的产业应用将继续保持高速的增长。

14.4.1. 政策环境

随着自然语言生成技术的发展，相关的政策、法律、法规从宏观层面逐渐深化到具体的算法和场景，并且逐步完善对于算法可控、可靠等角度的政策规范，对自然语言生成技术的产业落地提供了政策指引。这里重点关注国家整体政策和智能写作应用需求最强烈的媒体行业政策动向。

国家政策积极引导和推动自然语言生成技术的发展。2017 年 7 月，国务院发布《新一代人工智能发展规划》，该文件奠定了国内人工智能发展政策的基础，其中对自然语言生成技术多有提及，例如“重点突破跨媒体统一表征、关联理解与知识挖掘、知识图谱构建与学习、知识演化与推理、*智能描述与生成等技术*”，以及“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*。”。2018 年 10 月，人民日报刊发《习近平：推动我国新一代人工智能健康发展》，强调“人工智能是新一轮科技革命和产业变革的重要驱动力量”，同时也特别指出“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能*安全*、*可靠*、*可控*”。2021 年 3 月，“十四五规划”中多次提及人工智能和自然语言处理技术，从自然语言生成和智能写作技术角度出发，可以从“十四五规划”中看到在数字产业化、优秀文化作品创作等重要发展点的应用潜力。

新闻媒体行业积极拥抱自然语言生成与智能写作技术。2018 年 6 月，中华全国新闻工作者协会会同国家新闻出版广电总局等单位发布《中国新闻事业发展报告（2017 年）》，其中判断未来的新闻事业将由智能驱动，“人工智能正在对*新闻生产*、分发、反馈各环节

产生革命性影响”，“在*机器人写稿*、语音互动、人脸识别等智能应用基础上，进一步加强终端智能化建设，推动新闻信息产品和终端迭代升级”。2021 年 10 月，国家广播电视总局发布《广播电视和网络视听“十四五”发展规划》，其中总结在十三五期间，“人工智能技术在广播电视*内容生产*、分发传输、监测监管、网络安全保障等领域的融合应用取得初步成效”。规划在十四五期间，“跟踪研究全媒体内容智能认知和生产处理技术”，建立“智能云采编、智能云制作”服务，以及推动“虚拟主播”等新形态。

14.4.2. 产业发展现状

自然语言生成常见的应用场景包括内容创作、智能对话等。在内容创作方面，主要用于简单类型文章的自动写作，如股市报道、科技快讯等，同时也作为辅助创作的重要工具，为创作者提供初稿、创作灵感等。在智能对话方面，可以应用于聊天、任务型对话、知识问答等任务，从而提升产品或服务的有趣程度、实用性和易用性。这些应用场景在媒体、金融、营销等行业已经形成广泛落地，比如智能写作在媒体、营销中有较为广泛的应用。而智能对话在消费电子、电商客服中均有重要的应用。近年来受益于技术的逐渐成熟，语言生成的应用规模、企业数量、应用落地均有显著的提升。

产业规模迅速增长。新一代人工智能将成为推动经济发展的重要动能，这一点是包括中国、美国、欧盟等世界重要经济体的共识。在科技部《科技创新 2030-“新一代人工智能”重大项目 2020 年度项目申报指南的通知》中启动的 22 个研究项目中，与知识图谱、NLP 相关的占比 45%，较 2018 年的占比（19%）翻倍增长。考虑到自然语言生成技术的发展趋势与应用前景，自然语言生成相关产业的规模将以领先其他 NLP 技术的速度加速提升。2020 年末，国内知识图谱与 NLP 行业应用市场在 AI 中的整体占比为 8.8%，带动相关产业规模 655.8 亿元。由此估算，自然语言生成带动相关产业规模当前为百亿级，并将在下一个五年（2022-2026）达到千亿级。

大型企业和创业公司广泛参与。近几年随着语言生成技术的快速发展，其潜力和对相关行业生产力的促进作用也开始得到市场的认可，无论百度、腾讯、阿里巴巴等大型企业，还是雨后春笋般出现的创业公司，都从语言生成技术中挖掘可转换为市场价值的内容，并提供了大量技术平台、场景服务等产品。大型企业通常对于技术有较为全面的布局，一方面将生成技术用于自身 C 端的产品，同时也通过智能云支持广泛的行业客户。而创业企业通常聚焦于具体技术或者场景，进而对行业客户提供针对性的解决方案。

基于深度学习的生成技术广泛落地。在 2017-2021 年的五年中，自然语言生成技术经历了从统计机器学习到神经网络深度学习的变革，特别是大规模预训练模型的出现，基于深度学习的生成效果优势更加明显。在 GPU 等芯片技术、深度学习计算框架等基础设施的同步升级支持下，在机器翻译、人机对话、自动摘要、自动写作等应用问题中已经广泛应用了基于预训练深度神经网络的自然语言生成模型，有力支持了相关应用的更新换代。

多模态生成应用受到关注。随着视频内容、元宇宙等行业热点的出现，多模态生成技术和应用受到越来越多的关注。在内容创作中，不少企业研发了多模态内容创作能力，比如基于数据自动生成视频，基于文章自动生成视频等功能，有效提升内容生产效率，提升内容传播效果。同时，虚拟主播、数字人等新型生成应用也逐渐受到关注，由于可视化的天然优势，虚拟主播和数字人能够显著提升智能助手的交互效果，扩大其接受度和使用场景。

14.4.3. 产业应用案例

近年来，自然语言生成和智能写作技术逐渐成为产业界重点关注的方向。国外对于智能写作技术的研究已有几十年的历史，积累了大量研究成果，并在内容资讯、金融财经、数字营销和行政办公等应用领域发挥了重要作用。国际 IT 巨头企业，包括谷歌、微软、IBM、Facebook 等均将智能写作作为重要研发方向，结合自身优势打造智能写作技术创新应用产品。国内智能写作技术相关产品研发起步不久，也取得了很多成果；百度、腾讯、字节跳动、金山等公司也纷纷布局智能写作，抢占新一轮科技变革的先机。国内具有代表性的产业落地应用示例主要有：

百度大脑推出的“智能创作平台”，集合自然语言处理和知识图谱技术，旨在成为更懂用户的智能创作助手，主要包含自动创作、辅助创作、多模态创作三大功能：自动创作，用户可以通过接入数据、配置专属写作模板，快速实现批量和自动生成文章的能力；辅助创作，包括热点发现、事件脉络、热词分析、文本纠错、用词润色、文本审核、文章分类等技术，提升内容创作效率。

腾讯新闻机器人 Dreamwriter 基于大数据分析平台，能够在短时间内选出新闻点、抓取相关资料，按照特定的新闻体裁成稿，适用于资料量巨大的财经资讯新闻，机器人

的优势在于能够全面地抓取资料，从抓取资料到成稿发布秒级耗时，时效性上优于人工。

智能写稿机器人小明 BOT 从 2016 年开始研发，它可以从数据出发，分析例如足球比赛这样的视频，通过文本生成的技术，生成一篇全方位的比赛报道，再利用文本摘要的技术，把它摘要成简短的文字，利用机器翻译成多种语言。最后我们通过语音合成技术把文本可以生成语音读出来。我们也通过 AR 的技术去生成一个虚拟的形象、虚拟的播音员，他可以带有表情、带有动作的把整个文章播报出来。

微软亚洲研究院研发的机器人小冰 2017 年创作了首部诗集《阳光失了玻璃窗》于出版；在金融领域，小冰能够为用户提供由人工智能技术生成的上市公告文本摘要；在新闻资讯领域，小冰实现了中英双语金融资讯的整理、加工和创作。另外，微软对联是由微软亚洲研究院自然语言计算组研发的计算机自动对联系统。用户给定上联，它能够自动提供若干下联供用户选择；当用户确定一副对联后，还能生成若干四字横批供用户参考。

金山智能写作产品以深度学习神经网络算法为核心，融合注意力机制和神经网络语言模型技术，利用无监督学习方法实现文档的空间向量表示，建立多通道搜索推荐架构，实现在海量文档中达到高效准确的信息追踪和语义匹配，为用户生成高质量的写作提纲和段落内容，帮助用户完成创作。目前，金山智能写作产品已经成功落地到政府信创项目、招商证券、国企央企公文写作项目中。

语仓科技研发的智能公文辅助写作平台，以党政机关公文写作为需求导向，建设了海量公文资源库，以自然语言生成、语义检索、智能校对、知识图谱等为关键技术，研发了公文智能辅助写作、文本智能校对、写作素材智能推荐、智能政策解读等核心功能。该平台可使复杂繁琐的公文拟制、审批和归档等工作过程大大缩短，提高公文处理质量和效率，提升政府部门公文管理的智能化水平。

未来，人工智能写作产业通过加速科研、技术与产业的深度融合和规模化应用，汇聚整合海量高价值数据，将会为政务管理、文化教育、医疗卫生、环境和资源保护等领域智能化发展提供广阔的空间，推进人工智能技术和相关产业的良性发展。

14.4.4. 产业发展趋势

自然语言生成技术处于迅速发展的上升期，随着安全性、低资源学习、预测效率等

方面的技术突破，自然语言生成技术将保持高速增长并扩大应用范围，最终深度变革内容生产、人机交互等数字社会的核心生产力。

安全、可控、可靠将成为语言生成技术产业落地的基础。与语言理解算法不同，语言生成算法往往由模型生成前所未见的结果，且很可能直接面向包括普通用户在内的算法使用者。随着语言生成算法产业落地影响面的快速扩张，算法安全性、可控性、可靠性的衡量、增强与承诺，将是必须优先被解决的重要问题。以智能写作等内容生产算法为例，生成算法可以有效提升内容生产效率和效果，已经是业界和学界的共识，但智能写作在内容生产中的广泛落地，必须配套研发专用的技术和应用方案，避免自动生成内容违反社会管理要求、控制内容生产的领域和数量、确保内容本身的事实可靠性等。因此，虽然目前语言生成算法的安全性、可控性、可靠性研究还远未成熟，但已经引起学界和工业界的广泛关注。

低资源任务泛化能力是生成技术产业落地需要突破的关键瓶颈。自然语言生成的实际问题呈现出显著的发散性，即对于特定的领域、应用场景、用户受众，即使在技术上归属于同一个问题，也无法简单使用同一个模型完成任务，需要针对具体需求做泛化迁移。以自动摘要为例，新闻摘要、对话摘要、网页摘要等不同的摘要需求，不仅领域、输入-输出文本形式不同，且摘要算法中重要内容理解与表达的内生逻辑也有很大差异。即使基于预训练-精调的范式，也需要较大数量的任务标注语料使生成模型完成任务迁移，但具体任务中摘要语料标注的成本极高，会造成显著的落地障碍。目前，这一问题正在通过大模型、多任务学习、指令控制生成等技术路线逐步探索，但在低资源条件下，统一描述语言生成需求、精细控制需求细节等应用难题仍有待解决。

