中文信息处理发展报告
(2016)

中国中文信息学会

中国 • 北京

2016.12

前言

当前已经进入以互联网、大数据和深度学习为标志的海量信息时代，互联网和机器学习技术的快速发展对中文信息处理提出了许多新的挑战。

《中文信息处理发展报告（2016）》是中国中文信息学会召集专家对本领域学科方向和前沿技术的一次梳理，我们的定位是**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人士简要介绍相关领域的基本概念和应用方向，向高校、科研院所和高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。

编撰《中文信息处理发展报告》的想法来源于中国中文信息学会主办的“中文信息处理战略研讨会”。2006 年 11 月 20 日召开的中国中文信息学会第六届理事会第一次常务理事会上，常务理事们建议择期召开“中文信息处理战略研讨会”，共同探讨中文信息处理未来的研究方向和发展战略。中国中文信息学会于 2007 年 4 月 20 日在广西南宁召开“第一届中文信息处理战略研讨会”，之后于 2012、2014、2016 年分别于江西婺源、贵州贵阳、海南海口连续召开了“中文信息处理战略研讨会”，在与会各位专家的热情支持、积极参与、认真准备和共同努力下，这几次会议都取得了圆满成功，促进了本领域前沿技术的创新与发展！

在这几次战略研讨会上与会专家学者从学科发展趋势、国内外研究热点、未来重要应用、产业融合与发展等多个角度，给出了各自独到的见解和思考，会上也就我国中文信息处理未来的发展机遇和挑战进行了热烈的研讨。

尽管每次会后我们都会将专家的报告在学会网站分享以飨读者，但是因为比较零散，无法形成系统的观点。在今年的战略研讨会上，我们提出了发布《中文信息处理发展报告》的工作计划，由学会学术工作委员会主任马少平教授负责。

2016 年 5 月 12 日在中科院软件所召开了工作会议（马少平、孙乐、宗成庆、赵军、张敏、张家俊、韩先培、刘康等），经讨论确定报告分为基础研究和应用研究及各自具体的研究方向，也确定了撰写的模板，主要包括：方向定义及研究目标、关键科学问题和研究内容、研究方法及国内外现状、总结及展望等，宗成庆研究员提供了机器翻译方向的模板供大家编撰时参考。会后我们邀请各个方向的著名专家撰写了各个方向的报告：

基础研究：

词法与句法分析：李正华、陈文亮、张民（苏州大学）

语义分析：周国栋、李军辉（苏州大学）

篇章分析：王厚峰、李素建（北京大学）

语言认知模型：王少楠，宗成庆（中科院自动化研究所）

语言表示与深度学习：黄萱菁、邱锡鹏（复旦大学）

知识图谱与计算：李涓子、侯磊（清华大学）

应用研究：

文本分类与聚类：涂存超，刘知远（清华大学）

信息抽取：孙乐、韩先培（中国科学院软件研究所）

情感分析：黄民烈（清华大学）

自动文摘：万小军、姚金戈（北京大学）

信息检索：刘奕群、马少平（清华大学）

信息推荐与过滤：王斌（中科院信工所）、鲁骁（国家计算机网络应急中心）

自动问答：赵军、刘康，何世柱（中科院自动化研究所）

机器翻译：张家俊、宗成庆（中科院自动化研究所）

社会媒体处理：刘挺、丁效（哈尔滨工业大学）

语音技术：说话人识别——郑方（清华大学）、王仁宇（江苏师范大学），语音合成——陶建华（中科院自动化研究所），语音识别——王东（清华大学）

文字识别：刘成林（中科院自动化研究所）

多模态信息处理：陈晓鸥（北京大学）

医疗健康信息处理：陈清财、汤步洲（哈尔滨工业大学）

少数民族语言信息处理：吾守尔•斯拉木（新疆大学）、那顺乌日图（内蒙古大学）、海银花（内蒙古大学）等

最后由张敏（清华大学）、韩先培（中国科学院软件研究所）、张家俊（中科院自动化研究所）、刘康（中国科学院自动化研究所）等对初稿反馈意见，校对统一成文。

由于时间仓促，难免有疏漏，甚至错误的地方，供有志于中文信息处理事业的同仁和青年学者们参考，进行更广泛的讨论和思考，期待在我们的共同努力下再创中文信息处理事业新的辉煌！

中国中文信息学会

2016 年 12 月

目录

第一章 词法和句法分析研究进展、现状及趋势

1.任务定义、目标和研究意义

自然语言处理中的自然语言句子级分析技术，可以大致分为词法分析、句法分析、语义分析三个层面。

第一层面的词法分析（lexical analysis）包括汉语分词和词性标注两部分。和大部分西方语言不同，汉语书面语词语之间没有明显的空格标记，文本中的句子以字串的形式出现。因此汉语自然语言处理的首要工作就是要将输入的字串切分为单独的词语，然后在此基础上进行其他更高级的分析，这一步骤称为分词（word segmentation 或 tokenization）。除了分词，词性标注也通常认为是词法分析的一部分。给定一个切好词的句子，词性标注的目的是为每一个词赋予一个类别，这个类别称为词性标记（part-of-speech tag），比如，名词（noun）、动词（verb）、形容词（adjective）等。一般来说，属于相同词性的词，在句法中承担类似的角色。

第二个层面的句法分析（syntactic parsing）是对输入的文本句子进行分析以得到句子的句法结构的处理过程。对句法结构进行分析，一方面是语言理解的自身需求，句法分析是语言理解的重要一环，另一方面也为其它自然语言处理任务提供支持。例如句法驱动的统计机器翻译需要对源语言或目标语言（或者同时两种语言）进行句法分析；语义分析通常以句法分析的输出结果作为输入以便获得更多的指示信息。

根据句法结构的表示形式不同，最常见的句法分析任务可以分为以下三种：(1) 短语结构句法分析（phrase-structure syntactic parsing），该任务也被称作成分句法分析（constituent syntactic parsing），作用是识别出句子中的短语结构以及短语之间的层次句法关系；(2) 依存句法分析（dependency syntactic parsing），作用是识别句子中词汇与词汇之间的相互依存关系；(3) 深层文法句法分析，即利用深层文法，例如词汇化树邻接文法（Lexicalized Tree Adjoining Grammar，LTAG）、词汇功能文法（Lexical Functional Grammar，LFG）、组合范畴文法（Combinatory Categorial Grammar，CCG）等，对句子进行深层的句法以及语义分析。

上述几种句法分析任务比较而言，依存句法分析属于浅层句法分析。其实现过程相对简单，比较适合在多语言环境下的应用，但是依存句法分析所能提供的信息也相对较少。深层文法句法分析可以提供丰富的句法和语义信息，但是采用的文法相对复杂，分析器的运行复杂度也较高，这使得深层句法分析当前不适合处理大规模数据。短语结构句法分析介于依存句法分析和深层文法句法分析之间。

自然语言处理的第三个层面是语义分析（semantic parsing）。语义分析的最终目的是理解句子表达的真实语义。但是，语义应该采用什么表示形式一直困扰着研究者们，至今这个问题也没有一个统一的答案。语义角色标注（semantic role labeling）是目前比较成熟的浅层语义分析技术。基于逻辑表达的语义分析也得到学术界的长期关注。

出于机器学习模型复杂度、效率的考虑，自然语言处理系统通常采用级联的方式，即分词、词性标注、句法分析、语义分析分别训练模型。实际使用时，给定输入句子，逐一使用各个模块进行分析，最终得到所有结果。近年来，随着研究工作的深入，研究者们提出了很多有效的联合模型，将多个任务联合学习和解码，如分词词性联合、词性句法联合、分词词性句法联合、句法语义联合等。联合模型通常都可以显著提高分析质量，原因在于：联合模型可以让相互关联的多个任务互相帮助，同时对于任何单任务而言，人工标注的信息也更多了。然而，联合模型的复杂度更高，速度也更慢。

本章集中讨论第一和第二层面的词法和句法分析技术。

2. 研究内容和关键科学问题

词法分析是将输入句子从字序列转化为词和词性序列，句法分析将输入句子从词序列形式转化为树状结构，从而刻画句子的词法和句法结构。目前，学术界和产业界主要研究数据驱动的分析方法，即在人工标注的分词、词性语料和树库上自动训练构建词法和句法分析系统。数据驱动的方法主要优势在于给定训练数据，不需要太多的人工干预，就能得到最终的系统。但是给定一个句子，可以产生数量众多符合词法和句法的分析结果。如何从中找到正确的分析结果是最主要的研究内容。

词法分析主要面临如下几个问题：

- 词的定义和生词问题：什么是词？词的定义标准是什么？这在语言学界和计算语言学界争论多年，到至今还没有一个统一的标准。由于汉语构词非常灵活，特别是在互联网时代，外来语、新词、热词不断出现，事实上，也不存在一个绝对统一的构词标准和分词规范。汉语的词是开放、动态的，不可能用一部静态词典包含所有的词。所以，用来描述生词和构词法的模型是非常重要的。
- 分词歧义问题：分词歧义是指在一个句子中，一个字串可以有不同的切分方法。例如，“乒乓球拍卖完了”，可以切分为“乒乓/球拍/卖/完/了”，也可以切分为“乒乓球/拍卖/完/了”。即使给定词的定义标准和一部覆盖面很广的词典，分词歧义问题也非常难解决，需要上下文语义知识的帮助才能解决。分词歧义进一步和生词问题交叉在一起，分词问题变得就更加复杂。
- 词性定义和词性兼类问题：词性类别远比词的个数要小，但词性的定义也不完全存在一个统一的信息处理用的国内和国际标准。词性兼类问题是词性标注面临的主要问题，需要更高层次的上下文信息来解决。
- 句法分析主要面临如下四个关键问题：
- 模型定义问题：如何为各候选句法树打分。由于符合语法规则的句法树数目非常多，因此要对每棵树进行评估计算它的分值。分值高低体现了该树是正确树的可能性大小。本项内容是研究如何将句法树的分值分解为一些子结构的分值。
- 特征表示问题：如何表示句法树。在模型定义中，句法树已经被分解成一些子结构。这些子结构如何被机器学习模型所识别，也就是特征表示问题。本项内容是研究采用哪些特征来表示每一部分子结构。
- 解码问题：如何寻找概率（或分值）最高的句法树。在给定所有子树的分值后，通过组合可以得到数目众多的不同分值树，搜索空间较大，无法通过简单比较得到分值最高的结果。本项内容是研究如何设计有效算法高效地搜索到分值最高的句法树。
- 训练算法问题：如何训练获取特征权重。在句法分析中通常有数以千万计的特征，这些特征的重要性存在差异，因此需要去学习它们的重要程度，即特征权重。本项内容主要是研究如何使用机器学习模型来有效的学习特征权重。

3. 技术方法和研究现状

本节分别介绍分词、词性标注和句法分析所用的主流技术方法和研究现状。

3.1 分词

汉语分词任务的目标是将输入的句子从汉字序列切分为词序列。分词是汉语处理的重要基础。在过去三十多年里，经过研究者们的不断摸索，汉语分词研究取得了全方位的发展。尤其是大规模人工标注数据的产生和基于人工标注数据的统计方法取代了基于词表和规则的方法，分词准确率取得了显著提升，分词模型也变得更加简单有效。2003 年国际中文分词公开评测任务开展以来，中文分词也吸引了更多研究者的关注。

3.1.1 主要分词方法

基于词典的最大匹配分词方法：1986 年，刘源、梁南元首次将最大匹配方法应用到中文分词任务。根据方向不同，最大匹配方法又可以分为前向和后向最大匹配方法两种。最大匹配方法是最有代表性的一种基于词典和规则的方法，其缺点是严重依赖词典，无法很好地处理分词歧义和未登录词。然而，由于这种方法简单、速度快、且分词效果基本可以满足需求，因此在工业界仍然很受欢迎。

全切分路径选择方法：其思想是所有可能的切分表示为一个有向无环图，每一个可能的切分词语作为图中的一个节点。有向图中任何一个从起点到终点的路径构成一个句子的词语切分，路径数目随着句子的长度指数增长。这种方法的目标是从指数级搜索空间中求解出一条最优路径。张华平、刘群(2002)最初根据 n 元语言模型及其他大规模统计信息等对每个节点和边赋予一定的权重。而 Andrew (2006)利用半马尔科夫条件随机场，基于人工标注数据的统计方法，对每个节点进行打分。

基于字序列标注的方法：对句子中的每个字进行标记，如四符号标记{B，I, E, S}，分别表示当前字是一个字的开始、中间、结尾，以及独立成词。这种方法首次由 Nianwen Xue et al.(2002)提出，之后研究者们尝试使用不同的序列标注模型，如最大熵、SVM、结构化感知器、CRF 等，不断提高分词效果。目前基于序列标注的方法在学术界仍然是分词主流方法。

基于转移的分词方法：这种方法借鉴了基于转移的依存句法分析的思路，从左到右扫描句子中的每一个字，将分词过程转化为一个动作{append, separate}序列，使用柱搜索获得最优动作序列(Zhang and Clark, 2007)。和基于序列标注的方法相比，基于转移的方法可以更灵活的融入各种特征，特别是基于词的特征，因此在学术界受到越来越多的关注。

3.1.2 分词主要研究进展

有效的特征集合：经过研究者们的摸索，对于中文分词模型，无论是基于字序列标注的方法还是基于转移的方法，都形成了一套有效稳定的特征集合，如 n 元字串、字的类别、叠字现象、偏旁部首作为形态信息等等。基于这些特征集合，后续研究者们可以很快实现出效果很好的分词系统。

基于词典的特征： 分词过程中，可以把“当前字开始的三个字构成的字串是否在词典中出现”这样的信息作为特征，加入到统计模型中，这种信息称为基于词典的特征。使用基于词典的特征，实际上是将基于词典的规则系统和基于统计的分词方法进行了软融合。研究者们发现，在处理跨领域文本时，如果有比较好的领域词典，基于词典的特征可以显著提高分词准确率(Pi-Chuan Chang et al., 2008; 张梅山等, 2012)。在 NLPCC-2015/2016 会议上组织的微博分词任务上，评测结果也同样验证了词典特征的有效性。

基于无标注数据的半指导特征：如何从大规模无标注数据中获得帮助，一直是研究界非常感兴趣的研究方向。近年来，研究者们探索出了很多有效的基于无标注数据的半指导特征，如两个字串之间的互信息（mutual information），一个字串左右邻接字的多样性（accessor variety），一个字串左右邻接标点符号的频率，字串在篇章中出现频率，汉字的左右边界熵，两个汉字的卡方统计量等 (Weiwei Sun and Jia Xu, 2011; 韩东煦、常宝宝, 2015)。研究表明，这些半指导特征可以显著提高分词准确率，尤其在领域移植的场景中处理有别于训练数据的文本时。