提升大模型预测性价比是扩大产业落地的关键动力。当前自然语言生成算法效果距离理想程度依然有一定差距，因此在产业应用中选择效果最强的模型将是必然趋势。然而，目前模型效果强大往往意味着模型规模巨大，导致了语言生成的规模化落地成本较高。以美国 OpenAI 的 GPT-3 千亿大模型为例，据外媒估计，部署预测模型需要的硬件投入至少 10 万美元，即使调用其价格约为 0.06 美元/750 英文词的在线预测接口，普通个人网站使用该接口也需要至少需要每月 4000 美元成本。因此，缩减下游任务模型规模、集中部署通用能力等降本增效方案的成熟度，将决定未来语言生成产业落地的速度。

14.5. 总结及展望

基于神经网络的语言生成模型、尤其是近年来快速发展的大规模预训练模型得到了快速发展，生成模型的性能也获得了大幅提升。许多产品发展应用场景如机器翻译、自动文摘、文案生成等，生成模型也开始扮演着越来越重要的作用。但是目前语言生成在可控性、知识运用、长文本生成、创造力和表现力等方面表现还有明显不足。

语言生成的不可控主要来自两个方面：模型生成概率的估计不够可靠和透明，以及从生成概率中采样时具有随机性。这可能会在实际应用过程当中会带来意想不到的后果。同时，使得生成结果满足某些文本属性（如情绪、风格等）也是可控性研究的重要方面，语言生成模型应该有能力适配到不同需求的下游任务中。

现有的语言模型缺乏常识知识，如目前最强大的语言生成模型 GPT3 会说出“在 Apple 发布会上发布林肯汽车”[16]，这使得生成结果的可信度显著降低。因此，如何将世界知识、常识知识或事实知识与语言生成模型结合在一起，对生成结果加以约束，也是亟待解决的研究问题。

生成高质量的长文本也是一项重要挑战，如故事生成、散文生成、现代诗歌生成等。尽管现有的生成模型在短文本生成上取得了显著进展，但在长文本生成上，即便最强大的 GPT2、GPT3 模型也还面临显著的问题：容易生成重复或者通用文本、缺乏连贯性和逻辑性等。这是由于这些通用的语言模型仅仅对词间的共现关系进行建模，而缺乏对话题转移、事件关系、篇章结构的宏观规划，从而缺少句间一致性和连贯性，并进一步导致了通用文本和重复内容的产生。

目前生成模型的泛化能力较弱，不具备创造性。许多开放端语言生成任务对创造性和表现力均有较高的要求，如故事生成、歌词生成等任务。但是模型往往难以生成数据集之外的新颖内容，如有趣的故事情节。

语言生成的自动评价也仍是一个巨大的挑战。人工评价往往成本较高、费时费力且难以复现，可靠的自动评价指标对于快速迭代和发展生成模型非常重要。现有的评价生成质量的自动指标往往依赖于有限的参考文本，难以对许多开放端生成任务做出有效的评价。相比于收集更多人工标注的参考文本，未来研究的方向更可能是发展更细粒度或者任务相关的评价指标，如在流畅度、语法性、连贯性、多样性、创造性等方面分别通过自监督学

习训练可学习的无参考自动评价指标，进而通过强化学习等手段促进生成模型的进步。这种方法不依赖人工评价、数据标注和任何具体模型，在泛化性、鲁棒性上可以实现更好的表现。

总之，目前最先进的语言生成模型仍然距离创作高质量文本有较大差距，在知识运用、规划等方面还存在显著不足。期待越来越多的研究者投入自然语言生成领域的研究，推动这一领域的研究发展和应用实践。

14.6. 参考文献

[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.

[3] Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." arXiv preprint arXiv:1908.08345 (2019).

[4] 67Fabbri A R, Han S, Li H, et al. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation[J]. arXiv preprint arXiv:2010.12836, 2020.

[5] Laban P, Hsi A, Canny J, et al. The summary loop: Learning to write abstractive summaries without examples[J]. arXiv preprint arXiv:2105.05361, 2021.

[6] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

[7] Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In International Conference on Machine Learning, pp. 11328-11339. PMLR, 2020.

[8] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

[9] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out, pp. 74-81. 2004.

[10] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72. 2005.

[11] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

[12] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

[13] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

[14] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9.

[15] Zhou, Chunting, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. "Detecting hallucinated content in conditional neural sequence generation." arXiv preprint arXiv:2011.02593 (2020).

[16] Haonan, Wang, Gao Yang, Bai Yu, Mirella Lapata, and Huang Heyan. "Exploring Explainable Selection to Control Abstractive Summarization." arXiv preprint arXiv:2004.11779 (2020).

[17] Chen, Jiaao, and Diyi Yang. "Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[18] Zhao, Lulu, Weiran Xu, and Jun Guo. "Improving abstractive dialogue summarization with graph structures and topic words." Proceedings of the 28th International Conference on Computational Linguistics. 2020.

[19] Chen, Jiaao, and Diyi Yang. "Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

[20] Xiachong, Feng, Feng Xiaocheng, and Qin Bing. "Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks." Proceedings of the 20th Chinese National Conference on Computational Linguistics. 2021.

[21] Lei, Yuejie, et al. "Hierarchical Speaker-Aware Sequence-to-Sequence Model for Dialogue Summarization." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[22] Liu, Zhengyuan, Ke Shi, and Nancy F. Chen. "Coreference-Aware Dialogue Summarization." arXiv preprint arXiv:2106.08556 (2021).

[23] Narayan, Shashi, et al. "Planning with Learned Entity Prompts for Abstractive Summarization." arXiv preprint arXiv:2104.07606 (2021).

[24] Wu, Chien-Sheng, et al. "Controllable Abstractive Dialogue Summarization with Sketch Supervision." arXiv preprint arXiv:2105.14064 (2021).

[25] Liu, Yizhu, Zhiyi Luo, and Kenny Zhu. "Controlling length in abstractive summarization using a convolutional neural network." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4110-4119. 2018.

[26] Takase, Sho, and Naoaki Okazaki. "Positional encoding to control output sequence length." arXiv preprint arXiv:1904.07418 (2019).

[27] Makino, Takuya, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. "Global optimization under length constraint for neural text summarization." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1039-1048. 2019.

[28] Bai, Yu, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. "Bridging the Gap: Cross-Lingual Summarization with Compression Rate." arXiv preprint arXiv:2110.07936 (2021).

[29] Jiang, Yichen, et al. "Enriching Transformers with Structured Tensor-Product Representations for Abstractive Summarization." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

[30] Dou, Zi-Yi, et al. "Gsum: A general framework for guided neural abstractive summarization." arXiv preprint arXiv:2010.08014 (2020).

[31] Zhu, Junnan, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. "Multimodal summarization with guidance of multimodal reference." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 9749-9756. 2020.

[32] Zhu, Junnan, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. "MSMO: Multimodal summarization with multimodal output." In Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 4154-4164. 2018.

[33] Maynez J, Narayan S, Bohnet B, et al. On Faithfulness and Factuality in Abstractive Summarization[J]. arXiv preprint arXiv:2005.00661, 2020.

[34] Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries[J]. arXiv preprint arXiv:2004.04228, 2020.

[35] Durmus E, He H, Diab M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization[J]. arXiv preprint arXiv:2005.03754, 2020.

[36] Matsumaru K, Takase S, Okazaki N. Improving Truthfulness of Headline Generation[J]. arXiv preprint arXiv:2005.00882, 2020.

[37] Kang D, Hashimoto T. Improved Natural Language Generation via Loss Truncation[J]. arXiv preprint arXiv:2004.14589, 2020.

[38] Cao, Shuyang, and Lu Wang. "CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization." arXiv preprint arXiv:2109.09209 (2021).

[39] Zhu, Junnan, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. "NCLS: Neural cross-lingual summarization." arXiv preprint arXiv:1909.00156 (2019).

[40] Zhu, Junnan, Yu Zhou, Jiajun Zhang, and Chengqing Zong. "Attend, translate and summarize: An efficient method for neural cross-lingual summarization." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1309-1321. 2020.

[41] Ouyang, Jessica, Boya Song, and Kathleen McKeown. "A robust abstractive system for cross-lingual summarization." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2025-2031. 2019.

[42] Duan, Xiangyu, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. "Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3162-3172. 2019.

[43] Cao, Yue, Hui Liu, and Xiaojun Wan. "Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6220-6231. 2020.

[44] Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown. "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization." arXiv preprint arXiv:2010.03093 (2020).

[45] Xu, Ruochen, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. "Mixed-Lingual Pre-training for Cross-lingual Summarization." arXiv preprint arXiv:2010.08892 (2020).

[46] Bai, Yu, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. "Bridging the Gap: Cross-Lingual Summarization with Compression Rate." arXiv preprint arXiv:2110.07936 (2021).

[47] Takase, Sho, and Naoaki Okazaki. "Multi-Task Learning for Cross-Lingual Abstractive Summarization." arXiv preprint arXiv:2010.07503 (2020).

[48] Bai, Yu, Yang Gao, and Heyan Huang. "Cross-Lingual Abstractive Summarization with Limited Parallel Resources." arXiv preprint arXiv:2105.13648 (2021).

[49] Zhang, Rui, and Joel Tetreault. "This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[50] Kano, Ryuji, et al. "Identifying Implicit Quotes for Unsupervised Extractive Summarization of Conversations." Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020.

[51] Zhang, Shiyue, et al. "EmailSum: Abstractive Email Thread Summarization." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

[52] Scialom, Thomas, et al. "MLSUM: The multilingual summarization corpus." arXiv preprint arXiv:2004.14900 (2020).

[53] Hasan, Tahmid, et al. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages." arXiv preprint arXiv:2106.13822 (2021).

[54] Cao, Yue, et al. "MultiSumm: Towards a unified model for multi-lingual abstractive summarization." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.

[55] Li, Manling, et al. "Timeline Summarization based on Event Graph Compression via Time-Aware Optimal Transport." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[56] Steen, Julius, and Katja Markert. "Abstractive Timeline Summarization." Proceedings of the 2nd Workshop on New Frontiers in Summarization. 2019.

[57] Liu, Yang, and Mirella Lapata. "Hierarchical transformers for multi-document summarization." arXiv preprint arXiv:1905.13164 (2019).

[58] Fabbri, Alexander R., Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model." arXiv preprint arXiv:1906.01749 (2019).

[59] Xu, Yumo, and Mirella Lapata. "Coarse-to-fine query focused multi-document summarization." In Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP), pp. 3632-3645. 2020.

[60] Chen, Xiuying, et al. "Learning towards Abstractive Timeline Summarization." IJCAI. 2019.

[61] Ansah, Jeffery, et al. "A graph is worth a thousand words: Telling event stories using timeline summarization graphs." The World Wide Web Conference. 2019.

[62] Amplayo, Reinald Kim, Stefanos Angelidis, and Mirella Lapata. "Aspect-Controllable Opinion Summarization." arXiv preprint arXiv:2109.03171 (2021).

[63] Yu T, Liu Z, Fung P. Adaptsum: Towards low-resource domain adaptation for abstractive summarization[J]. arXiv preprint arXiv:2103.11332, 2021.

[64] Fu X, Zhang Y, Wang T, et al. RepSum: Unsupervised Dialogue Summarization based on Replacement Strategy[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 6042-6051.

[65] Liu P J, Chung Y A, Ren J. SummAE: Zero-shot abstractive text summarization using length-agnostic auto-encoders[J]. arXiv preprint arXiv:1910.00998, 2019.

[66] Bing, Lidong, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J. Passonneau. "Abstractive multi-document summarization via phrase selection and merging." arXiv preprint arXiv:1506.01597 (2015).

[66] Liang P, Jordan M I, Klein D. Learning semantic correspondences with less supervision[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 91-99.

[67] Chen D L, Mooney R J. Learning to sportscast: a test of grounded language acquisition[C]//Proceedings of the 25th international conference on Machine learning. 2008: 128-135.

[68] Lebret R, Grangier D, Auli M. Neural Text Generation from Structured Data with Application to the Biography Domain[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 1203-1213.

[69] Novikova J, Dušek O, Rieser V. The E2E Dataset: New Challenges For End-to-End Generation[C]//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. 2017: 201-206.

[70] Banik E, Gardent C, Kow E. The KBGen Challenge[C]//Proceedings of the 14th European Workshop on Natural Language Generation. 2013: 94-97.

[71] Gardent C, Shimorina A, Narayan S, et al. Creating training corpora for nlg micro-planning[C]//55th annual meeting of the Association for Computational Linguistics (ACL). 2017.

[72] Wiseman S, Shieber S M, Rush A M. Challenges in Data-to-Document Generation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2253-2263.

[73] Puduppully R, Dong L, Lapata M. Data-to-text Generation with Entity Modeling[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2023-2035.

[74] Chen M, Wiseman S, Gimpel K. WikiTableT: A large-scale data-to-text dataset for generating Wikipedia article sections[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 193-209.

[75] Parikh A, Wang X, Gehrmann S, et al. ToTTo: A Controlled Table-To-Text Generation Dataset[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 1173-1186.

[76] Nan L, Radev D, Zhang R, et al. DART: Open-Domain Structured Data Record to Text Generation[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 432-447.

[77] Li T, Fang L, Lou J G, et al. TWT: Table with Written Text for Controlled Data-to-Text Generation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 1244-1254.

[78] Chen W, Chen J, Su Y, et al. Logical Natural Language Generation from Open-Domain Tables[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7929-7942.

[79] Chen Z, Chen W, Zha H, et al. Logic2Text: High-Fidelity Natural Language Generation from Logical Forms[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 2096-2111.

[80] Mei H, Bansal M, Walter M R. What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 720-730.

[81] Puduppully R, Dong L, Lapata M. Data-to-text generation with content selection and planning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 6908-6915.

[82] Gong H, Bi W, Feng X, et al. Enhancing content planning for table-to-text generation with data understanding and verification[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 2905-2914.

[83] Su Y, Vandyke D, Wang S, et al. Plan-then-Generate: Controlled Data-to-Text Generation via Planning[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 895-909.

[84] Bai Y, Li Z, Ding N, et al. Infobox-to-text Generation with Tree-like Planning based Attention Network[C]//IJCAI. 2020: 3773-3779.

[85] Liu T, Wang K, Sha L, et al. Table-to-Text Generation by Structure-aware Seq2Seq Learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 4881-4888.

[86] Gong H, Feng X, Qin B, et al. Table-to-Text Generation with Effective Hierarchical Encoder on Three Dimensions (Row, Column and Time)[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3143-3152.

[87] Distiawan B, Qi J, Zhang R, et al. GTR-LSTM: A triple encoder for sentence generation from RDF data[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1627-1637.

[88] Zhao C, Walker M, Chaturvedi S. Bridging the structural gap between encoding and decoding for data-to-text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2481-2491.

[89] Xing X, Wan X. Structure-Aware Pre-Training for Table-to-Text Generation[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP

2021. 2021: 2273-2278.

[90] Li L, Ma C, Yue Y, et al. Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5979-5989.

[91] Nie F, Wang J, Yao J, et al. Operation-guided Neural Networks for High Fidelity Data-To-Text Generation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3879-3889.

[92] Jain P, Laha A, Sankaranarayanan K, et al. A Mixed Hierarchical Attention Based Encoder-Decoder Approach for Standard Table Summarization[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 622-627.

[93] Qin G, Yao J G, Wang X, et al. Learning latent semantic annotations for grounding natural language to structured data[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3761-3771.

[94] Sha L, Mou L, Liu T, et al. Order-Planning Neural Text Generation from Structured Data[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 5414-5421.

[95] Bao J, Tang D, Duan N, et al. Table-to-Text: Describing Table Region with Natural Language[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 5020-5027.

[96] Shen X, Chang E, Su H, et al. Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7155-7165.

[97] Song L, Wang A, Su J, et al. Structural Information Preserving for Graph-to-Text Generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7987-7998.

[98] Li L, Wan X. Point precisely: Towards ensuring the precision of data in generated texts using delayed copy mechanism[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1044-1055.

[99] Li Z, Lin Z, Ding N, Zheng H, et al. Triple-to-Text Generation with an Anchor-to-Prototype Framework[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020: 3780-3786.

[100] Wang P, Lin J, Yang A, et al. Sketch and Refine: Towards Faithful and Informative Table-to-Text Generation[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 4831–4843.

[101] Liu T, Zheng X, Chang B, et al. Towards Faithfulness in Open Domain Table-to-text Generation from an Entity-centric View[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 13415-13423.

[102] 73 Puduppully R, Dong L, Lapata M. Data-to-text Generation with Entity

Modeling[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2023-2035.

[103] Ghosh S, Qi Z, Chaturvedi S, et al. How Helpful is Inverse Reinforcement Learning for Table-to-Text Generation?[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 71-79.

[104] Iso H, Uehara Y, Ishigaki T, et al. Learning to Select, Track, and Generate for Data-to-Text[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2102-2113.

[105] Feng X, Sun Y, Qin B, et al. Learning to Select Bi-Aspect Information for Document-Scale Text Content Manipulation[C]//AAAI. 2020: 7716-7723.

[106] Dhingra B, Faruqui M, Parikh A, et al. Handling Divergent Reference Texts when Evaluating Table-to-Text Generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4884-4895.

[107] Rebuffel C, Scialom T, Soulier L, et al. Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 8029–8036.

[108] Faille J, Gatt A, Gardent C. Entity-Based Semantic Adequacy for Data-to-Text Generation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 1530-1540.

[109] Ma S, Yang P, Liu T, et al. Key Fact as Pivot: A Two-Stage Model for Low Resource Table-to-Text Generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2047-2057.

[110] Chen Z, Eavani H, Chen W, et al. Few-shot nlg with pre-trained language model [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 183-190.

[111] Gong H, Sun Y, Feng X, et al. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1978-1988.

[112] Su Y, Meng Z, Baker S, et al. Few-Shot Table-to-Text Generation with Prototype Memory[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 910-917.

[113] Zhao W, Liu Y, Wan Y, et al. Attend, Memorize and Generate: Towards Faithful Table-to-Text Generation in Few Shots[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 4106-4117.

[114] Perez-Beltrachini L, Lapata M. Bootstrapping Generators from Noisy Data[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long

Papers). 2018: 1516-1527.

[115] Fu Z, Shi B, Lam W, et al. Partially-Aligned Data-to-Text Generation with Distant Supervision[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9183-9193.

[116] Suadaa L H, Kamigaito H, Funakoshi K, et al. Towards table-to-text generation with numerical reasoning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1451-1465.

[117] Chen W, Tian J, Li Y, et al. De-Confounded Variational Encoder-Decoder for Logical Table-to-Text Generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5532-5542.

[118] Jianing Zhou, and Suma Bhat. "Paraphrase Generation: A Survey of the State of the Art." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5075-5086. 2021.

[119] Yao Fu, Yansong Feng, and John P. Cunningham. "Paraphrase Generation with Latent Bag of Words." Advances in Neural Information Processing Systems 32 (2019): 13645-13656.

[120] Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. "Decomposable Neural Paraphrase Generation." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3403-3414. 2019.

[121] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. "Neural Paraphrase Generation with Stacked Residual LSTM Networks." In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2923-2934. 2016.

[122] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. "A deep generative framework for paraphrase generation." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.

[123] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. "Learning to Paraphrase for Question Answering." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 875-886. 2017.

[124] Shuguang Zhu, Xiang Cheng, Sen Su, and Shuang Lang. "Knowledge-based question answering by jointly generating, copying and paraphrasing." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2439-2442. 2017.

[125] Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. "Improving statistical machine translation with a multilingual paraphrase database." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1379-1390. 2015.

[126] Brian Thompson and Matt Post. "Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 90-121. 2020.

[127] Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. "Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6806-6817. 2020.

[128] Jonathan Berant, and Percy Liang. "Semantic parsing via paraphrasing." In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1415-1425. 2014.

[129] Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3609-3619. 2019.

[130] Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. "Paraphrase Augmented Task-Oriented Dialog Generation." In ACL. 2020.

[131] Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, pages 834–842. The Association for Computer Linguistics.

[132] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, New York, NY, USA - August 24 - 27, 2014, pages 1156–1165. ACM.

[133] Igor A. Bolshakov and Alexander F. Gelbukh. 2004. Synonymous paraphrasing using wordnet and internet. In Natural Language Processing and Informa- tion Systems, 9th International Conference on Applications of Natural Languages to Information Sys- tems, NLDB 2004, Salford, UK, June 23-25, 2004, Proceedings, volume 3136 of Lecture Notes in Com- puter Science, pages 312–323. Springer.

[134] David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA. The Association for Computational Linguistics.

[135] Shashi Narayan, Siva Reddy, and Shay B. Cohen. 2016. Paraphrase generation from latent-variable pcfgs for semantic parsing. In INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK, pages 153–162. The Association for Computer Linguistics.

[136] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. "Paraphrase generation as monolingual translation: Data and evaluation." In Proceedings of the 6th International Natural Language Generation Conference. 2010.

[137] Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. "Combining multiple resources to improve SMT-based paraphrasing model." In Proceedings of ACL-08: HLT, pp. 1021-1029. 2008.

[138] Colin Bannard and Chris Callison-Burch. "Paraphrasing with bilingual parallel corpora." In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 597-604. 2005.

[139] Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. "Leveraging multiple MT engines for paraphrase generation." In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 1326-1334. 2010.

[140] Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. "Exploring diverse expressions for paraphrase generation." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3173-3182. 2019.

[141] Yue Cao and Xiaojun Wan. "DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 2411-2421. 2020.

[142] Qiongkai Xu, Juyan Zhang, Lizhen Qu, Lexing Xie, and Richard Nock. "D-page: Diverse paraphrase generation." arXiv preprint arXiv:1808.04364 (2018).