基于自然标注数据的学习方法：网页源文本中包含了大量的 html 标记，指定了文字在网页中的角色、超链接、显示位置或显示格式，而这些标记无形中也隐含了分词边界信息。研究者们将这种隐含的分词边界信息称为自然标注，将包含自然标注信息的文本转化为局部标注数据，加入到模型训练数据中，显著提高了分词效果 (Wenbin Jiang et al., 2013; Yijia Liu et al., 2014)。

基于异构标注数据的学习方法：汉语数据目前存在多个人工标注数据，然而不同数据遵守不同的标注规范，因此称为多源异构数据。近年来，学者们就如何利用多源异构数据提高模型准确率，提出了很多有效的方法，如基于指导特征的方法、基于部分词（subword）的方法、基于成对序列标注的方法。

基于深度学习的分词方法：近几年，深度学习方法为分词技术带来了新的思路，直接以最基本的向量化原子特征作为输入，经过多层非线性变换，输出层就可以很好的预测当前字的标记或下一个动作。在深度学习的框架下，仍然可以采用基于子序列标注的方式，或基于转移的方式，以及半马尔科夫条件随机场。深度学习主要有两点优势:1）深度学习可以通过优化最终目标，有效学习原子特征和上下文的表示；2）基于深层网络如 CNN、RNN、LSTM 等，深度学习可以更有效的刻画长距离句子信息。

词法句法一体化建模：随着计算资源的飞速发展和对机器学习模型的理解更加深入，研究者们提出了很有效的统计模型，直接从字开始对句子进行分析，输出分词、词性、句法的结果。多年前也有研究者提出词法句法一体化分析，如最有代表性的 NLPWin，但是均采用基于规则的方法。基于统计模型的一体化建模可以让词法句法分析互相影响，互相提高，显著提高了词法和句法的分析效果。尤其值得提出的是，研究者们提出进一步分析词语内部结构，有效缓解了数据稀疏问题。

国际公开评测任务：自从第一届 SIGHAN Bakeoff 2003 开始 ，国内学术界多年来不断组织针对分词的国际公开评测任务，组织方提供由多个机构提供的公开训练和评测数据集，吸引了国内外研究机构、大学和公司参赛，吸引了大量研究者从事相关研究，极大的促进了分词技术的发展。近年来，分词评测任务的关注点已经由传统的规范文本上的分词方法转向面向非规范文本和跨领域文本上的分词方法。2010 年 CIPS-SIGHAN 将测试语料分为四个领域：文学、计算机、医药、金融，引导研究者们对分词的领域迁移问题进行研究。2012 年 CIPS-SIGHAN 首次针对微博文本开展分词评测。2015 年 NLPCC 会议也开始组织面向微博文本的分词和词性标注联合评测任务。修驰 (2013, 表 1-5)对历届 SIGHAN 评测及相关数据给出了一个完整的总结。

分词开源软件开放：近年来，随着国内科研水平的提高，国内学术界纷纷开放研究相关的代码、数据，供其他研究者使用。其中，据笔者所知，影响较大、使用人数较多的几个分词系统包括**中科院计算所的 ICTLAS 分词系统**、**哈工大语言技术平台 LTP**、**清华大学自然语言处理工具包**、**海量云分词**等。

3.1.3 分词目前面临的主要挑战

分词歧义消解：分词歧义是指在一个句子中，一个字串可以有不同的切分方法。例如，“乒乓球拍卖完了”，可以切分为“乒乓/球拍/卖/完/了”，也可以切分为“乒乓球/拍卖/完/了”，类似的例子还有“门把手弄坏了”。虽然基于人工标注数据的统计方法能够解决很大一部分分词歧义，然而当面临一些训练语料中没有出现过的句子（或子句）时，基于统计的方法可能会输出很差的结果。

未登录词（新词）识别：**未登录词（out-of-vocabulary, OOV）指未在训练数据中出现过的词，而新词指日常生活中人们新创的一些词（也可能是旧词新意）**。大部分未登录词是专有名词，包括人名、地名、机构名等。黄昌宁、赵海 (2007)发现，未登录词（新词）识别错误对分词效果有着很大的影响。一般的专有名词还有一定的构词规律，如前缀后缀有迹可循。而新词则五花八门，如新术语、新缩略语、新商品名、绰号、笔名等。据 2002 年统计，每年会产生超过 800 个新的中文词（修驰, 2013）。直到目前为止，未登录词识别，尤其是新词识别，仍然是分词研究面临的最大挑战。尤其是在领域移植的情境下，当测试文本与训练数据的领域存在较大差异的时候，未登录词的数量增多，导致分词效果变差。

错别字、谐音字规范化：当处理不规范文本（如网络文本和语音转录文本）时，输入的句子中不可避免会存在一些错别字或者刻意的谐音词（如“香菇”->“想哭”;“蓝瘦”->“难受”;“蓝菇”->“难过”等等）。这些错别字或谐音字对于分词系统造成了很大的困扰。

分词粒度问题：分词粒度的选择长期以来一直是困扰分词研究的一个难题。选择什么样的词语切分粒度，是和具体应用紧密相关的。另外，Sproat et al. (1997)研究发现，即使是以汉语为母语的人，对于汉语词语认识的一致也只有 0.76。汉语语法教科书中对“词语”的定义是“语言中有意义的能单独说或用来造句的最小单位”，然而这种定义的实际操作性很差。实际操作时，如语料标注过程中，研究者们往往把“结合紧密、使用稳定”视为分词单位的界定准则，然而人们对于这种准则理解的主观性差别较大，受到个人的知识结构和所处环境的很大影响(黄昌宁、赵海, 2007)。这样就导致多人标注的语料存在大量不一致现象，

即表达相同意思的同一字串，在语料中存在不同的切分方式，如“我国”和“我/国”。修驰(2013, 表 1-3)粗略估计发现，在 SIGHAN Bakeoff-2005 采用的 PKU 训练语料中，有约 3%的字可能存在切分不一致的问题。考虑到目前分词模型的准确率已经可以达到 95%（F 值）以上，切分不一致的问题可能导致语料本身无法可信地评价模型。

3.2 词性标注

给定一个切好词的句子，词性标注的目的是为每一个词赋予一个类别，这个类别称为词性标记（part-of-speech tag），比如，名词（noun）、动词（verb）、形容词（adjective）等。

3.2.1 词性标注主要方法

词性标注是一个非常典型的序列标注问题。最初采用的方法是隐马尔科夫生成式模型，然后是判别式的最大熵模型、支持向量机模型，目前学术界通常采用结构感知器模型和条件随机场模型。近年来，随着深度学习技术的发展，研究者们也提出了很多有效的基于深层神经网络的词性标注方法。

3.2.2 词性标注研究的近几年主要进展

词性标注和句法分析联合建模：研究者们发现，由于词性标注和句法分析紧密相关，词性标注和句法分析联合建模可以同时显著提高两个任务准确率。

异构数据融合：汉语数据目前存在多个人工标注数据，然而不同数据遵守不同的标注规范，因此称为多源异构数据。近年来，学者们就如何利用多源异构数据提高模型准确率，提出了很多有效的方法，如基于指导特征的方法、基于双序列标注的方法、以及基于神经网络共享表示的方法。

基于深度学习的方法:传统词性标注方法的特征抽取过程主要是将固定上下文窗口的词进行人工组合，而深度学习方法能够自动利用非线性激活函数完成这一目标。进一步，如果结合循环神经网络如双向 LSTM，则抽取到的信息不再受到固定窗口的约束，而是考虑整个句子。除此之外，深度学习的另一个优势是初始词向量输入本身已经刻画了词语之间的相似度信息，这对词性标注非常重要。

3.3 句法分析

语言语法的研究有非常悠久的历史，可以追溯到公元前语言学家的研究。不同类型的句法分析体现在句法结构的表示形式不同，实现过程的复杂程度也有所不同。因此，科研人员采用不同的方法构建符合各个语法特点的句法分析系统。下文主要对句法分析技术方法和研究现状进行总结分析。

3.3.1 依存句法分析

依存语法历史悠久，最早可能追溯到公元前几世纪 Panini 提出的梵文语法。依存语法的现代理论起源于法国语言学家 Lucien Tesniere 的工作。依存语法存在一个共同的基本假设：句法结构本质上包含词和词之间的依存（修饰）关系。一个依存关系连接两个词，分别是核心词（head）和依存词（dependent）。依存关系可以细分为不同的类型，表示两个词之间的具体句法关系。目前，依存语法标注体系已经被自然语言处理领域的许多专家和学者所接受和采纳，应用于不同语言中，并不断地发展和完善。研究者们提出并实现了多种不同的依存分析方法，达到了较好的准确率。计算自然语言学习国际会议 CoNLL 举办的公开评测任务中，2006、2007 连续两年举行了多语依存句法分析评测，对包括汉语在内的十几种语言进行依存分析；2008、2009 年则对依存分析和语义角色标注联合任务进行评测。国内外多家大学、研究机构和商业公司都参加了这些评测任务。这些评测一方面提供了多种语言的标准评测数据集，另一方面提供了研究者们就依存句法分析进行集中交流、讨论的平台。

依存句法分析的形式化目标是针对给定输入句子$x = w_0 w_1 \ldots w_i \ldots w_n$，寻找分值（或概

率）最大的依存树d^*：

$$d^* = \text{argmax}_{d \in Y(x)} \text{Score}(x, d; \theta) \quad (1.1)$$

其中，$Y(x)$表示输入句子x对应的合法依存树集合，即搜索空间；θ 为模型参数，即特征权重向量。

目前研究主要集中在数据驱动的依存句法分析方法，即在训练实例集合上学习得到依存句法分析器，而不涉及依存语法理论的研究。数据驱动的方法的主要优势在于给定较大规模的训练数据，不需要过多的人工干预，就可以得到比较好的模型。因此，这类方法很容易应用到新领域和新语言环境。数据驱动的依存句法分析方法主要有两种主流方法：基于图（graph-based）的分析方法和基于转移（transition-based）的分析方法。

基于图的依存句法分析方法：基于图的方法将依存句法分析问题看成从完全有向图中寻找最大生成树的问题。一棵依存树的分值由构成依存树的几种子树的分值累加得到。根据依存树分值中包含的子树的复杂度，基于图的依存分析模型可以简单区分为一阶和高阶模型。高阶模型可以使用更加复杂的子树特征，因此分析准确率更高，但是解码算法的效率也会下降。基于图的方法通常采用基于动态规划的解码算法，也有一些学者采用柱搜索(beam search)来提高效率。学习特征权重时，通常采用在线训练算法，如平均感知器（averaged perceptron）。

基于转移的依存句法分析方法：基于转移的方法将依存树的构成过程建模为一个动作序列，将依存分析问题转化为寻找最优动作序列的问题。早期，研究者们使用局部分类器（如支持向量机等）决定下一个动作。近年来，研究者们采用全局线性模型来决定下一个动作，一个依存树的分值由其对应的动作序列中每一个动作的分值累加得到。特征表示方面，基于转移的方法可以充分利用已形成的子树信息，从而形成丰富的特征，以指导模型决策下一个动作。模型通过贪心搜索或者柱搜索等解码算法找到近似最优的依存树。和基于图的方法类似，基于转移的方法通常也采用在线训练算法学习特征权重。

多模型融合的依存句法分析方法：基于图和基于转移的方法从不同的角度解决问题，各有优势。基于图的模型进行全局搜索但只能利用有限的子树特征，而基于转移的模型搜索空间有限但可以充分利用已构成的子树信息构成丰富的特征。详细比较发现，这两种方法存在不同的错误分布。因此，研究者们使用不同的方法融合两种模型的优势，常见的方法有：stacked learning；对多个模型的结果加权后重新解码(re-parsing)；从训练语料中多次抽样训练多个模型(bagging)。

3.3.2 短语结构句法分析

短语结构句法分析的研究基于**上下文无关文法**（Context Free Grammar，CFG）。上下文无关文法可以定义为四元组<T, N, S, R>，其中 T 表示终结符的集合（即词的集合），N 表示非终结符的集合（即文法标注和词性标记的集合），S 表示充当句法树根节点的特殊非终结符，而 R 表示文法规则的集合，其中每条文法规则可以表示为 $N_i \rightarrow \gamma$，这里的 γ 表示由非终结符与终结符组成的一个序列（允许为空）。

根据文法规则的来源不同，句法分析器的构建方法总体来说可以分为两大类：人工书写规则和从数据中自动学习规则。人工书写规则受限于规则集合的规模：随着书写的规则数量的增多，规则与规则之间的冲突加剧，从而导致继续添加规则变得困难。

与人工书写规模相比，自动学习规则的方法由于开发周期短和系统健壮性强等特点，加上大规模人工标注数据，比如宾州大学的多语种树库的推动作用，已经成为句法分析中的主流方法。而数据驱动的方法又推动了统计方法在句法分析领域中的大量应用。为了在句法分析中引入统计信息，需要将上下文无关文法扩展成为**概率上下文无关文法**（Probabilistic Context Free Grammar，PCFG），即为每条文法规则指定概率值。概率上下文无关文法与非概率化的上下文无关文法相同，仍然表示为四元组<T, N, S, R>，区别在于概率上下文无关文法中的文法规则必须带有概率值。获得概率上下文无关文法的最简单的方法是直接从树库中读取规则，利用最大似然估计（Maximum Likelihood Estimation，MLE）计算得到每条规则的概率值。使用该方法得到的文法可以称为简单概率上下文无关文法。在解码阶段， CKY

等解码算法就可以利用学习得到的概率上下文无关文法搜索最优句法树。虽然基于简单概率上下文无关文法的句法分析器的实现比较简单，但是这类分析器的性能并不能让人满意。性能不佳的主要原因在于上下文无关文法采取的独立性假设过强：一条文法规则的选择只与该规则左侧的非终结符有关，而与任何其它上下文信息无关。文法中缺乏其它信息用于规则选择的消歧。因此后继研究工作的出发点大都基于如何弱化上下文无关文法中的隐含独立性假设。

针对这个问题，研究人员提出了两种截然不同的改进思路。一是词汇化（Lexicalization）方法，在上下文无关文法规则中引入词汇的信息；二是符号重标记（Symbol Refinement）方法，通过对非终结符的改写（细化或者泛化）而引入更多的上下文信息。词汇化方法首先由 Magerma 在 1995 年取得较大进展，但是该方法真正突破性的工作是 Collins 句法分析器，它第一次将英语句法分析的性能提高到接近 90%的性能。后续工作包括 Charniak 句法分析器以及两阶段的重排序（reranking）句法分析器。符号重标记方法的动机在于人工标注的树库中的非终结符可能过于粗粒或者过于细粒化，这两种情况都不利于统计学习，因此有必要对树库中的非终结符标记进行标注。最简单的重标注方法是将任意节点的父亲节点的文法标记挂载到儿子节点的非终结符标记上以便扩大上下文的范围，例如 Johnson 的工作。在 Klein 和 Manning 的工作中，通过人工方法对非终结符进行细分类，获得了性能上的提升，但是他们的方法需要语言学知识的支持。Matsuzaki 等人首次使用了自动方法对非终结符进行重标记，将树库中出现的每个非终结符标记固定地分为八类。上述所述的符号重标记方法相对于简单概率上下文无关文法都获得了性能上的提升，但是与基于词汇化的句法分析器相比并不具有性能上的优势。直到 2006 年，Petrov 和 Klein 用期望最大化（Expectation Maximization，EM）方法对树库中的类别标记进行自动的切分-合并（splitting-merging），这样得到的文法所对应的句法分析器获得了优于词汇化句法分析器的性能。其后，Petrov 等人对上述模型进行了进一步的改进，并且提出了基于隐含标记文法的由粗到精（coarse-to-fine）解码算法。在研究了单系统之后，Petrov 进一步研究了用于系统的融合的 product 模型。其出发点是期望最大化方法的参数初始化过程对最终得到的句法分析器的性能有较大的影响。Product 模型随机选择多组不同的初始化参数，每组参数对应一个模型，最终得到的句法分析器由这些模型组合得到。当前主流的句法分析模型，无论底层的机器学习方法（生成模型或者判别模型）或是所采用的系统框架（单系统、多系统融合或者两阶段的重排序方法），本质上都可以归到基于词汇化方法或者基于符号重标记方法的句法分析器。

3.3.3 深层文法句法分析

相对前两种句法分析，深层文法句法分析的研究相对较少。本节简要介绍词汇化树邻接文法（Lexicalized Tree Adjoining Grammar，LTAG）、词汇功能文法（Lexical Functional Grammar，LFG）和组合范畴文法（Combinatory Categorial Grammar，CCG）。

词汇化树邻接文法，简称 LTAG，是对树邻接文法（TAG）进行词汇化扩展得到的。树邻接文法包含两种基本树（Elementary Tree）：初始树（Initial Tree）和辅助树（Auxiliary Tree）。初始树是非递归的最小语言结构，比如 NP、PP 等等。它的特点是：1）所有内部节点的标签都是非终结符；2）所有叶子节点的标签是终结符或者是可用于替换的非终结符。辅助树是一种递归结构。它的特点是：1）所有内部节点的标签都是非终结符；2）除了一个节点的其他叶子节点的标签是终结符或者是可用于替换的非终结符，另外那个非终结符节点被称为 Foot Node，用于和其他子树相连；3）那个 Foot Node 和相连子树的 Root Node 的标签相同。在树邻接文法中，有两种子树操作：替换（Substitution）和插接（Adjunction）。词汇化语法是给所有基本树都和具体词关联起来，使得树更加具有个性化。

词汇功能文法，简称 LFG，是一种短语结构文法。在上个世纪七十年代，美国计算语言学家 Joan Bresnan 和 Ronald Kaplan 提出词汇功能文法，属于形式文法。LFG 把语言看成是由多维结构组成的，每一维都用特殊规则、概念和格式表示成一个特殊结构。LFG 包含两种最基本的结构：1）F-结构，用于表示语法功能；2）C-结构，用于表示句法功能。除此之外还有一些其他结构，用于表示浅层信息，例如谓词论元关系等。LFG 研究核心目标是构建一个理论语言学家感兴趣的深层文法模型，同时是计算语言学家能接受的格式严格且能高效

分析的模型。因此，基于 LFG 的分析器已经被应用于机器翻译，例如 TranSphere 和 Lekta。

组合范畴文法，简称 CCG，是一种类型驱动的词汇化文法，通过词汇范畴显式地提供从句法到语义的接口，属于短语结构文法。CCG 的基本操作包括：1)原子范畴(Atomic Category)，用于表达基本的词汇类别和句法功能；2）组合范畴（Function Category），由原子范畴构成，通常用 X/Y 或 X\Y 来表示可以向左或者向右寻找变元 Y 来获得组合 X。

基于深层文法的句法分析器也取得一些进展。例如，Boullier 和 Sagot 构建基于 LFG 的分析器-SxLFG。WenduanXu 等人借鉴依存分析模型，采用 Shift-reduce 框架构建高效的 CCG 分析器取得很好的效果。在树库语料方面，大多数深层文法树库是通过从 PTB 自动转换得到的，而黄昌宁老师在清华中文树库基础上结合中文特点，探索如何构建中文 CCG 树库。

3.3.4 基于深度学习的句法分析

近年来，深度学习（Deep Learning）在句法分析课题上逐渐成为研究热点，主要研究工作集中在特征表示方面。传统方法的特征表示主要采用人工定义原子特征和特征组合，而深度学习则把原子特征进行向量化，在利用多层神经元网络提取特征。所谓向量化就是把词、词性等用低维、连续实数空间上的向量来表示，从而便于寻找特征组合与表示，同时容易进行计算。

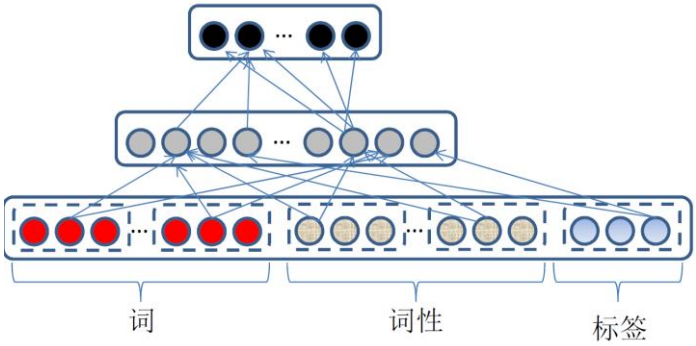


图 1 特征向量表示

在图 1 中，把词、词性、类别标签等原子特征表示为向量，然后利用多层网络进行特征提取。深度学习在特征表示方面有如下优点：1）只需要原子特征。这些原子特征以前是通过人工的自由组合形成最终的一元特征、二元特征、三元特征、四元特征甚至更多元的组合。这种人工组合最后取得较好的效果，但是事实上我们不知道怎么组合能形成最佳的特征集合。深度学习将所有的原子特征向量化之后，直接采用向量乘法以及非线性等各种运算从理论上能实现任意元的特征组合。2）能使用更多的原子特征。比如基于图的模型中，在建立弧时，不仅仅使用左边第一个词、右边第一个词等原子特征，还可以使用左边整个词序列、右边整个词序列的特征。研究人员把这种基于深度学习的特征表示方法分别应用在基于图的句法分析模型和基于转移的句法分析模型上，实验结果表明深度学习方法开始在句法中发挥作用。目前基于深度学习方法得到越来越多的研究者关注，在国际主流会议（ACL、NAACL、EMNLP）上发表了众多工作。

4. 技术展望与发展趋势

纵观词法和句法分析研究发展的态势和技术现状，以下研究方向或问题将可能成为未来研究必须攻克的堡垒：

- 深度学习和传统方法相结合的问题：最近几年在 ACL/EMNLP/NAACL 等会上，深度学习热潮席卷了 NLP 的各个任务包括句法分析。目前深度学习相关的研究工作主要是在特征表示上。而传统方法在这个方面的研究已经经历很长时间并取得较好的成果。如何把两种方法进行有效结合将是一个很有意思的研究课题。
- 多粒度分词：从目前研究来看，采用单一粒度分词存在两个问题。第一，采用单一

粒度分词规范时，标注人员对于分词规范的理解存在差异，因此会影响人工标注数据的质量。第二，不同的上层应用对于分词粒度的需求不同，有些应用甚至需要不同粒度的分词结果，从而从不同的角度对句子进行分析和理解；故而，作为汉语处理的第一步，采用单一粒度的分词规范也无法充分满足上层应用的需要。如果采用多粒度分词，数据标注时，便可以弱化分词规范，可以让标注者根据自己对词语定义的理解进行标注，模棱两可的地方允许提供多个标注结果。然而，采用多粒度分词也会引入一些新的问题和挑战：1）需要设计新的分词方法：目前的分词方法，如基于字序列标注的方法和基于转移的方法，均只支持单一粒度的分词形式；2）需要设计合理的多粒度分词评价方法；3）多粒度分词规范下，如何进行词性标注、句法分析，也是一个很有意思的问题。值得一提的是，海量云分词系统目前支持两种粒度的分词结果，其中粗粒度主要服务于一些上层应用如文本分类、信息检索、机器翻译等，使得上层应用可以直接在粗粒度词语上进行匹配或分析；而细粒度则可以缓解数据稀疏问题，解决粗粒度词语在训练数据中没有出现的问题。另外，研究者们近年来尝试进一步对词内部结构进行分析，也可以认为是多粒度分词方面的工作。

- 面向非规范文本的分词：数据和方法。目前的分词系统主要是在规范新闻文本为主的数据上训练，当不规范文本时，分析准确率急剧下降。因此，如何提高分词系统在文本领域、类型切换时的鲁棒性，是一个值得深入研究的问题。一方面，我们需要针对非规范文本，有选择的人工标注一定规模的数据，用于模型的评价和训练。近两年，学术界已经开始组织微博文本的分词公开评测，提供了一定规模的人工标注数据用来训练和评价模型，吸引研究者们展不规范文本上分词方法的探索。
- 分词、新词发现（词语归一化）交互建模：如上所述，未登录词识别，尤其是新词识别，对分词效果的影响很大。参考人在理解一个句子的过程可以发现，新词识别和分词是一个紧密联系，互相交互的过程。如果一个句子中出现了一个新词时，人在理解句子时会尝试多种分词结果，甚至会综合句法结构、语义结构是否合理，从而判断出这个新词是否应该是一个词，如果确认应该是一个新词，又会根据上下文尝试理解这个新词的含义。所以人在理解语言时，是词法、句法、语义不断交互的过程。然而，我们很难实现这么复杂的模型。但是目前来看，我们至少应该探索如何让分词和新词发现有效交互起来，从而显著提高分词效果 (Tao Qian et al., 2015)。
- 面向非规范文本的词性标注：和分词任务相似，目前的词性标注系统主要是在规范新闻文本为主的数据上训练，当不规范文本时，分析准确率急剧下降。因此，如何提高词性标注系统在非规范文本上的准确率，是一个值得深入研究的问题。近年来，学术界已经开始关注这个问题，也提出了不少有效的方法。比如 2012 年 Google 组织的 parsing the web 评测，将英文网络文本的词性标注作为一个预处理任务；2015 年 NLPCC 组织了微博文本的分词和词性标注公开评测。然而，目前这个问题还远远没有解决，一方面需要构建更多的训练和评价数据，另一方面也需要提出新的方法和模型。
- 词性标注的数据标注问题：词性标注人工标注数据目前也存在一些问题，可能会影响面向不规范文本的词性标注研究。第一，数据中往往会包含存在一些标注不一致现象，如 CTB 中“新华社”被大量标为两种词性“NR”（专有名词）和“NN”（普通名词）。第二，词性标注标签集合如何设计？这个问题取决于词性标注服务于什么任务。从底层语言分析的角度看，CTB 中词性标注主要服务于句法分析，因此对很多虚词根据其句法功能做了很多细分；而北大人民日报语料中，词性标注主要服务于命名实体识别等信息抽取任务，因此对名词的类别进行了细分。
- 互联网文本分析和领域自适应问题：句法分析面临的一个重要问题是面对和训练数据很接近的新闻领域规范文本时性能较好，但用于其它领域或类型的文本（比如口语化文本）时，准确率则急剧下降。随着大规模网络数据的出现，依存句法分析的重要挑战是如何精准分析有别于传统新闻文本的网络文本。

通过对词法分析和句法分析技术在最近 20 来年的发展现状和趋势分析，我们有理由相信，随着机器学习、半监督学习和词法分析等相关技术的快速进展，词法、句法分析这个自

然语言处理的基础任务性能将会进一步提升，会对机器翻译、信息抽取和对话系统等上层自然语言处理任务的性能提高产生更大的推动作用，进而以自然语言处理为核心技术的产业应用将更加广阔。

第二章 语义分析研究进展、现状及趋势

1. 任务简述、目标和研究意义

语义分析（Semantic Analysis）指运用各种机器学习方法，学习与理解一段文本所表示的语义内容。语义分析是一个非常广的概念，任何对语言的理解都可以归为语义分析的范畴。一段文本通常由词、句子和段落来构成，根据理解对象的语言单位不同，语义分析又可进一步分解为词汇级语义分析、句子级语义分析以及篇章级语义分析。一般来说，词汇级语义分析关注的是如何获取或区别单词的语义，句子级语义分析则试图分析整个句子所表达的语义，而篇章语义分析旨在研究自然语言文本的内在结构并理解文本单元（可以是句子从句或段落）间的语义关系。

简单地讲，语义分析的目标就是通过建立有效的模型和系统，实现在各个语言单位（ 包括词汇、句子和篇章等）的自动语义分析，从而实现理解整个文本表达的真实语义。

从理论上讲，语义分析涉及语言学、计算语言学、人工智能、机器学习，甚至认知语言等多个学科，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，即有利于推动相关学科的发展，揭示人脑实现语言理解的奥秘。在应用上，语义分析一直是自然语言处理的核心问题，它有助于促进其他自然语言处理任务的快速发展。比如，语义分析在机器翻译任务中有着重大的应用：在过去 20 多年的发展历史中，统计机器翻译主要经历了基于词、基于短语和基于句法树的翻译模型。目前已有相关研究将词汇级语义应用于统计机器翻译，并取得一定的性能提高，基于句子级、甚至篇章级语义的统计翻译一直是未来的研究方向。再比如，基于语义的搜索一直是搜索追求的目标。所谓语义搜索，是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

语义分析同时还是实现大数据的理解与价值发现的有效手段。伴随着互联网技术的迅猛发展和普及以及用户规模的爆发式增长，互联网已经步入了“大数据”时代，大数据已成为我们面临的常态问题，语义分析与大数据在某种程度上其实是互为基础的。一方面，如果想得到更准确的语义分析结果，需要大数据的支持，即从大数据中挖掘出并形成更大、更齐全、更准确的知识库，而知识库对语义分析的性能有着重要的影响。另一方面，如果想从大数据库中挖掘出更多、更有用的信息，人们需要用到语义分析等自然语言处理技术。总得来说，大数据为语义分析的发展提供了契机，但离开语义分析，基于大数据的信息获取、挖掘、分析和决策等其他应用，也将变得寸步难行。