[143] Wenqing Chen, Jidong Tian, Liqiang Xiao, Hao He, and Yaohui Jin. " A Semantically Consistent and Syntactically Variational Encoder-Decoder Framework for Paraphrase Generation." In Proceedings of the 28th International Conference on Computational Linguistics, pp. 1186-1198. 2020.

[144] Mingtong Liu, Erguang Yang, Deyi Xiong, Yujie Zhang, Yao Meng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2020. A learning-exploring method to generate diverse paraphrases with multi-objective deep reinforcement learning. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2310–2321.

[145] Zhe Lin, and Xiaojun Wan. "Pushing Paraphrase Away from Original Sentence: A Multi-Round Paraphrase Generation Approach." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1548-1557. 2021.

[146] Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31.

[147] Zibo Lin, Ziran Li, Ning Ding, Hai-Tao Zheng, Ying Shen, Wei Wang, and Cong-Zhi Zhao. 2020. Integrating linguistic knowledge to sentence paraphrase generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8368– 8375.

[148] Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Pa- pers), pages 196–206.

[149] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1875–1885.

[150] Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5972–5984.

[151] Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 238–252.

[152] Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. Transactions of the Association for Computational Linguistics, 8:329–345.

[153] Shaohan Huang, Yu Wu, Furu Wei, and Zhongzhi Luan. 2019. Dictionary-guided editing networks for paraphrase generation. In Proceedings of the AAAI Con- ference on Artificial Intelligence, volume 33, pages 6546–6553.

[154] Amirhossein Kazemnejad, Mohammadreza Salehi, and Mahdieh Soleymani Baghshah. 2020. Paraphrase generation by learning how to edit from samples. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6010– 6021.

[155] Zhe Lin, Yitao Cai, and Xiaojun Wan. "Towards Document-Level Paraphrase Generation with Sentence Rewriting and Reordering." In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 1033-1044. 2021.

[156] Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. "Unsupervised Paraphrasing by Simulated Annealing." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 302-312. 2020.

[157] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. "Paraphrasing revisited with neural machine translation." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 881-893. 2017.

[158] Yitao Cai, Yue Cao, and Xiaojun Wan. "Revisiting Pivot-Based Paraphrase Generation: Language Is Not the Only Optional Pivot." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4255-4268. 2021.

[159] Huang, M., Zhu, X., & Gao, J. (2020). Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems (TOIS), 38(3), 1-32.

[160] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., ... & Matsuo, Y. (2014, August). Towards an open-domain conversational system fully based on natural language processing. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 928-939).

[161] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., ... & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. arXiv preprint arXiv:1506.06714.

[162] Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364.

[163] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[164] Hadsell, R., Chopra, S., & LeCun, Y. (2006, June). Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1735-1742). IEEE.

[165] Cheng, Pengyu , et al. "CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information." (2020).

[166] Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403.

[167] Lee, S., Lee, D. B., & Hwang, S. J. (2020). Contrastive learning with adversarial perturbations for conditional text generation. arXiv preprint arXiv:2012.07280.

[168] Mi, F., Chen, L., Zhao, M., Huang, M., & Faltings, B. (2020). Continual learning for natural language generation in task-oriented dialog systems. arXiv preprint arXiv:2010.00910.

[169] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[170] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[171] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[172] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... & Hon, H. W. (2019). Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197.

[173] Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536.

[174] Bao, S., He, H., Wang, F., Wu, H., & Wang, H. (2019). Plato: Pre-trained dialogue generation model with discrete latent variable. arXiv preprint arXiv:1910.07931.

[175] Kim, B., Ahn, J., & Kim, G. (2020). Sequential latent knowledge selection for knowledge-grounded dialogue. arXiv preprint arXiv:2002.07510.

[176] Meng, C., Ren, P., Chen, Z., Sun, W., Ren, Z., Tu, Z., & Rijke, M. D. (2020, July). Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1151-1160).

[177] Zhao, X., Wu, W., Xu, C., Tao, C., Zhao, D., & Yan, R. (2020). Knowledge-grounded dialogue generation with pre-trained language models. arXiv preprint arXiv:2010.08824.

[178] Boussaha, B. E. A., Hernandez, N., Jacquin, C., & Morin, E. (2019). Deep retrieval-based dialogue systems: a short review. arXiv preprint arXiv:1907.12878.

[179] Zheng, C., & Huang, M. (2021). Exploring prompt-based few-shot learning for grounded dialog generation. arXiv preprint arXiv:2109.06513.

[180] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980.

[181] Gu, X., Yoo, K. M., & Lee, S. W. (2021). Response Generation with Context-Aware Prompt Learning. arXiv preprint arXiv:2111.02643.

[182] Gatt A, Krahmer E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation[J]. Journal of Artificial Intelligence Research, 2018, 61: 65-170.

[183] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv: 1301.3781, 2013.

[184] Guan J, Mao X, Fan C, et al. Long Text Generation by Modeling Sentence-Level and Discourse-Level Coherence[J]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 6379--6393.

[185] Fan A, Lewis M, Dauphin Y. Hierarchical Neural Story Generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 889-898.

[186] Guan J, Huang F, Zhao Z, et al. A knowledge-enhanced pretraining model for commonsense story generation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 93-108.

[187] Xu P, Patwary M, Shoeybi M, et al. Controllable Story Generation with External Knowledge Using Large-Scale Language Models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 2831-2845.

[188] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[189] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.

[190] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7871-7880.

[191] Mostafazadeh N, Chambers N, He X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 839-849.

[192] Fan A, Lewis M, Dauphin Y. Hierarchical Neural Story Generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 889-898.

[193] Louis A, Sutton C. Deep Dungeons and Dragons: Learning Character-Action Interactions from Role-Playing Game Transcripts[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 708-713.

[194] Rae J W, Potapenko A, Jayakumar S M, et al. Compressive Transformers for Long-Range Sequence Modelling[C]//International Conference on Learning Representations. 2019.

[195] Akoury N, Wang S, Whiting J, et al. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation[J]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9157-9166.

[196] Guan J, Feng Z, Chen Y, et al. LOT: A Benchmark for Evaluating Chinese Long Text Understanding and Generation[J]. arXiv preprint arXiv:2108.12960, 2021.

[197] Dou Y, Forbes M, Koncel-Kedziorski R, et al. Scarecrow: A framework for scrutinizing machine text[J]. arXiv preprint arXiv:2107.01294, 2021.

[198] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[199] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[200] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[201] Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), pp.423-443.

[202] Li, L., Tang, S., Deng, L., Zhang, Y. and Tian, Q., 2017, February. Image caption with global-local attention. In Thirty-first AAAI conference on artificial intelligence.

[203] Yu, J., Li, J., Yu, Z. and Huang, Q., 2019. Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology, 30(12), pp.4467-4480.

[204] Lu, X., Wang, B., Zheng, X. and Li, X., 2017. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4), pp.2183-2195.

[205] Sharma, P., Ding, N., Goodman, S. and Soricut, R., 2018, July. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).

[206] Wang, B., Ma, L., Zhang, W. and Liu, W., 2018. Reconstruction network for video captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7622-7631).

[207] Aafaq, N., Akhtar, N., Liu, W., Gilani, S.Z. and Mian, A., 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12487-12496).

[208] Krishna, R., Hata, K., Ren, F., Fei-Fei, L. and Carlos Niebles, J., 2017. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision (pp. 706-715).

[209] Zhou, L., Xu, C. and Corso, J.J., 2018, April. Towards automatic learning of procedures from web instructional videos. In Thirty-Second AAAI Conference on Artificial Intelligence.

[210] Cheng, W.F., Wu, C.C., Song, R., Fu, J., Xie, X. and Nie, J.Y., 2018. Image inspired poetry generation in xiaoice. arXiv preprint arXiv:1808.03090.

[211] Liu, D., Guo, Q., Li, W. and Lv, J., 2018, July. A multi-modal chinese poetry generation model. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[212] Zhipeng, G., Yi, X., Sun, M., Li, W., Yang, C., Liang, J., Chen, H., Zhang, Y. and Li, R., 2019, July. Jiuge: A human-machine collaborative chinese classical poetry generation system. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 25-30).

[213] Xu, L., Jiang, L., Qin, C., Wang, Z. and Du, D., 2018, April. How images inspire poems: Generating classical chinese poetry from images with memory networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[214] Liu, B., Fu, J., Kato, M.P. and Yoshikawa, M., 2018, October. Beyond narrative description: Generating poetry from images by multi-adversarial training. In Proceedings of the 26th ACM international conference on Multimedia (pp. 783-791).

[215] Liu, L., Wan, X. and Guo, Z., 2018, October. Images2poem: Generating chinese poetry from image streams. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1967-1975).

[216] Wu, C.C., Song, R., Sakai, T., Cheng, W.F., Xie, X. and Lin, S.D., 2019, October. Evaluating Image-Inspired Poetry Generation. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 539-551). Springer, Cham.

[217] Huang T H, Ferraro F, Mostafazadeh N, et al. Visual storytelling[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1233-1239.

[218] Gonzalez-Rico D, Fuentes-Pineda G. Contextualize, show and tell: A neural visual storyteller[J]. arXiv preprint arXiv:1806.00738, 2018.

[219] Kim T, Heo M O, Son S, et al. Glac net: Glocal attention cascading networks for multi-image cued story generation[J]. arXiv preprint arXiv:1805.10973, 2018.

[220] Wang X, Chen W, Wang Y F, et al. No metrics are perfect: Adversarial reward learning for visual storytelling[J]. arXiv preprint arXiv:1804.09160, 2018.

[221] Hu J, Cheng Y, Gan Z, et al. What makes a good story? designing composite rewards for visual storytelling[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 7969-7976.

[222] Chen Z, Zhang X, Boedihardjo A P, et al. Multimodal storytelling via generative adversarial imitation learning[J]. arXiv preprint arXiv:1712.01455, 2017.

[223] Hsu C Y, Chu Y W, Yang T L, et al. Stretch-VST: Getting Flexible With Visual Stories[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021: 356-362.

[224] Chen, H., Huang, Y., Takamura, H. and Nakayama, H., 2021. Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling. arXiv preprint arXiv:2102.02963.

[225] Elliott, D. and Kádár, A., 2017. Imagination improves multimodal translation. arXiv preprint arXiv:1705.04350.

[226] Elliott, D., Frank, S., Sima'an, K. and Specia, L., 2016. Multi30k: Multilingual english-german image descriptions. arXiv preprint arXiv:1605.00459.

[227] Wang, D. and Xiong, D., 2021, January. Efficient Object-Level Visual Context Modeling for Multimodal Machine Translation: Masking Irrelevant Objects Helps Grounding. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 4, pp. 2720-2728).

[228] Lei, J., Yu, L., Bansal, M. and Berg, T.L., 2018. Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696.

[229] Hu, R., Singh, A., Darrell, T. and Rohrbach, M., 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9992-10002).

[230] Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A. and Essa, I., 2019, May. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2352-2356). IEEE.