目前，语义分析技术还不完美，特别是在句子级和篇章级，它仍面临很多具体问题和困难。本文对语义分析研究的主要内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势，分别从词汇级、句子级和篇章级三个层次做简要介绍。

2. 研究内容和关键科学问题

本节我们分别从三个层次描述语义分析内容和关键科学问题:词汇级、句子级和篇章级。

2.1 词汇级语义分析

词汇层面上的语义分析主要体现在如何理解某个词汇的含义，主要包含两个方面：一，在自然语言中，一个词具有两种或更多含义的现象非常普遍。如何自动获悉某个词存在着多种含义，以及假设已知某个词具有多种含义，如何根据上下文确认其含义，这些都是词汇级

语义研究的内容。在自然语言处理领域，这又称为词义消歧。二，如何表示并学习一个词的语义，以便计算机能够有效地计算两个词之间的相似度。

（1）词义消歧

词汇的歧义性是自然语言的固有特征。词义消歧根据一个多义词在文本中出现的上下文环境来确定其词义，作为各项自然语言处理的基础步骤和必经阶段被提出来。词义消歧包含两个必要的步骤：(1)在词典中描述词语的意义；(2)在语料中进行词义自动消歧。例如“苹果”在词典中描述有两个不同的意义：一种常见的水果；美国一家科技公司。对于下面两个句子：

她的脸红得像苹果。

最近几个月苹果营收出现下滑。

词义消歧的任务是自动将第一个苹果归为“水果”，而将第二个苹果归为“公司”。从上面的例子中我们发现，词义消歧主要面临如下两个关键问题：（1）词典的构建；（2）上下文的建模。

（2）词义表示和学习

对于词义表示，早期的做法将某个词义表示为从该词义在同义词网络中出现的位置到该网络根之点之间的路径信息。词义表示的另一个思路是将其数字化。最直观、也是到目前为止最常用的词表示方法是 one-hot 表示方法，这种方法把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。不难想象，这种表示方法存在一个重要的问题：任意两个词之间都是孤立的。造成的结果是：光从两个向量中看不出两个词是否有关系，即使这两个词是同义词，例如“计算机”和“电脑”、“上海”和“上海市”。

随着机器学习算法，目前更流行的词义表示方式是词嵌入（Word Embedding，又称词向量）。其基本想法是：通过训练将某种语言中的每一个词映射成一个固定维数的向量，将所有这些向量放在一起形成一个词向量空间，而每一向量则可视为该空间中的一个点，在这个空间上引入“距离”，则可以根据词之间的距离来判断它们之间的（词法、语义上的）相似性。

2.2 句子级语义分析

句子级的语义分析试图根据句子的句法结构和句中词的词义等信息，推导出能够反映这个句子意义的某种形式化表示。根据句子级语义分析的深浅，又可以进一步划分为浅层语义分析和深层语义分析。

（1）浅层语义分析

语义角色标注(Semantic Role Labeling，简称 SRL)是一种浅层的语义分析。给定一个句子，SRL 的任务是找出句子中谓词的相应语义角色成分，包括核心语义角色（如施事者、受事者等）和附属语义角色（如地点、时间、方式、原因等）。根据谓词类别的不同，又可以将现有的 SRL 分为动词性谓词 SRL 和名词性谓词 SRL。

目前 SRL 的实现通常都是基于句法分析结果，即对于某个给定的句子，首先得到其句法分析结果，然后基于该句法分析结果，再实现 SRL。这使得 SRL 的性能严重依赖于句法分析的结果。例如，相关研究表明基于自动句法分析的中文 SRL 的性能要低于基于正确句法分析 SRL 性能 F1 值大概 20%。因此，减轻 SRL 性能对句法分析性能的依赖是 SRL 研究的一个关键问题。

同时，在同样的句法分析结果上，名词性谓词 SRL 的性能要低于动词性谓词 SRL。例如，在中文 PropBank 上，使用相同的训练、开发集和测试集上，中文名词性谓词的 SRL 性能 F1 值较中文动词性谓词 SRL 性能 F1 值低约 20%。其主要原因为：

- 尽管采用相同的数据集，名词性谓词的标注实例数仍远小于动词性谓词。
- 名词性谓词的角色识别更加困难。根据名词性谓词的角色标注原则，即使某个名词为动词的派生词，并不是该名词的所有修饰成分都将被标注为该名词的语义角色。
- 名词性谓词与其角色之间的结构更加灵活与复杂。
- 名词性谓词的识别远远要比动词性谓词的识别困难。

因此，如何提高名词性谓词 SRL 性能也是 SRL 研究的另一个关键问题。

（2）深层语义分析

不难看出，浅层语义分析主要围绕着句子中的谓词，为每个谓词找到相应的语义角色。给定一个句子，深层的语义分析（有时直接称为语义分析,Semantic Parsing）不再以谓词为中心，而是将整个句子转化为某种形式化表示，例如：谓词逻辑表达式（包括 lambda 演算表达式）、基于依存的组合式语义表达式（dependency-based compositional semantic representation）等。以下给出了 GeoQuery 数据集中的一个中英文句子对，以及对应的一阶谓词逻辑语义表达式：

中文： 列出 在 科罗拉多 州 所有 的 河流

英文: Name all the rivers in Colorado

语义表达式: answer(river(loc_2(stateid('colorado'))))

虽然各种形式化表示方法采用的理论依据和表示方法不一样，但其组成通常包括关系谓词（如上例中的 loc_2、 river 等）、实体（如 colorado）等。语义分析通常需要知识库的支持，在该知识库中，预先定义了一序列的实体、属性以及实体之间的关系。虽然形式化表示方法的不同也体现在语义分析方法有所差异，但概况地讲，深度语义分析主要面临如下二个关键问题。

- **普通文本到实体/关系谓词之间的映射**。自然语言的一个主要特点在于其表达形式的丰富多样性，对同样的表达意思（如某个语义表达式），不仅可以使用不同的语言进行表达。如何建立普通文本到实体/关系之间的映射是一个关键问题。
- **面向开放领域的语义分析**。受标注语料的限制，目前的很多语义分析研究都限于某一特定领域。随着面向开放领域的知识库的构建及完善，如 Freebase 等，人工大规模标注涉及各领域的语义表达式是个费时费力的过程。为此，需要探索基于半监督或无监督的语义分析研究。

2.3 篇章级语义分析

篇章是指由一系列连续的子句、句子或语段构成的语言整体单位，在一个篇章中，子句、句子或语段间具有一定的层次结构和语义关系，篇章结构分析旨在分析出其中的层次结构和语义关系。具体来说，给定一段文本，其任务是自动识别出该文本中的所有篇章结构，其中每个篇章结构由连接词，两个相应的论元，以及篇章关系类别构成。篇章结构可进一步分为显式和隐式，显式篇章关系指连接词存在于文本中，而隐式篇章关系指连接词不存在于文本中，但可以根据上下文语境推导出合适的连接词。对于显式篇章关系类别，连接词为判断篇章关系类别提供了重要依据，关系识别准确率较高；但对于隐式篇章关系，由于连接词未知，关系类别判定较为困难，也是篇章分析中的一个重要研究内容和难点。

3. 技术方法和研究现状

与其他自然语言处理任务类似，目前主流的语义分析方法也是基于统计的方法，该方法以信息论和数理统计为理论基础，以大规模语料库为驱动，通过机器学习技术自动获取语义知识。本节仍分别从词汇级、句子级以及篇章级出发，结合第二节讨论的研究内容，简述语义分析的技术方法和研究现状。

3.1 词义消歧

词义消歧的研究通常需要语义词典的支持，因为词典描述了词语的义项区分。英语的词义消歧研究中使用的词典主要是 WordNet，而中文使用的词典有 HowNet，以及北京大学的“现代汉语语义词典”等。除词典外，词义标注语料库标注了词的不同义项在真实文本中的使用状况，为开展有监督的词义消歧研究提供了数据支持。常见的英文词义标注语料库包括 Semcor（普林斯顿大学标注）、DSO（新加坡国立大学标注）以及用于 Senseval 评测的语料库等。在中文方面，哈尔滨工业大学和北京大学分别基于 HowNet 和北大“现代汉语语义词

典”标注了词义消歧语料库。

词义消歧的研究是自然语言处理的一项基础关键，根据所使用的资源类型不同，可以将词义消歧方法分为以三类：

（1）基于词典的词义消歧

基于词典的词义消歧方法研究的早期代表工作是 Lesk 于 1986 的工作。给定某个待消解词及其上下文，该工作的思想是计算语义词典中各个词义的定义与上下文之间的覆盖度，选择覆盖度最大的作为待消解词在其上下文下的正确词义。但由于词典中词义的定义通常比较简洁，这使得与待消解词的上下文得到的覆盖度为 0，造成消歧性能不高。

（2）有监督词义消歧

有监督的消歧方法使用词义标注语料来建立消歧模型，研究的重点在于特征的表示。常见的上下文特征可以归纳为三个类型：（1）词汇特征通常指待消解词上下窗口内出现的词及其词性；（2）句法特征利用待消解词在上下文中的句法关系特征，如动－宾关系、是否带主/宾语、主/宾语组块类型、主/宾语中心词等；（3）语义特征在句法关系的基础上添加了语义类信息，如主/宾语中心词的语义类，甚至还可以是语义角色标注类信息。

最近随着深度学习在自然语言处理领域的应用，基于深度学习方法的词义消歧成为这一领域的一大热点。深度学习算法自动的提取分类需要的低层次或者高层次特征，避免了很多特征工程方面的工作量。

（3）无监督和半监督词义消歧

虽然有监督的消歧方法能够取得较好的消歧性能，但需要大量的人工标注语料，费时费力。为了克服对大规模语料的需要，半监督或无监督方法仅需要少量或不需要人工标注语料。例如 Yarowsky（1995）仅需要使用少量的人工标注语料作为种子数据，Ng 等（2003）从词对齐的双语语料抽取种子数据。Resnik（1997）根据词的不同歧义往往也体现在句法搭配上的差异这一思想，通过计算“语义优选强度”和“选择关联度”在大规模语料中自动获取句法结构的语义优选，然后用之于词义消歧。一般说来，虽然半监督或无监督方法不需要大量的人工标注数据，但依赖于一个大规模的未标注语料，以及在该语料上的句法分析结果。另一方面，待消解词的覆盖度可能会受影响。例如，Resnik（1997）仅考察某部分特殊结构的句法，只能对动词、动词的主词/宾语、形容词修饰的名词等少数特定句去位置上的词进行消歧，而不能覆盖所有歧义词。

3.2 词嵌入学习

词嵌入的学习通常与语言模型是捆绑在一起的，即训练语言模型的同时也学习和优化了词嵌入，具体请参见《语言表示与深度学习研究进展、现状与未来发展趋势》一章内容。

3.3 语义角色标注

语义角色标注的任务明确，即给定一个谓词及其所在的句子，找出句子中该谓词的相应语义角色成分。语义角色标注的研究热点包括基于成分句法树的语义角色标注和基于依存句法树的语义角色标注。同时，根据谓词的词性不同，又可进一步分为动词性谓词和名词性谓词语义角色标注。尽管各任务之间存在着差异性，但标注框架类似。以下以基于成分句法树的语义角色标注为例，任务的解决思路是以句法树的成分为单元，判断其是否担当给定谓词的语义角色。系统通常可以由三部分构成：

- 角色剪枝. 通过制定一些启发式规则，过滤掉那些不可能担当角色的成分。
- 角色识别. 在角色剪枝的基础上，构建一个二元分类器，即识别其是或不是给定谓词的语义角色。
- 角色分类. 对那些是语义角色的成分，进一步采用一个多元分类器，判断其角色类别。

在以上的框架下，语义角色标注的研究热点是如何构建角色识别和角色分类分类器。常用的方法有基于特征向量的方法和基于树核的方法。

在基于特征向量的方法中，最具有代表性的 7 个特征，包括成分类型（constituent type）、谓词子类框架（subcategorization）、成分与谓词之间的路径（parse tree path）、成分与

谓词的位置关系（constituent position）、谓词语态（predicate voice）、成分中心词（constituent head word）和谓词本身（predicate）。这 7 个特征随后被作为基本特征广泛应用于各类基于特征向量的语义角色标注系统中，同时后续研究也提出了其他有效的特征。

作为对基于特征向量方法的有益补充，核函数的方法挖掘隐藏于以句法结构中的特征。例如，可以利用核函数 PAK（Predicate/Argument Structure Kernel）来抓取谓词与角色成分之间的各种结构化信息。此外，传统树核函数只允许“硬”匹配，不利于计算相似成分或近义的语法标记，相关研究提出了一种基于语法驱动的卷积树核用于语义角色标注。

在角色识别和角色分类过程中，无论是采用基于特征向量的方法，还是基于树核的方法，其目的都是尽可能准确地计算两个对象之间的相似度。基于特征向量的方法将结构化信息转化为平面信息，方法简单有效；缺点是在制定特征模板的同时，丢弃了一些结构化信息。同样，基于树核的方法有效解决了特征维数过大的问题，缺点是在利用结构化信息的同时会包含噪音信息，另外计算开销远大于基于特征向量的方法。

3.4 句子级深层语义分析

虽然在句子级的语义分析任务中，语义的表示形式具有多样性，造成分析方法也会存在着差异。但根据所使用的资源类型不同，可以将语义分析方法分为以下三类：

（1）基于知识库的语义分析

该类型的语义分析利用的资源只有知识库（如 DBpedia、Freebase、Yoga 等），而没有人工标注的语义分析语料。知识库中通过三元组等形式记录了一系列的事实。对某个给定的句子，语义分析通过某种转换技术，将句子分析为一系列知识库中已定义的元组，并构成一个实体关系图。

（2）有监督语义分析

有监督的语义分析需要人工标注的语义分析语料支持。在人工标注的语义分析语料中，为每个自然语言句子人工标注了其语义表达式。常见的有监督语义分析方法包括采用同步上下文无关文法（Synchronous Context-Free Grammar，简称 SCFG）和 CCG 文法（Combinatory Categorical Grammar）。基于 SCFG 的方法把语义分析理解为一个机器翻译过程，即把语义表达式看作是翻译的目标语言，只不过该目标语言具有一定的结构要求。SCFG 方法采用同步上下文无关文法来同时构建源端句子和目标端翻译，当源端句子形成时，目标端的翻译也自动获得。基于 CCG 的方法通过添加逻辑表达式扩展了传统的 CCG 表示，例如：$flight \vdash N:\lambda_x.flight(x)$，其中 flight 指单词，$N$ 指 CCG 框架下的句法结构，而 $\lambda_x.flight(x)$ 指该单词对应的逻辑表达式。语义分析的过程也类似于基于 SCFG 的方法，即通过单词及其 CCG 句法结构形成句子，而通过逻辑表达式形成语义表达式。

（3）半监督或无监督语义分析

无监督的语义分析方法不需要利用人工标注的语义分析语料，仅利用知识库（或数据库）中的实体名/关系名等，也不利用知识库中的记录的事实。无监督的语义分析方法虽然不使用人工标注的语料，但通常会采用 EM 算法，在每轮算法迭代中，对句子进行语义分析，并且选择置信度高的句子及其语义分析结果作为自训练数据集。

3.5 篇章分析

相对于词汇级和句子级的语义分析，篇章级语义分析还处理初始阶段。目前的篇章语义分析主要还是围绕着判定子句与子句的篇章语义关系。以下我们先按照篇章分析标注体系来描述其主要实现方法，然后再描述中文篇章分析。

（1）基于 Penn Discourse TreeBank 的篇章分析

PDTB 作为目前最大的篇章语料库，在其上开展的篇章分析的研究也越来越多。一个基于 PDTB 的端对端的篇章分析通常划分为四个子任务，分别是：（1）篇章连接词识别；（2）论元(Argument)识别；（3）显式篇章关系识别；（4）隐式篇章关系识别。下面我们分别阐述这四个子任务。

篇章连接词识别。此步骤的主要任务是确定文本中的连接词。由于连接词通常是一个闭

集合，一般的做法是先收集所有的候选连接词，然后采用一个二元分类器判断其是否为连接词。基于 PDTB，在自动句法上的连接词识别 F1 值可以达到 94%左右。由于连接词识别的任务相对较简单，所取得的性能已达到实用阶段，通常它并不是篇章分析的研究重点。

- 论元识别。此步骤的主要任务是在文本中识别连接词的两人个论元 (Arg1 和 Arg2)，即识别它们的跨度。通常的作法与语义角色标注方法类似，即把连接词看作是语义角色标注中的谓词，然后以句法树中的成分为识别单元，判断其是否为该连接词的论元；最后对判断为同一论元的成分，采用后处理的方式优化得到该论元的跨度。不难看出，论元识别的性能依赖于句法树的质量。例如，采用精确匹配的评测情况下，利用正确句法树和自动句法树取得的论元识别性能大概是 54%和 40%。因此，对于论元的识别，特别是在要求精确匹配的情况下，仍然是一个很有挑战性的研究问题。
- 显式篇章关系识别。在获得一个显式篇章关系的情况下，识别其类别可以看作是一个多元分类问题。由于连接词本身与某个关系识别具有较强的相关性，因此，显式篇章关系的识别任务相对较简单，不是篇章分析研究的重点。
- 隐式篇章关系识别。在 PDTB 中，如何连续的上下两个句子不构成显式篇章关系的话，通常会形成一个隐式篇章关系。由于不存在显式连接词，隐式篇章关系类别的识别是一个非常困难的问题，因为这往往需要理解两个句子，才能正确判断其篇章关系类别。虽然可以简单地把其看作是一个分类问题，但制定有效的特征仍是一开放问题，传统方法定义的特征包括词汇、词性、句法特征等，同时为减轻数据稀疏性问题，往往采用词类或词聚类信息。目前，也有一些基于词嵌入深度学习的隐式篇章关系识别研究，但其性能提高仍然有限。基于 PDTB，隐式篇章关系识别的准确率只有 45%左右，远低于显式篇章关系识别性能，是目前篇章分析的一个研究重点和难点。

（2）基于 RST 的篇章分析

相对于 PDTB，基于 RST 篇章分析研究相对较少和较早。潜在的原因可能在于 RST 语料发布时间比 PDTB 发布时间要早，但在规模上却不及 PDTB，而且 PDTB 是建立在宾州树库基础上的。基于 RST 的篇章分析主要包含有两个子任务:(1)篇章基本单元 (Element Discourse Unit, 简称 EDU) 识别;（2）篇章结构生成，即对每一个过程的输出采用自底向上方法，为功能子句对确定一个最可能的修辞关系。

- EDU 识别。EDU 识别是指对给定句子，识别哪些为 EDU。通常的做法是通过识别 EDU 边界，从而获取 EDU。识别 EDU 边界可以采用基于规则的方法和基于统计模型的方法等。例如，基于序列标注的篇章分割模型，使用词汇和句法特征，采用 CRF，学习得到的 EDU 准确率达 94%。
- 篇章结构生成。在识别 EDU 的基础上，篇章结构生成任务包括识别 EDU 之间的关系，形成篇章树。例如，Hernault 等在 RST 上实现了基于 SVM 的篇章结构分析器 HILDA。该分析器对篇章切分和关系识别均使用 SVM 分类器，采用贪婪的自底向上的方法构建篇章结构树，篇章结构树构建的时间复杂度取决于输入文本的长度。HILDA 在树构建和篇章关系分析上取得较好的效果，结构识别 F1 值为 72%，完整句法树识别 F1 值为 47%。

（3）中文篇章分析

中文篇章分析起步较晚，目前的研究成果不多。主要是在语料资源建设方面，在借鉴和参考了英文篇章标注体系的基础上，标注了一批中文篇章分析资源，主要包括如下：

- 基于 RST 体系的标注。中国传媒大学和南京师范大学均分别基于英文 RST 框架标注了中文语料，但他们的研究都表明英文 RST 的很多篇章关系无法在中文中找到与之对应的关系。
- 基于 PDTB 体系的标注。LDC 发布了基于英文 PDTB 标注体系的中文篇章分析语料，该语料是基于中文宾州树库之上标注的。PDTB 标注体系是以连接词为谓词标注其论元结构，但由于中文中连接词大量缺省，该标注体系表现出很大不适应，此外，PDTB 标注体系通常不能构建一个完整的篇章结构树。哈尔滨工业大学发布了 HIT-CDTB 篇章分析语料，该语料包含 525 篇标注文本，语料文本来源于 OntoNotes4.0，覆盖了句群关系、复句关系、分句关系等多级信息。整体上，HIT-CDTB

还是参照了英文 PDTB 的标注体系。

- 基于连接依存树的标注。苏州大学发布了基于连接依存树的中文篇章结构表示体系的中文篇章分析语料（Chinese Discourse TreeBank，CDTB），该标注体系借鉴了 RST 和 PDTB 体系优点，并结合中文的特点。CDTB 采用自顶向下的标注策略，对每一段内容先找出其最上层关系，然后递归地对切分后的内容进行标注。目前 CDTB 共包含 500 个文档，全部来自宾州大学汉语树库（Penn Chinese Treebank，简称 CTB），每个段落标注为一棵连接依存树，共有效标注 2342 个篇章（段落）。CDTB 标注内容包括：连接词、连接词位置、连接类型、关系类型、中心位置、子句切分位置、子节点、父节点等。CDTB 共标注关系 7310 个，其中显式关系 1814 个，隐式关系 5496 个，篇章结构层次最大为 9 层。

目前基于中文篇章语料的篇章分析研究不多，CoNLL 2016 评测任务包含了端到端的中文篇章分析，但参赛小组数量明显少于英文篇章分析。

4. 技术展望与发展趋势

基于对语义分析研究内容和技术现状分析和总结，我们可以看出，语义分析后续研究的发展趋势主要包括：

- 短语/句嵌入的学习。目前词嵌入已经在自然语言处理领域有了广泛的应用。如何为更大粒度的语言单位（如短语，甚至句子等）学习得到其相应的嵌入表示，已经成为目前研究的一个热点。短语/句嵌入有着广泛的应用，例如，由于缺少连接词，隐式篇章关系识别变得非常困难。为更好的理解两个论元之间的篇章关系，仅靠一些表面特征（如词汇、句法等特征）是远远不够的。
- 基于句子级语义分析的篇章融合。目前的语义分析都是以句子为基本单位。以语义角色标注为例，由于缺省（特别是中文）现象，很多时候谓词的角色并没有出现在谓词所在的句子中，而是位于前面句子中。目前的语义角色标注通常都直接忽略了此类情况。此外，指代消解从篇章的角色串起了一序列实体，如何融合指代消解的识别结果和语义角色标注的识别结果来展现篇章的语义，或将成为篇章级语义分析的一个值得关注的研究方向。
- 中文篇章分析。目前，中文的篇章分析研究成果不多。由于中文与英文的差异性，中文篇章分析不能完全套用英文篇章分析的方法。一方面，随着中文篇章分析语料的发布和不断累加，另一方面，随着中文基础研究技术的成熟，中文篇章分析将会成为大家关注的一个研究方向。
- 非规范文本的语义分析。微博、Twitter、Facebook 等社交媒体网站产生大量的口语化、弱规范甚至不规范的短文本，这些具有实时性、数量众多的社交媒体短文本是非常具有研究和应用价值的，被广泛用于情感分析和事件发现等任务。目前的语义分析技术几乎都是面向规范化的文本，直接应用于非规范文本上将不可避免地导致低性能问题。因此，如何针对非规范文本的语义分析也必将成为未来的一个研究热点。

Go on!!!

第三章 语篇分析研究进展、现状及趋势

1.任务定义、目标和研究意义

语篇分析又称话语分析或篇章分析，是对“语篇”整体进行的分析，包括语篇基本单元之间的关系，不同语篇单元的成份间关联以及语篇所含的信息等等。

语篇是由一个以上的句子（sentence）或语段（utterance）构成的。一篇文章、一段会话等都可以看成语篇。构成语篇的句子（或语段）彼此之间在形式上相互衔接，在意义上前后连贯。

自然语言处理的发展对语篇分析的研究起到了重要的推动作用。大多数应都是在语篇层面上的，很少针对单个词、短语或者句子，如机器翻译、文本摘要、自动会话、机器阅读理解等，这些应用都需要利用语篇信息。下面是一个汉英翻译的例子：

选自“人民日报”的一段话：

①吉林省梨树县女农民蔡淑珍，②过去不懂技术，③养鸡鸡死，④养兔兔亡，⑤赔了几万元，⑥险些寻了短见。

若用 google 翻译器翻译成英文，翻译结果为：

①Lishu female farmer Cai Shuzhen,②in the past I do not understand technology, ③chicken chicken died, ④raising the death bunny, ⑤lost several million, ⑥almost to find a short-sighted.

翻译结果出现了多类错误，下面仅从语篇关系的角度对错误作简单分析：

首先，汉语的①与②联合构成一个完整的句子，①充当其中的主语。但是，在翻译结果中，②独立翻成了一个完整句子，这使得①的翻译成为多余；其次，②中的“I”也不正确。翻译器将②看成缺失主语（零形式），并将零形式补充为“I”（当然，这一补充是错误的）。如果翻译器有原文语篇分析的功能，能分析出①与②共同构成一个句子，这一错误就可以避免。再看后面 4 句，翻译结果中③④⑤⑥之间没有形成关联关系，这不符合英语的表达习惯；也没有补充缺失的零形式（主语）。可以推断，翻译器没有很好甚至缺乏对这 4 句进行语篇分析。整体上看，汉语的表述既简洁又明了，人们阅读时很容易理解。但翻译之后，句子之间的逻辑关系在翻译中完全没有体现出来，连贯性弱，理解非常困难。

另一方面，语篇分析也可以为词、短语和句子的分析提供更多有用的信息。例如，句子“你能穿多少就穿多少”，如果独立理解，至少有二种意思：一种是表示尽量少穿，另一种则是尽量多穿。这二种相反的意思对应着二种不同的分词结果：“你/ 能/ 穿/ 多/ 少/ 就/ 穿/ 多/ 少”和“你/ 能/ 穿/ 多少/ 就/ 穿/ 多少”。若不利用上下文信息，很难判断哪一种是合适的。

2.研究内容和关键科学问题

2.1 研究内容

语篇分析是指超越单个句子范围的各种可能分析，包括句子（语段）之间的关系以及关系类型的划分，段落之间的关系的判断，跨越单个句子的词与词之间的关系分析，话题的继承与变迁等。

由于研究目标的不同，人们研究的内容也不同。语篇的衔接关系分析主要是分析词汇（或短语）之间的语义关联，如同义关系，反义关系，整体与部分关系、远程搭配关系以及指代

关系等，话题的演化与推进很大程度上要借助于衔接性分析的结果。这就需要研究词汇层面的各种关系计算方法和支撑相关计算所需要的语言资源建设。语篇的连贯性则主要以句子为出发点研究前后文的关联，以期揭示内容的推演过程以及前后文之间的逻辑关系，这就需要研究合理的关系表示和关系分析的有效计算模型。

2.2 关键科学问题

(1) 理论体系的研究。这需要解决如下关键科学问题：其一，语篇的基本单元是什么，如何界定；其二，语篇基本单元与内层的词汇或短语之间的关系如何刻划，与同层的其它语篇基本单元之间的关系又如何表示；进一步，语篇基本单元是否需要复合成更大的单元以及如何复合；其三，语篇的若干基本问题（如指代问题，话题的推演问题）如何体现在理论体系中。

(2) 计算模型的构建。这需要解决如下关键科学问：其一，语篇的基本单元之间的关系如何分析，其分析手段有哪些；其二，在语篇基本单元之间关系不明显（尤其是汉语）的情况下，如何有效分析它们是否存在关系以及关系类型是什么。其三，指称语（Reference Expression）识别。指称语通常是不能通过自身的词义确定其意义（如代词）的表述，在形式上由名词（名词短语）或代词表示。如果是名词（短语），如何与一般的名词（短语）区分。其四，同指消解（Coreference Resolution）。如何判断两个指称语之间是否具有同指关系。

3. 技术方法和研究现状

3.1 篇章性、连贯性与衔接性理论

语篇理论的研究主要集中于上一世纪 80 年代。对于语篇的判断准则，较具代表性的工作是 Beaugrande 于 1981 年提出的语篇篇章性（textuality）标准，共包括 7 条：衔接性，连贯性，意图性，可接受性，信息性，情境性以及篇际性。Beaugrande 的体系非常大，但具体操作上有些困难，在计算中不太容易把握。