[231] Lin, X., Bertasius, G., Wang, J., Chang, S.F., Parikh, D. and Torresani, L., 2021. VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7005-7015).

[232] Ni M, Huang H, Su L, et al. M3p: Learning universal representations via multitask multilingual multimodal pre-training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3977-3986.

[233] Hu R, Singh A. UniT: Multimodal Multitask Learning with a Unified Transformer[J]. arXiv preprint arXiv:2102.10772, 2021.

[234] Luo H, Ji L, Shi B, et al. Univl: A unified video and language pre-training model for multimodal understanding and generation[J]. arXiv preprint arXiv:2002.06353, 2020.

[235] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models[J]. arXiv preprint arXiv:2109.01134, 2021.

[236] Xia Q, Huang H, Duan N, et al. Xgpt: Cross-modal generative pre-training for image captioning[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2021: 786-797.

[237] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long Zhou, Michele Tufano, Ming Gong, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng, Shengyu Fu, Shujie Liu. CodeXGLUE: A Machine Learning Benchmark Dataset for Code Understanding and Generation. NeurIPS 2021.

[238] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, Wojciech Zaremba. Evaluating Large Language Models Trained on Code. arXiv, 2021.

[239] Colin B. Clement, Shuai Lu, Xiaoyu Liu, Michele Tufano, Dawn Drain, Nan Duan, Neel Sundaresan, Alexey Svyatkovskiy. Long-Range Modeling of Source Code Files with eWASH: Extended Window Access by Syntax Hierarchy. EMNLP 2021.

[240] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, Ming Zhou. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. EMNLP 2020.

[241] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, Ming Zhou. GraphCodeBERT: Pre-training Code Representations with Data Flow. ICLR 2021.

[242] Yue Wang, Weishi Wang, Shafiq Joty, Steven C.H. Hoi. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. EMNLP 2021.

[243] Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, Kai-Wei Chang. Unified Pre-training for Program Understanding and Generation. NAACL 2021.

[244] Bailin Wang, Wenpeng Yin, Xi Victoria Lin, Caiming Xiong. Learning to Synthesize Data for Semantic Parsing. NAACL 2021.

[245] Colin B. Clement, Dawn Drain, Jonathan Timcheck, Alexey Svyatkovskiy, Neel Sundaresan. PyMT5: multi-mode translation of natural language and Python code with transformers. EMNLP 2020.

[246] Daya Guo, Alexey Svyatkovskiy, Jian Yin, Nan Duan, Marc Brockschmidt, Miltiadis Allamanis. Learning to Generate Code Sketches. arXiv 2021.

[248] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[249] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.

[250] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.

[251] Nema P, Khapra M M. Towards a better metric for evaluating question generation systems[J]. arXiv preprint arXiv:1808.10192, 2018.

[252] Raykar V C, Yu S, Zhao L H, et al. Learning from crowds[J]. Journal of Machine Learning Research, 2010, 11(4).

[253] Popović M. chrF++: words helping character n-grams[C]//Proceedings of the second conference on machine translation. 2017: 612-618.

[254] He T, Chen J, Ma L, et al. ROUGE-C: A fully automated evaluation method for multi-document summarization[C]//2008 IEEE International Conference on Granular Computing. IEEE, 2008: 269-274.

[255] Lo C. YiSi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 507-513.

[256] Zhao T, Lala D, Kawahara T. Designing precise and robust dialogue response evaluators[J]. arXiv preprint arXiv:2004.04908, 2020.

[257] Stent A, Marge M, Singhai M. Evaluating evaluation methods for generation in the presence of variation[C]//international conference on intelligent text processing and computational linguistics. Springer, Berlin, Heidelberg, 2005: 341-351.

[258] Ma Q, Wei J, Bojar O, et al. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 62-90.

[259] Zhang Y, Vogel S, Waibel A. Interpreting bleu/nist scores: How much improvement do we need to have a better system?[C]//LREC. 2004.

[260] Callison-Burch C, Osborne M, Koehn P. Re-evaluating the role of BLEU in machine translation research[C]//11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.

[261] Sai A B, Gupta M D, Khapra M M, et al. Re-evaluating adem: A deeper look at scoring dialogue responses[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6220-6227.

[262] Kilickaya M, Erdem A, Ikizler-Cinbis N, et al. Re-evaluating automatic metrics for image captioning[J]. arXiv preprint arXiv:1612.07600, 2016.

[263] Kryściński W, Keskar N S, McCann B, et al. Neural text summarization: A critical evaluation[J]. arXiv preprint arXiv:1908.08960, 2019.

[264] ANANYA B. SAI, et al., A Survey of Evaluation Metrics Used for NLG Systems, arXiv: 2008.12009v2

第十五章　情感计算研究进展、现状及趋势

15.1. 研究背景与意义

人类情感是人们相互交往中主动选择和创造的结果，它是通过特定的人类行为和符号来表现、传达和显示的。因此，“情感”实际上是社会意义和各种符号价值的载体与承担者。人类的认知，行为以及社会组织的任何一个方面几乎都受到情感的影响。1985 年，人工智能的奠基人之一 Minsky 就明确指出：“问题不在于智能机器能否有情感，而在于没有情感的机器能否实现智能”。但由于当时技术限制，赋予计算机或机器人以人类式情感的研究并未受到广泛关注。1995 年情感计算的概念由 Picard 首次提出，并于 1997 年正式出版《Affective Computing（情感计算）》。在书中，她指出“情感计算就是针对人类的外在表现，能够进行测量和分析并能对情感施加影响的计算”，开辟了计算机科学的新领域，其思想是使计算机拥有情感，能够像人一样识别和表达情感，从而使人机交互更自然。

简单来说，情感计算研究就是试图创建一种能感知、识别和理解人的情感，并能针对人的情感做出智能、灵敏、友好反应的计算系统。显然，情感计算是个复杂的过程，不仅受时间、地点、环境、人物对象和经历的影响，而且要考虑表情、语言、动作或身体的接触。因此，在智能人机交互的研究中，拥有对情感的识别、分析、理解、表达的能力也应成为智能机器必不可少的一种功能。例如：在管理行业，通过情感计算获得领导者与员工的情绪，从而提升企业的整体效率；在贸易方面，通过客户评价文本分析客户的情感进行精准促销，可以更精准的帮助企业树立自己的品牌；在健康领域，基于医患对答的情感预测可以帮助医生分析病人心理，辅助进行心理访谈，进而诊治心理疾病和平复自杀等消极情绪。为实现真正的人工智能，必须要实现融合智能与情感的自然人机交互。

此外，情感计算是一个多学科交叉的崭新的研究领域，它涵盖了传感器技术、计算机科学、认知科学、心理学、行为学、生理学、哲学、社会学等方面。情感计算的最终目标是赋予计算机类似于人的情感能力。要达到这个目标，许多技术问题有待解决。这些技术问题的突破对各学科的发展都产生巨大的推动作用。以下分别从情感认知、文本情感计算、多模态情感计算等领域的问题挑战、技术方法、发展趋势等对情感计算的研究进行探讨。

情感综合了行为、思想和感觉，是人们对待事物的表达方式。认知是人们对某个物体，对某件事情所理解的程度。因此，情感认知即个体对这种表达方式的认识程度和理解程度。情感认知的主要研究目标是通过外在情感信息（如面部表情、唇动、声音、姿势等）和内在情感信息（如心率、脉搏、血压、体温等）来识别和推断行动者的情感状态。情感认知的研究与发展不仅是人与人之间社会关系维系的重要课题，更是人机情感交互的关键。情感认知技术能够让机器感知到人们的情感状态，从而提高机器的人性化水平，在疾病和压

力识别、课堂反馈、安全驾驶和用户体验等多个领域都有广泛的应用。

文本情感计算的主要任务是研究自然语言中的主观信息（如情感、情绪、态度、评价等）的提取、分析、理解和生成。文本作为人类表达情感情绪的重要载体，文本情感计算是情感计算的一个重要组成部分，也是自然语言处理、文本挖掘等领域的重要内容。文本情感计算可以视为以主观信息为对象的自然语言处理技术。自然语言处理包含自然语言理解、自然语言生成、知识图谱等领域。同样地，文本情感计算也涵盖文本情感分析、情感文本生成、情感图谱构建、论辩挖掘等方面的研究，在舆情分析、心理健康监测、评论分析与生成、商业决策等方面有着广泛应用。

人类在表达情感时，通常以多种模态的方式呈现。单模态的情感分析并不符合人类对情感的感知与表达模式，当人类主观上对情感信号加以掩饰或者单一通道的情感信号受到其他信号影响时，情感分析性能将会明显下降。单模态信息量不足且容易受到外界各种因素的影响，如面部表情容易被遮挡、语音容易受噪声干扰。考虑到各个模态之间的情感表达的互补性，多模态融合的情感计算研究正日益受到重视。由于不同表现方式在表达情感信息时存在一定的互补作用，多模态情感分析更加完整，具有更好的鲁棒性，也更加符合人类自然的行为表达方式。近年来，学术界和工业界将目光转向多个模态信息融合的情感分析，利用各个模态信息之间能互补性得到性能更优的情感计算方法和系统。

15.2. 领域发展现状与关键科学问题

15.2.1. 情感认知

现阶段的情感认知研究主要集中在对面部表情、语音情感、生理信号的情感认知。

面部表情识别是指通过面部肌肉的变化识别特定的情感状态。由于面部表情是最容易控制的一种，而且受先天生理影响，单纯的面部表情识别准确性并不高，但是相应的识别模型则比较简单。如 Paul Ekm 等提出的面部动作编码系统(FACS)，描述了基本情感以及对应的产生的肌肉运动的动作单元。依据 FAcs 系统制造的面部识别器，仿真测试准确率可以达到 98%以上。但面部识别器的处理效率较低，对于处理连续表情还存在一定困难。目前大部分有关人脸表情的分析与识别主要针对基本表情的分析识别，使用的方法大致归为两类：基于静态图像（单一图像）的方法和基于动态图像序列的识别方法。

语音情感识别是指由计算机自动识别输入语音的情感状态。语言除了包含语义信息，还包含具有情感的语速、语调等信息。通过利用声学和语言学来描述说话方式的计算机应用程序“情感编辑器”，除了在输人情感参数之外还进行了语法语义的分析，对语音频率和音量进行控制，对语音形成较好的情感识别和合成效果。近年来，语音情感识别研究工作在情感描述模型的引入、情感语音库的构建、情感特征分析等领域的各个方面都得到了发

展。

生理信号情感识别是指通过内部的生理反应（如呼吸、心跳等）来识别情感状态。情感生理信号的研究重点在普适性理论研究，而后是个性化的研究，并且目前已经大多转向应用性的研究方面。但是，人的生理信号比起面部表情和语音，识别难度更大，所以目前生理模式的情感识别研究还处于初级阶段，哪些信号可以转化为情感参数、信号各个方面的权重、比例应该是多少，这些都还需要进行进一步的研究和探索。