在构建理论模型方面，典型的工作有 Kamp 的语篇表示理论 DRT(Discourse Representation Theory)，Grosz 的中心理论 CT（Center Theory）以及 Mann 和 Thompson 的修饰结构理论 RST（Rhetorical Structure Theory)。DRT 受到蒙太古（Montague）语法理论的影响，采用了类似模型论的思想，通过谓词演算主要解释回指问题和语义推理关系；CT 则通过句子或语段（utterance）的中心以及中心变化的规律来实现指代消解，也为话题（在 CT 中表示中心）的推演分析提供了基础，BFP 算法（Brennan, Friedman 和 Pollard）是 CT 实现的经典算法。RST 则以句子（或语段）为语篇的基本单元，构建语篇的结构关系，这类似于将一个句子以其中的词为基本单元构建句法结构。总体上看，RST 主要考虑的是语篇的整体连贯性，DRT 和 CT 更多地考虑了衔接和局部连贯。

此外，Hallidy 和 Hason 专门针对衔接性问题，提出了五种衔接关系：连接(Conjunction)、指代(Reference)、省略(Ellipsis)、替换(Substitution)和词汇衔接(Lexical Cohesion)关系。除了连接（Conjunction）之外，其余的都可以看成为广义词汇衔接关系。所谓广义词汇衔接，是指通过相同或相关联的词表示上下文的概念关系。衔接性是从词汇或短语层面来描述语篇中的话题或概念之间关系的，在计算上相对容易实现。例 2 给出了衔接性的说明：

例 2. ⑴**社交**的吃**饭**种类虽然**复杂**，⑵性质极其**简单**。⑶把**饭**给自己有**饭**的人吃，⑷那是请**饭**；⑸自己有**饭**可吃而去吃人家的**饭**，⑹那是**赏面子**。⑺**交际**的微妙不外乎此。⑻反过来说，⑼把**饭**给没**饭**吃的人吃，⑽那是**施食**，⑾**赏面子**就一变而成**丢脸**。⑿这便是慈善**救济**，⒀算不上**交际**了。（钱钟书：《吃饭》）

其中，“饭”出现了 8 次，“赏面子”出现了 2 次，“交际”出现了 2 次；此外，还有与“赏面子”相关的“丢脸”，与“交际”相关的“社交”以及与“复杂”相关的“简单”、与“施食”相关的“救济”等，重复出现或相关词出现都是词汇的衔接性表现。通过相同词和相关词的多次出现，将语篇所强调的概念或话题呈现出来。

在国内，宋柔教授针对语篇关系提出了广义话题结构理论。基本思想是以标点句为基础，从话题-说明关系的视角出发，描述话题的流水结构。

语篇研究绕不开的一个问题是代词问题。在 Hallidy 和 Hason 提出的 5 种衔接性中，指代（Reference）和省略（Ellipsis）主要由代词或 0-形式表示的。DRT 和 CT 的一个直接任务就是指代消解，Mann 和 Thompson 研究连贯性的动机之一也是同指或指代问题。广义话题理论也通过嵌入结构对 0-形式给予了解释。

3.2 语篇结构分析技术

语篇结构自动分析的发展很大程度上得益于两个有代表性的语篇关系库：宾州语篇树库（Penn Discourse Treebank，PDTB）和 RST 树库（Rhetorical Structure Theory-Discourse Treebank，RST-DT）。PDTB 将句子看作论元（Argument），主要标注论元对之间的关系和可能的连接词，把一个大的语篇分解成平面化的论元对，语篇标注层次较浅，可以看作是浅层语篇结构标注。RST-DT 基于修辞结构理论（Rhetorical Structure Theory，RST）构建，由美国南加州大学和美国国防部共同创建，通过修辞关系对语篇结构进行描述，将整个语篇构建成一棵有层次的 RST 树。近年来，有学者提出利用依存结构进行语篇结构的分析，在已有的 RST-DT 的基础上转换建立依存树库。二元非对称的依存结构解释了篇章的深层关系，保留了 RST 树中的大部分信息。因为相对简单的结构，可以直接分析各个单元之间的关系，使机器自动分析工作能更容易开展。

（1）RST 树结构分析：在 RST 树结构的分析过程中，关联性强的单元先通过修辞关系进行组合，形成大的语篇单元，大的语篇单元再形成更大的语篇单元，直至形成一棵覆盖语篇所有单元的树。一般来说，在同一段落内的语篇单元关系要强于不同段落之间的语篇单元。因此，通常是先进行段内的语篇单元关系分析，之后再分析段落之间的关系。RST 树的构成过程，类似于短语结构树的分析过程，语篇单元（通常是小句）类似句法中的词，这就可以借鉴语法分析中的很多算法，例如移进归约算法、CYK 算法等。

Maucu 在 1999 年首次使用了有导机器学习方法构造了一个移进归约（shift-reduce）语篇分析器，他使用了 C4.5 决策树方法，从标注数据中自动学习规则。

也有研究者把语篇结构分析看作一步或多步的分类问题。例如，用一个二分类器判断邻近的哪些单元应该合并，之后多分类器判断合并后的单元属于何种修辞关系。这种简单的策略在语篇结构分析上获得了不错的结果。也有研究人员针对说明书类的文章进行了语篇分析，提出了一种基于移进归约的分析方法，并用分类器进行关系标注。分类器可以从包括组合语义在内的大量特征中学习一阶逻辑规则。

近年来，深度学习技术也开始在语篇结构分析中尝试。例如，基于词向量和句法分析技术，通过递归神经网络（Recursive Neural Network）或循环神经网络（Recurrent Neural Network）得到语篇单元的特征表示，之后再利用分类器进行邻近语篇单元的修辞关系判别和标注。

（2）依存结构树分析：依存分析结果也是一颗树，但和 RST 树不同，句子之间直接建立依存关系，不再含有中间节点。语篇依存结构的单元之间以非对称的二元依存关系连接。其中，称依靠单元为“从属单元”（subordinate），称被依靠的单元为“中心单元”（head）。利用语篇依存树表示依存结构时，需要在依存树起始位置插入一个人工单元，称之为 e0，并视之为该语篇的根（Root），以此简化定义与计算过程。

为了自动分析语篇的依存结构，可以借鉴句法依存分析技术，如 Eisner 算法或最大生成树方法，利用两个句子间的特征以及最大间隔的学习方法得到一棵最优的语篇依存树。还可以借鉴依存句法分析的其他技术，如基于转换的句法分析等，选择适合篇章的特征进行语篇分析。

（3）隐含语篇关系分析：在进行语篇关系的确定时，由于隐含语篇关系缺乏可直接作

为特征的显式连接词，相对于含连接词的显式关系分析更具挑战性。

对于隐式关系的识别，研究者们一般采用有监督学习的思路，将其作为一个分类问题进行处理。包括设计各类特征训练分类器，尝试使用机器学习的各类方法来提升分类器的效果。近年来，也开始有研究者尝试使用深度神经网络的方法，基本思想是，首先对句子对进行建模，学习其语义表示，之后再对两个表示进行组合，最终将组合后的特征向量输入，预测得到隐式语篇关系的分布。

由于标注语料的不足，半监督的方法也开始用于隐式关系的识别。这一类方法通过同时使用标注和无标注的语料，从大规模的无标注语料中学习语篇特征，并能够较大幅度的提高对一些出现频率较少的关系的识别效果。有研究者首先在无标注的语料上训练可以预测句子间连接词的模型，再将这个模型得到的预测结果，作为隐式关系识别的一个特征。这种方法利用了显式连接词与隐式关系之间的联系，巧妙地从无标注的数据中获得了有助于识别隐式关系的信息。

3.3 指代消解技术

语篇的衔接性分析是指从词汇层面分析语篇内的概念关系，主要包括指代关系、省略关系、替换关系和词汇衔接性。词汇衔接又表现为词汇重复、同义或近义、反义、上下位义、整体与部分以及搭配六种。具有衔接关系的词可以通过一个链表示，称为词汇链。

通过分析各个词汇链跨越的句子范围，可以判断相关概念或话题的持续情况。利用词汇链，可以获取文本的关键词集合；形成文本的摘要；在时序性报道中，还可以检测新话题，跟踪已出现的话题。

指代或同指关系是构建词汇链的一个重要内容。如果两个词或短语具有同指关系，那么它们应该属于同一词汇链。

所谓同指，是指两个名词之间，或者名词与代词之间具有相同的指称语义（referent）。同指关系是等价关系。指代是指由一个代词来表示上下文中某个名词或名词短语所指示的实体或实体的某个部分。语篇中第一次指示实体的名词或名词短语称为先行语（antecedent）。先行语通常会先于代词出现，此时的指代关系也称为回指（anaphora）；在有些情况下，也可能先出现代词，后出现先行语，此时的指代关系则称为预指（cataphoric reference）。

同指消解有很多方法，机器学习方法仍然是主流。基本思想是将每个指称语表示为特征（组合），再根据特征组合判断指称语之间是否具有同指（或等价）关系。这可以看成分类问题，按二分类情况判断为同指或不同指；也可以看成为排序问题，计算当前指称语与其它指称语之间的排序关系，将排序最后的作为同指关系；还可以看成聚类问题，即，对所有指称语进行聚类，形成若干聚类子集，位于同一子集的看成同指关系。

4. 总结与展望

语篇分析包括衔接性和连贯性两个方面。衔接性通过词汇（或短语）之间的关系来表示上下文的关联；而连贯性则通过句子或者句群之间的关系表示关联。

在分析语篇衔接性时，需要先区分词（短语）表示的是词义还是指称义。如果是词义，还需要进行词义消歧和词义的相似性或相关性计算；如果是指称义，则要进行同指或指代的消解。

近些年来，语篇分析（无论是结构分析还是同指消解）集中在分析技术的研究上。在具体实现中，主要采用了机器学习的方法，特别是有指导的学习方法，包括近几年较热的深度神经网络方法。

然而，这些年针对语言自身问题的研究较少，新的理论模型不多。加强语篇理论模型的研究对语篇分析尤为重要，这应成为今后研究的重点。

具体对汉语而言，还有一个更基本的问题是语篇基本单元的界定。目前，有两种最具代表性的观点：（a）以小句为基本单元；（b）以标点句（逗号，分号，句号等表示的词序列）

为基本单元。对于（a）而言，在语篇分析之前必须给出小句的判断方法。一种最直接的方法就是句法分析，但句法分析本身又是很难的问题。对（b）而言，仍然面临（a）的问题，汉语中标点句可能只是一个短语，如例 1 中的①，而语篇关系应该在句法之上，否则，就意味着还需要分析句法结构。汉语的语篇分析需要破解这一怪圈。

此外，汉语还有一个非常突出的问题是 0-指代问题。相比西方语系讲究句子结构的完整性，汉语则更加追求经济性或简洁性，即能省则省，其结果会导致大量的 0-形式，见例 3。

例 3. ①我自来是**如是**，②从会吃饮食时便吃药，③到今未断。④请了多少名医，⑤修方配药，⑥皆不见效。

其中的③，既缺失了主语，也缺失了宾语；对于⑤而言，缺失的主语由④的宾语表示，目前的指代消解方法不容易正确判断。如果用中心理论，⑤缺失的主语优先和④的主语一致，这就会出现错误。此外，例 1 中翻译的很多错误也是 0-形式导致的。

由于汉语存在上述明显的特点，汉语的语篇分析需要为此开展专门的研究。

第四章 语言认知模型研究进展、现状及趋势

1. 任务定义、目标和研究意义

认知语言学（cognitive linguistics）是认知科学（cognitive science）与语言学交叉的一个研究分支，是研究人脑的思维、心智、智能、推理和认识等认知机理及其对语言进行分析和理解过程的一门学问。该分支诞生于上个世纪 70 年代，至 80～90 年代取得了较大的发展，但到目前为止尚未形成一个完整的、系统的学科，不同学科的学者对其理解也是智者见智、仁者见仁，因此尚没有一个关于认知语言学的严密而完整的定义。认知语言学家王寅教授曾根据自己的理解将认知语言学定义为：坚持体验哲学观，以身体体验和认知为出发点，以概念结构和意义研究为中心，着力寻求语言事实背后的认知方式，并通过认知方式和知识结构等对语言做出统一解释的、新兴的、跨领域的学科。

语言认知计算模型就是刻画人脑语言认知和理解过程的形式化模型。四十多年来，认知语言学与神经科学、语言心理学和计算语言学等学科交叉，从各种不同的角度、以不同的目的不断探索着人类思维的奥秘，试图将人脑学习、分析和理解语言的思维过程通过形式化的模型描述出来。理想情况下，希望建立可计算的、复杂度可控的数学模型，以便在计算机系统上实现对人脑语言理解过程的模拟。尤其近几年随着人工智能研究的再度兴起，人们在对人脑理解语言的生物过程尚不清楚的情况下，也在尝试通过模拟人脑神经系统的结构和功能，或者借鉴人脑的某些认知行为（记忆、编码、搜索、概念形成、缺省推理、隐喻投射、概念整合等）的表现，或受人脑某些功能和表现的启发，建立实用、有效的自然语言处理模型或方法，实现所谓的“类脑语言信息处理”。

从事认知语言模型和类脑语言信息处理方法研究具有极其重要的理论意义和应用价值，它不仅可以从本质上揭示人脑进行语言学习、思维和推理的机理，探索大脑实现语义、概念和知识计算的奥秘，而且可以了解人类某些与语言能力相关的疾病形成的原因，对于改善人类的健康，提高计算机信息处理的能力，促进社会的发展，都具有非常重要的意义。

2. 研究内容和关键科学问题

认知语言模型和类脑信息处理都是比较宽泛的概念，拥有不同学科背景的学者所关注的研究内容和科学问题不尽相同。从大的方面讲，我们可以粗略地将相关研究内容归纳为两大类别：人脑处理语言的认知机理和类脑语言信息处理方法。

2.1 人脑处理语言的认知机理

有关人脑处理语言机理的研究开始于十九世纪，主要研究脑损伤病人的语言交流能力，通过对失语症患者的研究发现了大脑产生语言的区域（布洛卡区）和理解语言的区域（威尔尼克区）。在随后一百多年的研究探索中，认知神经科学家和语言心理学家等对人脑的生理结构、神经元信号和语义表征、记忆及整合等问题进行了大量研究，主要内容包括：

- **对人脑的结构和语言进化的过程进行研究：**如人脑的功能区域化和各区域间的解剖连接关系，大脑两半球间的整合作用产生的认知系统模块化组织，左脑半球的拓扑优势，连接语言和动作的大脑机制等；语言环境对人的语言认知和理解能力的影响，人脑的语言重建能力，脑的发育与第二语言习得等。
- **通过采集分析在某种语言环境下人脑的生理数据，研究人脑对语音、词汇、句法和语义的理解机理：**如词汇与大脑处理区域的对应关系，语义处理的脑信号，词汇