情感认知的关键科学问题在于通过各类传感器获取由人类情感引起的生理指标或者行为特征发出的信号（例如语音、面部表情、手势、姿态、脑电波、脉搏等），以建立可计算的情感模型。

15.2.2. 文本情感计算

早期的文本情感分析技术主要针对文本情感的分类，其方法主要分为基于情感字典的规则化方法和基于情感特征的统计机器学习方法。传统情感分析方法在特定领域下构建情感词典，依据情感词与文本的映射关系能够实现快速自动情感分析。但由于细粒度、多领域及多方面自适应情感常识的缺乏，难以支撑多领域情感分析。

随着深度学习的深入发展，大量的神经网络模型被引入到情感分析任务中，包含卷积神经网络、循环神经网络、递归神经网络、注意力机制网络等。近年来，随着预训练语言模型的兴起，以 BERT 和 GPT 为代表的预训练语言模型在不同的情感分析任务中均取得了较大的成功。当前基于深度学习的情感分析方法依赖于大量高质量标注训练样本，人工标注成本昂贵，同样面临难以实现多领域及多方面自适应的实时在线情感分析的挑战。为了弥补情感计算依赖大规模标注数据、具有强领域特性的特点中，常常会引入外部的情感知识库提供监督信息，提高模型的泛化性能。然而，当前常采用的外部知识库存存在以下三个问题：（1）缺乏领域适应性：当前常用的情感词典常常只适用于某领域，缺乏领域泛化能力。（2）缺乏方面适应性：在同一领域中，同一情感词在不同方面的情感极性可能会不同。在现存外部知识库缺乏方面泛化能力。（3）缺乏情感推理能力：现存的情感词典以及外部知识库往往只建立词语与情感的一对一的映射关系，无法建模情感词间关系、方面词间关系，以及方面词与情感词的动态多关系。

情感文本生成任务的目标是让模型生成符合指定的情感类别的文本。具体而言，生成的文本应当表达出任务指定的情感类别，如开心、难过、愤怒等，这既可以通过情感相关的关键词体现（如开心与“享受”、难过与“哭泣”等），也可以通过隐喻等手法体现（如在难过的情感类别下，“我的心头阴霾不散”）。该任务的挑战有两点：（1）如何保证模型生成的文本语法正确、通顺连贯。（2）在保证语法性的前提下，生成文本应该蕴含指定的情感类别，并避免产生与指定情感类别矛盾的表述，以防造成歧义。

论辩分析是文本情感计算一个新兴的研究领域。近年来，计算论辩学研究将人类关于

逻辑论证的认知模型与计算模型结合起来，以提高人工智能自动推理的能力。论辩挖掘是计算论辩中的重要任务，以文本中包含论辩性内容的部分作为研究对象，旨在自动化识别论辩性文本的结构，论辩语义单元直接的逻辑交互关系等。论辩文本中往往呈现逻辑推理过程，因此语义结构复杂；其文本内容有高度的领域相关性，对于方法的领域迁移性提出了很高的要求；论辩文本体现了人类高级的认知能力，是对人类世界理解的综合运用，依赖于知识融合。

15.2.3. 多模态情感计算

多模态情感计算在模态融合方面，包括了基于特征层的融合、基于模型层的融合和基于决策层的融合。其中，基于模型层的融合策略得到了更多关注；在建模方法方面，随着深度学习技术的发展和数据资源的扩增，基于深度神经网络的多模态情感识别方法在学术界和工业界广泛应用，在建模过程中通过有效融合场景、个体差异、时序上下文等先验信息，进一步提升情感分析系统的性能；在应用场景方面，除了当前最为主流的情绪和倾向性分析，面向压力、精神状态、维度情感、专注度、言语置信度等多模态情感计算问题，近年来得到了广泛关注，在教育、安全、医疗、金融等领域有着广泛的需求。

当前多模态情感计算需要解决的科学问题主要包括;（1） 多模态情感计算数据库普遍面临着数据稀疏和标注不确定的问题，如何从这些低资源的数据中学习到有效、鲁棒的情感表征；（2）在模态信息缺失、互斥、冗余等条件下，如何设计高效的融合算法来整合不同模态的信息以提高多模态情感分析的准确率；（3）基于多模态信息对情感表达含义进一步理解，如何有效融合语义信息进行多尺度情感的准确理解，实现在认知层面的情感分析；（4）情感是随时间连续变化的，不同情感可能同时出现，如何基于多模态信息实现细微情感的准确表征。

15.3. 关键技术进展及趋势

15.3.1. 情感认知

当前人工智能情感认知模型所面临的最大挑战在于：要为情感认知找到适当的计算表征。在直觉理论中，人们使用两种或多种标记来区分情感是不够全面的，如“生气”与“不生气”、“开心”与“不开心”，尽管这种区分方式在许多情感分析中被广泛使用。事实上，定义表征空间是概率建模的一个重要前提，表征空间允许从情感中抽取样本并将其边缘化。即使是在某些高维度的空间中，其向量也可能是不充分的。

具体而言，比如关于“生气”可能存在三种场景：1）A 在生气；2）A 生气是因为他得知了一些不太好的结果；3）A 生气是因为不公平导致了不太好的结果。这三种场景在定性

上会存在不同，而这种不同则会导致对情感的评价也不尽相同，最终造成不同的行为后果。这种观点需要一种更为丰富的情感表征理论来解释目标相关信息和事件相关信息，事实上这是当前贝叶斯模型所没有涵盖的地方。因此，当前人工智能情感认知计算模型需要为情感认知及其评价选择一种适当的表征方式，用以获得对他人情感认知的理解，并可以进行有效地计算。

情感认知技术发展脉络主要集中在情感信息的获取和情感建模两个方面。情感信息的获取主要分为可以被自然观察到的情感信息（声音、手势和面部表情等）和需要特殊测试设备才能获取的情感信息（心跳速率、脉搏和温度等）。情感建模主要包括离散状态计算模型、情感空间计算模型和基于规则的模型。情感建模的技术从最普遍使用的 OCC 情感识别模型，逐渐完善衍生出基于事件评价的情感模型以及 EMA 模型等。现有关注问题主要是如何抽取有效的特征参数并运用恰当的模型来表达这些特征参数和情感之间的关联性。由于最终采集到的情感数据主要通过音频或者视频的形式进行储存和分析，因此目前主流方法主要从音频情感认知，视频情感认知和多模态情感认知三个层面分析。

音频情感认知的声学特征分析主要围绕韵律、频谱和音质特征。研究者已经发现很多声学特征与情感状态有关，如持续时间、语速、基音频率、共振峰、强度、Mel 频率倒谱系数（MFCC）等。研究人员将它们表示为固定维数的特征向量，其中的各个分量为各声学参数的统计值，包括平均值、方差、最大或最小值、变化范围等。近年来，神经网络提取优良特征参数的能力越来越受到关注。深度语音情感特征是基于语音信号或者频谱图，并通过语音情感识别相关任务学习到的深度特征。目前应用比较广泛的是通过语音事件检测或者语音情感识别等任务，采用在大规模的训练数据学习到的深度语音特征作为语音情感特征，比如 VGGish 和 wav2vec。

视频情感认知中局部二值模式（LBP）、局部相位量化特征（LPQ）、Gabor 特征被广泛应用于静态图像的情感识别工作中；时序信息为情感识别提供了关键信息，许多基于上述特征的时空特征，如 LBP-TOP。计算机视觉中常用的方向梯度直方图（HOG）描述子、尺度不变特征变换（SIFT）描述子、词袋模型（BoW）和 Gist 描述子均在情感识别工作中有所涉及。另一类是基于深度神经网络的深度情感特征。深度情感特征主要从人脸情感识别数据集上训练的模型中进行抽取，比如目前应用广泛的深度特征是从人脸情感识别数据集（比如 FER+）上训练的 VGGNet、DenseNet 等神经网络模型中抽取。

多模态信息的分析方法有很多，从信息融合层次来看，多模态信息融合的方法主要有决策层融合和特征层融合，也有一些学者将这两个融合方式混合使用。决策层融合方式操作方便灵活，允许各个模态采用最适合的机器学习算法进行单独建模。特征层融合的通常做法是将各个通道的特征相串联，组合成一个长的特征向量，然后再将该特征向量放入机器学习算法进行分类或是回归输出。最新的认知神经科学表明，大脑在整合多感官信息时存在多阶段融合的现象，受此启发，研究者提出了多阶段多模态情感融合方法。首先训练

一个单模态模型，然后将其隐含状态与另一个模态特征拼接再训练双模态模型，以此类推得到多模态模型。

情感认知计算的发展趋势主要体现在三个方面：首先，必须优先基于对自然主义数据进行认知建模，如静态面部表情和实验场景是我们研究的重要出发点。对于未来的研究工作来说，重要的是要观察在自然语境中如何对他人的情感进行建模，如观察某人无脚本的独白；其次，开展融合面部表情、语音、姿势、文本和生理信号等的多模态情感认知研究也必将是未来重要的发展趋势。多模态情感融合的关键在于实现了跨模态之间的有效整合以获得多模态信息的互补，从而比单模态情感识别具有更大的优势；最后，情感是一个时序变化的行为，其演变都会经历一定的时间，因此需要考虑情感信息的前后依赖性。在模型中引入注意力机制，通过全局上下文信息自动学习不同帧对于情感识别的重要性得到相匹配的权重系数，可以实现更有针对性的情感建模，显著提高情感识别的性能。

15.3.2. 文本情感计算

目前文本情感分析最关注的问题集中于属性级情感分析和情绪分析理解两个任务。前者是细粒度的情感分析，后者是除了情绪分析之外，还需对情绪原因进行理解和推理。属性级情感分析的主要研究任务包括属性抽取、属性级情感分类、属性情感配对抽取以及属性观点情感三元组抽取等。另一方面，情感文本生成的技术在早期大多基于 RNN 语言模型的方法。近年来，随着预训练模型的发展，情感可控的文本生成逐渐以 GPT 等预训练模型作为基座，并取得了更强大的效果。现有研究主要关注如何建模情感的表达过程、让文本生成受控于指定情感；以及如何丰富情感表达的方式和内容，以提高生成的多样性和信息量。

在属性抽取和属性级情感分类等传统方向上，主流方法包括基于卷积或循环神经网络的方法、基于注意力机制的方法以及基于图神经网络的方法；而在最新的属性情感配对抽取以及属性观点情感三元组抽取等方向上，主流方法包括基于机器阅读理解的方法、基于表格的方法以及基于 Seq2Seq 的生成式方法等。对于情绪分析于理解，主要的研究任务包括文本情绪分类、对话情绪识别、情绪原因抽取。在情绪分类以及对话情绪识别方向上，主流方法包括基于循环神经网络的方法、基于注意力机制的方法以及基于图神经网络的方法。在最新的情绪原因抽取方向上，主流方法包括基于卷积或循环神经网络的方法、基于自注意力机制网络的方法以及基于外部知识融合的图神经网络方法等。

情感文本生成的研究中，针对如何建模情感的表达过程这一问题，由于情感表达具有显性（如情感关键词）和隐性（如隐喻）的特点，情感表达也是一个动态的过程，因此现有研究大多采用将拷贝网络与动态记忆单元相结合的方式。一方面，拷贝网络可以显式地在生成文本中插入情感词，另一方面，动态记忆单元可以控制表达情感的过程，在已生成出表达情感的词语后，适时控制生成过程的结束。针对情感生成文本的多样性和丰富性问

题，由于模型的输入信息十分有限（只有指定的情感类别），因此现有研究大多利用外部知识丰富情感表达的内容。例如，通过在常识知识图谱检索与情感类别相关的实体（如难过与“分手”、“失业”等）来提升生成文本的信息量。

论辩挖掘的研究主要经过以下几个阶段：（1）理论迁移：对经典论辩理论的迁移和改造使其具备可计算的特点。（2）单体式论辩文本理解：研究论辩基本单元识别和关系分类方法，设计到不同领域的小规模语料标注。（3）交互式论辩文本理解：针对多人参与的论辩场景，研究文本分析框架以及论辩方法。（4）论辩文本自动生成：针对某一个特定主题或者其它用户的一段论辩性文本，自动化生成论辩内容。目前的研究热点为交互式论辩文本理解和论辩文本自动生成两个部分。在初期，学者们采用基于特征工程的论辩文本理解方法，近几年基于神经网络的文本编码解码框架开始成为主流。

文本情感计算的发展趋势主要体系在以下几个方面。现阶段以 BERT 和 GPT 为代表的预训练语言模型在不同的情感分析任务中均取得了成功，但是大部分工作仍是采用预训练加微调的范式。这种范式的缺陷在于语言模型在预训练过程中是脱离于下游情感分析任务的。为了解决此缺陷，基于提示（prompt）的学习的范式可能会成为一个比较有发展潜力的研究方向，如何针对下游不同的情感分析任务设计符合预训练语言模型训练目标的 prompt 是值得深入探究的问题。

情感文本生成未来技术发展有两方面的趋势。一是利用大型预训练模型内部的知识。在不引入外部信息的情况下，使得生成文本在情感可控的前提下更加多样、丰富。近期基于提示学习的方法展现出触发大模型内部知识的潜力，未来的情感文本生成的研究或许可以与提示学习方法相结合。二是高效地融合外部知识信息。外部知识信息往往能够提供更好的可控性。然而在基座模型越来越大的趋势下，传统的为小模型所设计融合外部信息的方法可能不再适用（受限于复杂度和效率），此时利用外部知识的方法需要更高的可拓展性。

此外，针对现有方法难以高效处理多领域及多方面自适应、情感常识离散、缺乏推理机制而难以进行情感推理等问题，其中的一个技术发展趋势是将情感词在多领域、多方面的动态情感倾向知识化。通过构建面向多领域多方面的情感知识图谱，利用知识图谱丰富的表达能力，可以实现领域细粒度情感知识化，通过情感常识关联整合、建模方面词和情感词之间的层级逻辑关系，形成情感知识图谱，有利于领域知识、方面知识及情感知识的动态关联、聚合以及推理，为情感计算的应用，如高效实时的在线情感分析、情感注入的对话系统、情感注入的故事生成等提供具有动态精准的领域自适应情感常识。

论辩挖掘的研究未来发展趋势主要包括：（1）不同场景和粒度的论辩性内容表示方法。从单一论点到论辩性段落再到同一主题下的多立场论点，到整个论辩性文本的知识库构建，这些都论辩性文本挖掘的核心问题，但相关的研究还很少。（2）大规模语料集合的构建。目前的论辩性文本研究很大程度上受到数据集合规模小、领域分散的限制，如何构建有标

注、无标注的大规模论辩性文本是一个重要课题。（3）论辩性文本生成机制和方法研究。相比叙述性文本，论辩性文本的产生更多的依赖于人类的逻辑推理能力，如何将推理方法融入到文本生成过程中对于论辩内容的自动生成至关重要。

15.3.3. 多模态情感计算

相对于单模态情感分析，多模态情感分析能够有效利用不同模态信息的协同互补，增强情感理解与表达能力。然而，受限于自然场景的复杂性和情感变化的多样性，多模态情感计算存在着诸多挑战：（1）多模态情感信息协同表征难，受限于部分模态信息缺失、跨模态信息不同步以及不同模态行为呈现的情感差异化等问题，制约了跨模态间情感信息的一致性抽取和呈现；（2）难以实现细粒度的多模态情感识别，当前主流多模态情感分析主要对正负倾向性或者基本情绪进行分类，难以有效对复杂细微情感进行准确跟踪，制约了对情感含义的准确分析；（3）针对多模态数据中的语义信息理解不充分，现有融合语义的情感分析，主要关注于文本中的语义信息，未能有效融合表情姿态和语气语调中的语义线索，影响了多模态语义信息的传递与理解；（4）面对碎片化、多源异构的跨模态海量数据，由于数据价值密度低，难以有效挖掘用户的隐藏情感；（5）标注多模态情感数据集成本高昂，缺乏高质量的标注数据，制约了多模态情感计算的落地应用。

在多模态融合策略方面，现有方法主要分为模型无关与模型依赖两种路线。前者不依赖于特定的学习算法，包含前期融合（特征级融合）、后期融合（决策级融合）、混合式融合三种策略。后者在构建学习模型的过程中显式地执行融合操作。对于浅层模型来说，常用的模型依赖策略包括基于核函数的融合和基于图的融合；对于近期流行的深层模型来说，则有基于神经网络的融合、基于注意力机制的融合、基于张量的融合等。随着 Transformer 架构和多模态预训练模型的兴起，当前主流的信息融合方法主要是基于模型的融合，并使用融合特征向量的方式去区分来自不同模态和信息源的特征，从而有效地建模这些复杂特征之间的关系。

在多模态情感识别建模方法方面，主要分为静态模型和动态模型，其区别在于模型是否具有建模情感时序上下文的能力。常用的静态模型包括支持向量机、高斯混合模型、AdaBoost、多层感知机等；为了组合不同分类器的优点，多分类器系统也在多模态情感识别领域得到了探索。情感是一个时序变化的行为，需要考虑情感信号的前后依赖性。传统的动态模型如隐马尔科夫模型和条件随机场，由于其可以对时序上下文信息建模的内在属性，取得了比静态模型更好的识别性能。然而这些模型考虑的前后时序信息较短，因此取得的效果有限。随着深度学习技术的发展和数据资源的扩增，基于深度神经网络的多模态情感识别方法得到了广泛关注，这类方法不仅可以学习到数据的深层非线性特征表示，而且能够有效处理情感的时序特性，在建模过程中通过有效融合场景、个体差异、时序上下文等先验信息，能够显著提升多模态情感识别的性能。在识别任务方面，当前主要任务是

情绪或倾向性分类，面向压力、精神状态、维度情感、专注度等复杂情感的识别，近年来得到了广泛关注。

目前主流的多模态情感生成方法是首先根据文本及其蕴含的情感合成语音，然后根据合成的语音及其蕴含的情感以及目标参考视频，生成人像视频。研究者们提出利用人脸特征点作为中间表示的两阶段方法，首先根据语音内容和语音情感信息生成人脸特征点的序列，然后进行空间、动作、情感表达的对齐，最后结合参考视频生成最终的人像视频。为了进一步增强情感表达的自然度，引入情感强度表征、情感强度排序等情感强度建模方法，对生成的情感强度进行控制；通过引入情感转移矩阵或时间序列建模等方法，可以使交互系统情感转移更平滑，进而获得稳定的情感表达。

在应用场景方面，早期的多模态情感计算应用主要在实验室条件下进行，如通过生理信号进行情绪监测，或通过学生的面部情感识别反应教学质量。随着技术的发展、设备的更新和数据的扩增，多模态情感计算应用逐步延伸至实际场景，如通过可穿戴设备记录多模生理信号，应用于自闭症治疗，也有研究人员构建具备一定情绪反馈能力的机器人，用于儿童陪伴和教育。此外多模态情感计算也广泛应用于网络舆情分析，识别评论中蕴含的情绪，以反映公众态度，从而获取信息以了解其演变过程。

多模态情感计算的发展趋势集中体现在以下三个方面。首先是如何融合语义信息进行多尺度情感准确理解，分别从倾向性、情绪状态、心理压力、精神状态、专注度等多个维度进行多模态情感分析，实现从情感感知到情感认知的跨越。第二个趋势是增强复杂环境下情感计算的鲁棒性，实现在非协作开放模式下，面向高维碎片化开源数据，实现目标对象情感状态的精准识别；与预训练及多任务联合训练等方法结合，实现更广泛场景下的多模态情感计算；第三个趋势是探索通用的多模态情感计算模型，通过适配多场景应用，实现多模态情感计算应用零成本迁移。加强情感计算的个性化表达能力，适配不同个体的情感状态，融入用户画像、人格特质等个性化特征，实现对不同对象情感的准确理解，满足个性化的情感计算需求，实现与人共情的突破。

15.3.4. 情感计算资源

情感计算作为人工智能重要的分支之一，经过多年发展沉淀，积累了大量具有实际意义且富有研究价值的任务及相关资源，可分为粗粒度情感分析，细粒度情感分析，隐式情感计算，图文视频类多模态情感分析以及生理信号类多模态情感分析这五类资源。

粗粒度情感分析主要用来判断文本整体情感倾向，表明一个人对某件事或对某个物体的整体评价，分析的粒度可以是文本的篇章、段落或句子。典型的句子级情感分析语料有斯坦福大学发布的 SST 数据集(The Stanford Sentiment Treebank)，主要是针对电影评论做句子级别的情感分类。对于粗粒度情感分析，早期的工作主要借助构建情感词典来进行。例如，段落篇章级情感分析主要是针对某个主题或事件进行情感倾向判断，一般需要构建对应事件的情感词典，如电影评论的分析，需要构建电影行业自己的情感词典；句子级的

情感分析则大多通过计算句子里包含的所有情感词的值来得到。较为广泛使用的情感词典有 SenticNet，其提供了一组语义、情感、极性关联的十万个自然语言概念，包含了一系列将常识推理、心理学、语言学和机器学习相结合的情感分析工具和技术。

相比之下，细粒度情感分析更加深入到每个句子里的具体评价对象中，分析其对应的情感极性，下面将称为方面词。细粒度情感分析基准数据集主要来源于国际语义评测 SemEval 和中文倾向性分析评测 COAE。SemEval 发布的数据集包括 Restaurant 和 Laptop 两个领域，分别标注了方面词项、方面类别、观点词项、情感类别。近年来有研究者对 SemEval 数据集进一步标注了隐式属性和隐式观点，构建了完整的情感标注体系。COAE 发布的数据集涉及电脑、手机等领域，分别标注了情感词及其情感极性，以及方面词项及其情感极性。除以上提及的主要数据集外，还有 Citysearch corpus、BeerAdovacat、Twitter 等英文数据集，以及国际自然语言处理与中文计算会议 NLPCC 发布的多方面多情感数据集 MAMS。以上数据集的数据来源主要为产品评论，且中文数据集的规模、标注规范性和完整性还滞后于英文数据集。未来可将数据来源扩展至微博等社交媒体平台，构建跨领域、多模态的方面级情感分析数据集，并进一步规范标注体系。