在大脑中的分布，名词和动词兼类词对命名实体和动作的神经关联的影响，预测与名词含义相关的脑区活动，语句理解过程中语义系统内的神经解剖学特性，利用功能磁共振成像（fMRI）和事件相关电位（ERP）等技术手段分析句子因果关系推理和事件相关的关系从句，简单句之间的概念关系，以及汉字的形、音处理，以及文化影响对处理算术问题的方式等。

这方面研究要解决的关键科学问题是：人脑进行语言理解的认知过程和机理是什么？什么生理因素或外部原因影响着人脑的语言认知能力和进化过程？

2.2 类脑语言信息处理

近年来，人们越来越多地关注如何通过研究人脑在某些任务上（如歧义消解、选择性限制、记忆容量等）的语言认知能力和表现，来建立语言信息处理和计算模型。

人脑处理语言的能力十分强大，可以快速、准确地阅读和理解各种类型的文本，即使文本中含有少量的错误，也不影响正常的理解，而且对语言的理解是增量式的，可以不断地将当前阅读的词汇语义与上文的含义进行整合，形成新的含义，并可在一定程度上推断下文的内容。另外，人脑在语言理解时只利用有限的记忆空间就可处理与当前词汇距离较远的词汇。如何对这些特性进行数学建模和模拟，一直是神经语言学家和计算语言学家关注的问题。

这一方向面临的科学问题是：是否可以对人脑执行语言理解的认知过程进行有效的数学建模？换句话说，语义和概念是否是可计算的？在尚不完全清楚人脑的语言认知机理和生物过程的情况下，如何在冯·诺伊曼计算机系统上模拟人脑的语言理解过程？大家知道，人脑在进行自然语言理解时，往往是利用多种信息（包括上下文、文本的主题、个人的背景知识等）综合完成的，而在不同情况下对不同信息的利用程度也不尽相同，有时仅需要很简单的语言学知识，有时却需要很深的语言学知识、甚至历史和文化知识，或者专业背景知识，才能对某段文字正确地理解。这些不同的知识之间是如何建立联系的？在什么情况下会被激活和利用？如何确定其参与理解和推理的程度？人脑中进行理解和推断的过程肯定不会是精确的数学计算和严格的逻辑推理，那么，如何通过数学建模的方式较为准确地模拟这个过程？这些都是需要进一步研究的问题。

3. 研究进展和现状

3.1 脑科学、认知神经科学与语言认知计算

正如前面所述，最早对脑科学和语言认知的研究是由医生和神经科学家进行的。他们使用了大量的实验和解剖学方法对非正常语言能力（如失语症患者）的大脑进行研究分析，发现了不同脑区与语言功能之间的关系，利用双通道理论初步解释了人脑语言处理的过程，提出了大脑腹侧通道负责语义处理、背侧通道将大脑运动区（包括额下回的听觉网络）与视听觉和感觉运动区形成联系的论断，发现了失语症的原因和不同类型的失语症与脑区的对应关系等。

脑生理结构和现代认知心理学的研究表明，语言信号中的不同词汇与大脑相应处理区域相对应，信号空间具有相似特征的信号被反映到脑皮质相近区域时，大致保留了信号空间的概率分布特征和拓扑结构特征，即大脑具有自动归类的功能。基于对正常的和脑损伤群体的行为和脑的研究证据，认知神经科学家发现语义记忆在大脑中是沿着特定的维度进行组织和表征的，对于物体和相应名词的语义，最重要的两个组织维度分别是人类所固有的感觉运动通道（如视觉、听觉、肢体运动等）和对生存与进化具有重要意义的领域（如动物、植物、工具、同类等）。以此为基础，结合神经影像学和计算机建模手段，认知神经科学家们已经能够在一定程度上从人脑活动模式中推测出其所正在观察的物体和正在思考的名词。

3.2 大脑语义整合的理论

此外，认知神经科学家和心理学家对于大脑语义整合的理论研究也有了一些进展。他们认为，语言处理过程至少涉及两种并行的过程：一个是语义记忆，这条通路负责检索单词间的语义特征、关联和语义关系；另一个是语义组合，这里至少有一个通路（可能是多条通路）负责将单词整合形成更高级的含义。2016 年 4 月《Nature》杂志封面刊发了题为“The Brain Dictionary”的轰动性成果，首次公布了大脑不同脑区对不同词汇语义的反应，用不同颜色将不同语义范畴的 985 个英语词汇在整个大脑上标识了出来，使人们看到了一幅五彩斑斓的大脑词汇分布图。这项研究的理论贡献在于发现了人脑进行语义加工时不仅在传统认为的布洛卡区，语义加工在大脑左右半球基本上是对称的，打破了传统研究中认为语言加工左半球偏侧化的情况。该研究得出了不同个体的大脑的整个语义网络“看起来特别相似”的结论。尽管有些结论和问题仍有待于进一步探索，但这毕竟向着揭示人脑记忆和理解语言奥秘的方向迈进了一步。

3.3 语言认知计算模型

在发明脑成像技术之前，语言心理学研究中普遍使用反应时和眼动追踪技术来研究人脑处理语言的过程，利用这些简单的技术取得了大量的成果。比较复杂的脑成像技术在最近几十年流行起来，如脑电图（electroencephalogram, EEG）、事件相关电位（event-related potential, ERP）、脑磁图（magnetoencephalography, MEG）、功能性磁核共振成像（functional magnetic resonance imaging, fMRI）和*正电子*发射型计算机断层显像（Positron Emission Computed Tomography, PET）等。这些技术各有优势，可以用不同的方式来测量大脑活动。虽然采集到的大脑活动信号中存在噪声，但是这些从人脑中直接采集的生理信号是最接近人脑活动的数据。因此，如何将生理信号用于语言认知计算模型的研究成为很多学者研究的焦点。例如，有学者在比较了 n-gram 语言模型，RNN 语言模型和基于概率短语结构语法（PSG）的语言模型之后发现，人在阅读句子时引发的事件相关电位与句子的信息量相关，不利用短语结构的语言模型可以更好地拟合采集的脑电数据；有人提出了一种多任务学习的模型，利用眼动数据和有标注的文本数据一起训练句子压缩模型，可有效提升模型的性能。也有学者通过对比人阅读时的脑成像数据和用统计语言模型对词汇的表示，发现基于循环神经网络的语言模型与脑成像数据拟合度更高等。

3.4 深度神经网络与自然语言处理

其实，1943 年心理学家 Warren McCulloch 和数理逻辑学家 Walter Pitts 提出的神经网络数学模型就是对神经元结构和信号传递方式给出的形式化数学描述，从而开创了人工神经网络的时代。在之后几十年的发展中，研究人员提出了许多重要的神经计算模型，包括：19 世纪 40 年代的 M-P 神经元和 Hebb 学习规则；50 年代感知器模型和自适应滤波器；60 年代的自组织映射网络和神经认知机等。由于训练过程容易出现过拟合现象，且参数训练速度慢，因而神经网络方法曾一度陷入低迷阶段。直到 2006 年，以 Hinton 为首的研究人员提出了深度信念网络，利用非监督的逐层贪心训练算法，解决了深度网络训练困难的问题，从此掀起了神经网络研究的一个新的高潮。

最近几年，深度神经网络模型在自然语言处理任务上也取得了不错的成绩，如 Yoon Kim 利用卷积神经网络（CNN）对文本进行分类，在 4/7 的文本分类任务上超过传统最好的成绩。Mikolov 等人利用循环神经网络模型构造语言模型，将华尔街日报任务上语音识别的错误率降低了 18%。Cho 等人提出了一种编码器—解码器模型，将源语言句子编码为固定长度的向量，然后再用这个向量解码出目标语言句子。在此基础上，Bahdanau 等人在编码器—解码器模型中加入对齐函数，使神经机器翻译在英法翻译任务的效果接近了传统短语翻译模型的方法，神经网络机器翻译的研究从此开始备受关注。另外，Weston 等人提出的记忆神经网

络和 Sukhbaatar 等人提出的端到端的记忆神经网络，在传统神经网络模型的基础上加入外部记忆模块，在自动问答等推理任务上取得了很好效果，等等。目前，在神经网络模型中融合记忆模块和注意力机制成为了研究的趋势。

3.5 研究现状

综上所述，人类对自身大脑的研究已经开启了一个霞光万道的新航程，取得了一大批令人振奋的成果。一方面，伴随神经影像技术（fMRI、MEG、ERP、PET 等）的发展，为计算建模研究提供了大量的经验数据，为基于心理语言学现象研究语言理解的机制，建立模拟人类智能的计算模型提供了良好的契机。神经科学对于人脑语言认知机理的研究为计算模型的建立提供了良好的生理学和心理学参照。以神经网络为代表的基于脑科学研究成果的若干认知方法和模型已经在自然语言处理中取得了可喜的效果。另一方面，计算模型也为语言认知机理的正确性和有效性检验提供了验证平台和工具，反过来推动语言认知机理研究的发展。当然，人脑毕竟具有极其复杂的结构，尤其在微观层面尚有太多的奥秘没有被揭开，语言理解过程涉及到多个脑区和数以亿计的神经元。可以说，目前人们对大脑处理语言的机理研究只是揭开了冰山一角，离真正认识大脑的语言处理机理并通过形式化数学方法准确地描述出来，还有非常遥远的道路要走。

4. 总结与展望

正如前面所述，语言认知模型的研究需要神经科学、认知科学、心理学和计算语言学等多种学科的协同努力。不同学科尽管面临不同的科学问题，但从揭示人脑理解语言的机理，并通过形式化模型准确地描述这一机理，最终构建有效的自然语言处理系统这个共同目标来看，努力的方向是一致的。粗略地讲，为了达到这个目标，未来将在如下几个方面进一步研究和探索：

- **从微观层面进一步研究人脑的结构，发现和揭示人脑理解语言的机理。**目前人们只是在宏观上大致了解脑区的划分和在语言理解过程中所起的不同作用，但在介观和微观层面，语言理解的生物过程与神经元信号传递的关系，以及信号与语义、概念和物理世界之间的对应与联系等，都是未知的奥秘。如何打通宏观、介观和微观层面的联系并给出清晰的解释，恐怕是未来必须解决的问题。介于微观和宏观之间的体系
- **建立完整的语言认知计算的理论体系和复杂度可控的形式化数学模型。**在语言认知模型研究方面，目前的工作基本上是离散的和实验性的，也就是说，研究人员针对某个具体的语言现象或某类语言学问题，通过大量的实验获取相关数据，然后加以分析、模拟和解释，有点类似于枚举的方法，而不同实验和不同问题之间往往没有必然的联系和相关性，所谓的模型基本都是解释性的，很难形成形式化的计算模型。因此，目前人们提到语言认知模型时往往泛指一个抽象的概念，而没有形成一整套完整的理论体系。因此，如何根据神经科学、认知科学、心理学、语言学和计算机科学等交叉领域研究的成果，建立完整的语言认知计算的理论体系和复杂度可控的形式化数学模型，也将是一项长期的艰巨而复杂的任务。
- **建立有效的、鲁棒、可解释的语言计算模型。**在类脑信息处理方面，目前的研究思路基本上是通过对人脑的某个功能或在某方面的行为表现进行建模和模拟，然后用于解决某个自然语言处理问题，所采用的理论工具是数理统计、信息论和机器学习。神经网络模型是一个比较成功的例子。神经网络模型在图像处理和语音识别等相关领域取得了很好的相效果，但这些任务大多解决的是“处理”层面的问题，如边界的切分、语音信号到文字的转换等，而上升到“语义理解”的层面还有太多的问题，如正确理解一幅图像所包含的语义和情感等，仍是极具有挑战性的问题。何况在自然语言处理中，神经网络模型远不如在图像处理和语音识别等任务中表现得那么好，究其原因还是因为神经网络模型未能真正从语义和概念的层面清楚地解释

其运算的道理和意义，而且在很多方面，并非受人脑工作方式的启发。如对于一个三岁的幼童，父母教给他某个词汇时只需要简单地告诉他（她）几个例子即可，根本不需要大规模标注样本，一个智力正常的孩童不仅可以学会新词，而且基本不会用错。可是，现有的神经网络模型一旦离开大规模训练样本，就变得毫无用处，而且当测试集与训练集有较大的差异时，模型的性能将急剧下降。这种现象明确地告诉我们，与人脑相比神经网络模型还“笨”得很！那么，随着对大脑认知机理研究的不断深入，如何建立更加有效的、鲁棒、可解释的计算模型，也是一个必须解决的问题。

随着计算机硬件和医学设备性能的提升，技术手段日渐强大，机器学习等大数据处理算法日臻成熟，更加深入地研究脑、了解脑和揭示脑的条件已经具备。近年来人工智能领域的一些突破性进展，如 IBM Watson 问答系统在“危险边缘”挑战赛中击败人类对手、谷歌公司利用深度学习和增强学习算法实现的 AlphaGo 系统在围棋项目上打败人类对手;微软小冰机器人以情感语料为基础，利用大数据知识搜索和深度神经网络机器学习方法等，建立了满足人的情感和心理需求的人机对话系统，这些成果让我们看到了未来智能信息处理的曙光。我们完全有理由相信，语言认知计算模型研究的春天已经到来，其研究成果必将在自然语言处理等相关领域中发挥重要的作用。

第五章 语言表示与深度学习研究进展、现状及趋势

1.任务定义、目标和研究意义

语言表示是对人类语言的一种描述或约定，是认知科学、人工智能等多个领域共同存在的问题。在认知科学里，语言表示是语言在人脑中的表现形式，关系到人类如何理解和产生语言。在人工智能里，语言表示主要指用于语言的形式化或数学的描述，以便在计算机中表示语言，并能让计算机程序自动处理。

人类语言不仅具有一定的语法结构，也蕴涵了其所表达的语义信息。但与计算机可以理解的人工语言（比如程序语言）不同，人类语言在语法和语义都充满了歧义性，需要结合一定的上下文和知识才能理解。这使得如何理解、表示以及生成自然语言变得极具挑战性。

从人工智能的角度，语言表示的研究内容可以定义为：如何设计一种计算机内部的数据结构来表示语言，以及语言和此数据结构之间的相互转换机制。

语言表示是自然语言处理以及语义计算的基础。语言具有一定的层次结构，具体表现为词、短语、句子、段落以及篇章等不同的语言粒度。为了让计算机可以理解语言，需要将不同粒度的语言都转换成计算机可以处理的数据结构。