隐式情感表达定义为“表达主观情感但不包含显式情感词的语言片段”。据统计，汉语中约有 15%-20%的语句采用客观陈述或借助修辞手法的方式来隐式地表达情感信息。相较于显式情感表达，隐式情感语料库的构建更具挑战。目前面向隐式情感计算的语料库构建尚处于起步阶段，但已受到许多研究者的高度关注。近年来由山西大学主办的 SMP-ECISA 隐式情感分析评测吸引了众多企业、高校参加，相应的评测语料也是现今使用最多的隐式情感计算数据集。此外，由于隐式情感与一些下游任务关联紧密，相关语料也同时标注了隐式情感以辅助相关任务的识别与分析过程，如隐喻计算、幽默计算等。因此，针对隐式情感计算的语料构建工作既能完善情感分析领域的研究方向，也可推动文本表示学习、文本语义理解等领域研究的发展。

随着技术的快速发展和信息的日益丰富，人们的情感表达方式逐渐多样。如何分析图文、视频等多模态数据中的情感已成为当前情感分析领域的机遇和挑战。针对多模态情感分析的迫切需求，卡耐基梅隆大学提出了一个大规模的多模态情感分析数据集 CMU-MOSEI，其中包含了来自 YouTube 的 3228 个自拍视角独白视频，具有清晰面部的表情。同时，数据集还包含了对应的 23453 条字幕文本，以及 COVAREP 抽取的声学特征等丰富信息作为补充特征。在标签方面，CMU-MOSEI 数据集具有情感、情绪两种标签，并对每种标签的情绪强弱进行衡量，从而可以支撑细粒度的情感分析任务。

目前主流的生理信号类多模态情感计算资源主要采用音、视频刺激方法诱发情绪，同步采集多模态生理信号，进而分析不同情绪下中枢神经系统和自主神经系统的反应，以实现基于多模态生理信号的情感识别。典型计算资源包括 DEAP、DECAF 等数据集。DEAP 数据集记录了 32 名被试在观看音乐视频片段时的 32 导联脑电、皮肤电、呼吸、皮肤温度、

心电、肌电、血容量脉冲、眼电等信号。所有信号采样率均为512Hz，脑电信号共有32导联。被试所观看每段视频时长约1分钟，共有40个视频片段。观看视频后，会根据自身感受从唤醒度、效价、喜欢或不喜欢、支配性和熟悉度等维度进行评分。考虑到被试个体的性别、年龄等因素均会对情绪激发产生重要影响，在未来的研究过程中，有必要深入考虑相关人口统计学信息的引入和建模。

15.4. 情感计算产业发展

近年来，情感计算在产业界得到了广泛的关注，各大企业不断挖掘其实际商业价值，在舆情分析、消费决策、个性化推荐、智能客服等领域均有广泛的应用。特别是2018年以来，基于预训练的语义理解获得了迅猛的发展，如Elmo、BERT、GPT、ERNIE等显著提升了包括情感计算任务在内的各类自然语言处理任务的效果。这也进一步推动了情感计算在业界的应用宽度，情感计算技术正大规模应用于政府、商业、金融、教育、医疗、娱乐等各类型行业。与此同时，企业也越来越重视在情感计算相关领域的学术研究，投入了大量的研究和工程力量，取得了丰富的研究和应用成果。近几年，在ACL、EMNLP、COLING、NAACL等会议上论文的统计，可以看到包括百度、腾讯、阿里巴巴、华为、科大讯飞、新浪等众多公司的学术研究成果。此外，企业发挥自身优势，建立了大量的校企实践基地、联合实验室等进行人才培养和技术储备，同时利用企业资源开放情感计算大模型平台、构建各类情感计算垂直任务的标准数据集等，取得了良好的效果。

当前，很多企业非常重视情感计算领域的人才和技术储备，设立了大量的校企联合实验室、实践培训基地等进行人才培养、项目研发以及应用落地。在这个过程中，企业结合自身优势扮演了重要的角色。

（1）海量情感计算的资源共享

相比于以高校和研究院为代表的学术界，企业拥有海量的数据资源，这为后续开展学术研究以及产品落地奠定了基础。以上海蜜度信息技术有限公司为例，据不完全统计，该企业每日可获取的数据包括文本内容覆盖全网信源，每日新增近2.5亿条全网公开评论内容；图片内容覆盖全网信源，每日新增近2.8亿张全网公开图片内容；视频内容包括每日新增数百万重点视频内容。同时，企业利用自身优势可以快速对上述数据进行垂直分类，并针对情感计算具体应用进行信息标注，并结合情感计算评测任务进行数据共享。

（2）面向情感计算的预训练模型研发

相比于通用预训练中主要关注事实型文本（如新闻、百科等），情感分析更侧重于分析主观型文本中蕴涵的情感和观点，因此很多企业利用自身算力优势，研发了专门面向情感分析研发情感预训练模型。例如，百度公司研发了一套情感预训练模型SKEP（Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis）。同时，为了更好地赋能整个行业，

百度开放了基于 SKEP 预训练的情感倾向分析、评论观点抽取、实体级情感分析等服务。目前，该平台已累计支持近10万用户，成为在情感分析领域技术布局最全面，业界使用最广泛的服务平台之一。

（3）多模态情感计算的研发引领

近几年，随着短视频以及Vlog等的盛行，更多的企业开展了面向多模态情感计算的研究，例如，北京知微公司围绕政企客户需求，探索文本、图片、音视频多模态情感计算技术及应用；拓尔思利用不同信源的多种模态信息进行跨语言跨媒体的事件追踪分析；华为研发了一套云情感计算工具，集成了包括文本，语音，图像，多模态情绪识别和情感合成等技术；腾讯音乐结合音频语文本进行细粒度情感属性抽取，并对用户进行音乐的精准推荐。

情感计算技术正进入到产业不断拓宽和研究深入融合的阶段。近几年，各行业对智能化业务处理要求的上涨，加速了情感计算技术与传统行业的融合。**在政府政务领域**，情感计算技术在文本内容敏感性研判、政策影响调研、上访事件处理、社会情绪洞察、以及智能决策支持等众多方面有效的辅助政府工作者提升政务处理效率和准确性；**在商业领域**，情感计算技术在企业口碑评价、产品细粒度抽取、产品精准推荐、消费辅助决策、以及商战情报分析等方面有效的辅助目标企业不断提升产品质量，提高企业形象；**在金融领域**，情感计算技术为量化投资提供了热点事件挖掘、舆情分析、情感属性细粒度抽取等多项重要因子，基于内容的用户画像、标签抽取等技术为大数据风控提供了重要支持，此外智能客服等产品也广泛应用于整个行业；**在教育领域，**情感计算技术在智能助教、学情分析、教学分析、教评考核等环节都发挥了重要作用，为全方位提升教育教学的智能化程度提供了重要的技术支持；**在医疗领域**，医疗决策支持、抑郁症患者心理引导、自闭症儿童筛查、病例只能抽取、问诊机器人等方面也深度集成了情感计算技术。随着情感计算技术的不断发展，企业的应用领域不断增加，服务范围进一步拓宽，并逐步满足用户实际需求。

经过这些年的发展，情感计算技术已经被越来越多企业应用，成为企业的核心竞争力，也有越来越多的企业通过校企合作项目、联合人才培养等形式在情感计算技术上进行大量储备，情感计算技术在性能上取得了突破，在实际应用中不断落地，在未来仍然拥有巨大的发展潜力。

15.5. 总结与展望

情感计算的研究可以为传统计算机（包括应用现有智能计算方法的计算机）增添具有感性思维的情感。可以认为，结合情感认知的情感计算是在人工智能理论框架下的一个质的进步。因为，基于情感认知的面部表情、语音情感、文本情感以及生理信号等情感研究，能赋予计算机拟人化的思维方式。从广度上讲它扩展并包容了情感智能，从深度上讲情感

智能在人类智能思维与反应中体现了一种更高层次的智能。因此，通过深入研究情感认知能促进情感计算的研究，从而为计算机的未来应用展现一种全新的方向。同时，由此引发出来的理论与应用问题会层出不穷。

目前，在文本情感计算的文本情感分析、情感文本生成、情感图谱构建等方面的研究已得到了广泛的关注。同时，大部分文本情感计算也被应用于基于社交媒体的舆情监测和治理，主要体现为情感量化和引导在社交媒体上的应用，可以看做是文本情感分析完成之后的统计和融合，再或者是基于启发式规则的简单处理，比如分析广大网友的情绪动态、监测具有特定情感倾向的。与此同时，伴随着深度学习技术的发展，基于情感知识的推理逐渐开始兴起。这方面的例子包括用户/产品评论的分析/生成、融入文本情感信息的推荐系统建模及其可解释性生成。这类应用基于带有情感的知识试图实现面向目标任务的推理，实现一些较为复杂的决策过程。相信，伴随着情感分析性能的提升，基于情感知识实现推理和决策将会是未来文本情感计算的一个重要方向，具有较大的发展空间和潜力。

虽然视觉、语音、文本等均能独立地表示一定的情感，但人的相互交流却总是通过信息的综合表现来进行。因此，多模态的情感分析更符合人类对情感的感知与表达模式。目前对多模态情感计算的研究主要集中于在情感识别和理解的方法上运用了模式识别、人工智能、语音和图像技术的大量研究成果，从而将不同模态的特征信息跟情感计算结合起来。然而，受到情感信息捕获技术的影响，以及缺乏大规模的情感数据资源，有关多模态特征融合的情感理解模型研究还有待深入。例如，融合语义信息进行多尺度情感准确理解、增强复杂环境下情感计算的鲁棒性、探索通用的多模态情感计算模型、等。这些技术的完善能进一步推动多模态情感计算的研究与发展。

基于对先进的情感计算技术开发，目前在情感计算领域的研究已取得了令人瞩目的成功。但是，受到情感信息捕获技术的影响，以及缺乏大规模的情感数据资源，有关提取更有效、更精确的情感特征，并将不同模态特征进行融合的情感计算模型研究还有待深入。展望未来，解决多模态情感分析问题需要更丰富的模态信息积累及不同模态之间的细粒度对齐，这无疑对于多模态信息的提炼与整合提出了更高的要求。在未来的研究过程中，有必要深入考虑相关人口统计学信息的引入和建模。与此同时，人类的情感远不止积极/消极等几种表现，如何设计更立体的情感标签与更丰富的模态信息，以更全面地涵盖人类的情感表现，无疑也是值得思考的问题。