早期的语言表示方法是符号化的**离散表示**。为了方便计算机进行计算，一般将符号或符号序列转换为高维的稀疏向量。比如词可以表示为 One-Hot 向量（一维为 1、其余维为 0 的向量），句子或篇章可以通过词袋模型、TF-IDF 模型、N 元模型等方法进行转换。但是离散表示的缺点是词与词之间没有距离的概念，比如“电脑”和“计算机”被看成是两个不同的词，这和语言的特性并不相符。这样，离散的语言表示需要引入人工知识库，比如同义词词典、上下位词典等，才能有效地进行后续的语义计算。一种改进的方法是基于聚类的词表示，比如 Brown 聚类算法，通过聚类得到词的类别簇来改进词的表示。

但是离散表示无法解决的“多词一义”问题，为了解决这一问题，可以将语言单位表示为连续语义空间中的一个点，这样的表示方法称之为连续表示。基于连续表示，词与词之间就可以通过欧式距离或余弦距离等方式来计算相似度。常用的连续表示有两种。

一种是应用比较广泛的分布式表示（Distributional Representations）。分布式表示是基于 Harris 的分布式假设，即如果两个词的上下文相似，那么这两个词也是相似的。上下文的类型可以为相邻词，所在句子或所在的文档等。这样我们就可以通过词与其上下文的共现矩阵来进行词的表示，即把共现矩阵的每一行看作对应词的向量表示。

另外一种是近年来在深度学习中使用的表示，即**分散式表示**（Distributed Representations）[1]。分散式表示是将语言的潜在语法或语义特征分散式地存储在一组神经元中，可以用稠密、低维、连续的向量来表示，也叫嵌入（Embeddings）。不同的深度学习技术通过不同的神经网络模型来对字、词、短语、句子以及篇章进行建模。除了可以更有效地进行语义计算之外，分散式表示也可以使特征表示和模型变得更加紧凑。

上述两种表示有一定的区别。分散式表示是指一种语义分散存储的表示形式，而分布式表示是通过分布式假设获得的表示。但这两者并不对立，比如 Skip-Gram、CBOW 和 glove 等模型得到词向量，即是分散式表示，又是分布式表示。

目前，语言表示在认知科学和人工智能领域都仍然没有一个完美的答案，仍面临很多具体问题和困难。接下来将对语言表示的主要内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势，做简要介绍。

[1]分散式表示也常叫做分布式表示，但为了和另一种分布式（Distributional）表示的区别，这里我们用分散式（distributed）表示。

2. 关键科学问题和研究内容

语言表示一直是人工智能、计算语言学领域的研究热点。从早期的离散表示到最近的分散式表示，语言表示的主要研究内容包括如何针对不同的语言单位，设计表示语言的数据结构以及和语言的转换机制，即如何将语言转换成计算机内部的数据结构（理解）以及由计算机内部表示转换成语言（生成）。概况地讲，语言表示主要面临如下关键科学问题：

2.1 语言表示的认知机理

语言是人类拥有的最复杂的认知功能之一，通过有限的符号来实现无限的表达能力。人脑通过对有限的词汇进行整合来表示复杂的人类思想。同时，语言长期来看是不断演化发展的，在短期又具有一定的稳定性。因此，一个高效的语言表示模型需要借鉴人类的认知机理。其中，比较关键的是人们对语言的理解需要大量的背景知识，因此语言表示和知识表示应该是相辅相成的。传统的离散语言表示方法在很多任务中也需要人工知识库（比如同义词词典、WordNet、HowNet 等）来提高效果。但是这些融合基本上都是词级别的。在句子或篇章级别的融合还比较困难。因此，如何构建语言表示和知识表示的联系，从人工知识库或大规模未标记语料来自动学习语言的表示，是语言表示研究的一个关键科学问题。

2.2 跨语种的统一语言表示

虽然不同语种的符号系统以及语法规则并不相同，但对世界以及人类思想的表达能力是相近的。因此，不同语种的语言表示也具有一定的相似性，即可以用同一种表示方式来刻画不同语言。比如在机器翻译中，可以考虑将同一含义的不同语言的句子表示为相同或相近的连续表示。因此，如何为不同语种构建一种统一的语言表示模型，利用不同语言之间的共性，从而提高各个语言的表示能力，也是语言表示需要研究的一个关键科学问题。

2.3 不同粒度单位的语言表示

在人类语言中，字、词、句子、篇章等不同粒度或层次的语言单位都需要结合不同的上下文进行理解。理解一段文本，需要理解其中每个词的语义。但每个词的理解也需要对整个文本的理解。比如，“一词多义”问题，必然要求从上下文语境中找出最相应的义项。而目前语言表示虽然在一定程度上解决了“一义多词”的问题，但“一词多义”问题也是语言表示研究中的一个重要内容。因此，结合语言本身的层次结构以及不同粒度文本之间的制约关系，构建一个多粒度文本的联合语义表示模型，也是语言表示研究中的一个关键科学问题。

2.4 基于少量观察样本的新词、低频词表示学习

目前，词的表示是通过大量的语料库学习得到的。受限于算法效率以及计算能力，目前的大多数语言表示方法经常忽略新词（或未登录词）或低频词的表示学习。即使对于低频词进行了特别处理，这些词的表示也难以像高频词那样很好地建模。然而语言中低频词往往富含有价值的信息，丢弃这些词也往往降低了语言表示的能力。而人们学习新词和低频词的方式并不是通过大量语料进行学习的，而是通过字典或少量观察样本进行学习。因此，对于新词或低频词，需要研究如何通过少量观察样本来学习新词和低频词的表示。这也是语言表示研究的关键科学问题之一。

3. 技术方法和研究现状

语言表示牵涉很多相关的研究领域，比如认知科学（语言在人脑中如何表示和加工？）、计算语言学（语言的本质是什么？）、人工智能（如何在计算机中表示语言？）等。本文主要从人工智能的角度综述语言表示的技术方法和研究现状。

语言表示方法大体上可以从两个维度进行区分。一个维度是按不同粒度进行划分，语言具有一定的层次结构，语言表示可以分为字、词、句子、篇章等不同粒度的表示。另一个维度是按表示形式进行划分，可以分为离散表示和连续表示两类。离散表示是将语言看成离散的符号，而将语言表示为连续空间中的一个点，包括分布式表示和分散式表示。

		表示模型	
		词	**句子、篇章**
离散表示	符号表示	One-Hot 表示	词袋模型，N 元模型
	基于聚类的表示	Brown 聚类	K-means 聚类
连续表示	分布式表示	潜在语义分析 潜在狄利克雷分配	
	分散式表示	Skip-Gram 模型 CBOW 模型	连续词袋模型，序列模型 递归组合模型，卷积模型

表 1. 语言表示模型划分

3.1 离散表示

离散表示是将语言看成离散的符号。以词为例，一个词可以表示为 One-Hot 向量（一维为 1 其余维为 0 的向量），也叫局部表示。离散表示的缺点是词与词之间没有距离的概念，这和事实不符。

一种改进的方法是基于聚类的词表示。其中一个经典的方法是 Brown 聚类算法，该算法是一种层次化的聚类算法。在得到层次化结构的词类簇之后，我们可以用根节点到词之间的路径来表示该词。

有了词的表示之后，我们可以进一步得到句子或篇章的表示。句子或篇章的离散表示通常采用词袋模型、N 元模型等。

3.2 连续表示

连续表示是将词表示为连续空间中的一个点，即为连续向量。这种表示的优势是词与词之间可以通过欧式距离或余弦距离等方式来计算相似度，可以有效地处理传统离散表示中的“一词多义”和“一义多词”问题。将词表示为连续向量也有很多方式，主要包括如下方式。

3.2.1 分布式表示

一种连续表示是基于 Harris 的分布式假设，即如果两个词的上下文相似，那么这两个词也是相似的。这样就可以通过共现矩阵的方式来进行词的表示，这类方法也叫分布式表示（Distributional Representations）。我们可以构建一个大小为 W×C 的共现矩阵 F，其中 W 是词典大小，C 是上下文。上下文的类型可以为相邻词、所在句子或所在的文档等。共现矩阵的每一行可以看作对应词的向量表示。基于共现矩阵，有很多方法来得到连续的词表示，

比如潜在语义分析模型（Latent Semantic Analysis，LSA）、潜在狄利克雷分配模型（Latent Dirichlet Allocation，LDA）、随机索引（random indexing）等。

如果我们取上下文为词所在的句子或篇章，那么共现矩阵的每一列是该句子或篇章的向量表示。结合不同的模型，我们很自然就得到了句子或篇章的向量表示。

3.2.2 分散式表示

另一种连续表示是分散式表示（Distributed Representations），即将语言表示为稠密、低维、连续的向量。根据所表示文本的颗粒度的不同，可以分为词、句子、篇章的表示。随着深度学习在自然语言处理中的研究不断深入，这种表示逐渐变得非常流行。

- 词表示

词的分布式表示也叫词嵌入（Word embeddings）。自然语言由词构成，深度学习模型首先需要将词表示为词嵌入。词嵌入向量的每一维都表示词的某种潜在的语法或语义特征。早期研究者并没有太多关注词嵌入的语言学解释，仅仅将其作为模型参数。因为词嵌入是一个稠密向量，这样不同词嵌入就存在了距离（或相似度）。一个好的词嵌入模型应该是：对于相似的词，它们对应的词嵌入也相近。因此很多研究者开始关注于如何得到高质量的词嵌入。研究者最早发现学习得到词嵌入之间存在类比关系。比如 apple-apples ≈ car-cars， man-woman ≈ king - queen 等。这些方法都可以直接在大规模无标注语料上进行训练。词嵌入的质量也非常依赖于上下文窗口大小的选择。通常大的上下文窗口学到的词嵌入更反映主题信息，而小的上下文窗口学到的词嵌入更反映词的功能和上下文语义信息。

在此基础上，也有研究者关注如何利用已有的知识库来改进词嵌入模型，结合知识图谱和未标注语料在同一语义空间中来联合学习知识和词的向量表示，这样可以更有效地实现词的嵌入。在一词多义方面，研究者将一个词的每个义项都用一个向量表示，在不同的上下文中选择一个最相关的义项进行更新。

- 句子表示

在自然语言处理中，很多任务的输入是变长的文本序列，而传统分类器的输入需要固定大小。因此，我们需要将变长的文本序列表示成固定长度的向量。

以句子为例，一个句子的表示（也称为编码）可以看成是句子中所有词的语义组合。因此，句子编码方法近两年也受到广泛关注。句子编码主要研究如何有效地从词嵌入通过不同方式的组合得到句子表示。其中，比较有代表性的方法有四种。

第一种是神经词袋模型，简单对文本序列中每个词嵌入进行平均，作为整个序列的表示。这种方法的缺点是丢失了词序信息。对于长文本，神经词袋模型比较有效。但是对于短文本，神经词袋模型很难捕获语义组合信息。

第二种方法是递归神经网络（Recursive Neural Network），按照一个给定的外部拓扑结构（比如成分句法树），不断递归得到整个序列的表示。递归神经网络的一个缺点是需要给定一个拓扑结构来确定词和词之间的依赖关系，因此限制其使用范围。

第三种是循环神经网络（Recurrent Neural Network），将文本序列看作时间序列，不断更新，最后得到整个序列的表示。

第四种是卷积神经网络（Convolutional Neural Network），通过多个卷积层和子采样层，最终得到一个固定长度的向量。

在上述四种基本方法的基础上，很多研究者综合这些方法的优点，结合具体的任务，已经提出了一些更复杂的组合模型，例如双向循环神经网络（Bi-directional Recurrent Neural Network）、长短时记忆模型（Long-Short Term Memory）等。

- 篇章表示

如果处理的对象是比句子更长的文本序列（比如篇章），为了降低模型复杂度，一般采用层次化的方法，先得到句子编码，然后以句子编码为输入，进一步得到篇章的表示。具体的层次化可以采用以下几种方法：

第一种是采用层次化的卷积神经网络，即用卷积神经网络对每个句子进行建模，然后以句子为单位再进行一次卷积和池化操作，得到篇章表示。

第二种是采用层次化的循环神经网络，即用循环神经网络对每个句子进行建模，然后再

用一个循环神经网络建模以句子为单位的序列，得到篇章表示。

第三种是混合模型，先用循环神经网络对每个句子进行建模，然后以句子为单位再进行一次卷积和池化操作，得到篇章表示。在上述模型中，循环神经网络因为非常适合处理文本序列，因此被广泛应用在很多自然语言处理任务上。

4. 技术展望与发展趋势

纵观语言表示研究的趋势和现状，语言表示的方法和手段一直随着人工智能研究的发展而不断进步。人工智能从早期的专家系统，到基于统计的方法，再到最近的深度学习方法，语言表示也从早期的离散表示，到分布式表示，再到最近的分散式表示。

目前，基于深度学习的方法在自然语言处理中取得了很大的进展，因此，分散式表示也成为语言表示中最热门的方法，不但可以在特定的任务中端到端地学习字、词、句子、篇章的分散式表示，也可以通过大规模未标注文本自动学习。分散式表示可以非常方便地应用在下游的各种自然语言处理任务上，并且可以端到端地学习，给研究者带来了很大的便利。但是分散式表示对以下几种情况还不能很好地处理，需要进一步解决。

- 语言中出现所有符号是否都需要使用统一的表示模型？比如，无意义的符号、变量、数字等。
- 新词以及低频词的表示学习方法。目前的表示学习方法很难对这些词进行很好的建模，而这些词都是极具信息量的，不能简单忽略。
- 篇章的语言表示。目前对篇章级别的文本进行建模方法比较简单，不足以表示篇章中的复杂语义。
- 语言表示的基础数据结构。除了目前的基于向量的数据结构之外是否有更好的表示结构，比如矩阵、队列、栈等。

随着深度学习、无监督学习、以及增强学习等技术的快速发展以及大量文本数据的涌现，语言表示作为自然语言处理中最基础的问题将会得到相当程度的解决，从而为下游的各种自然语言处理任务，诸如机器翻译、自动文摘、文本分类、自动问答等，提供有效的表示基础。

第六章 知识图谱研究进展、现状及趋势

1.任务定义、目标和研究意义

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。其中，概念是指人们在认识世界过程中形成对客观事物的概念化表示，如人、动物、组织机构等。实体是客观世界中的具体事物，如篮球运动员姚明、互联网公司腾讯等。事件是客观事件的活动，如地震、买卖行为等。关系描述概念、实体、事件之间客观存在的关联关系，如毕业院校描述了一个人与他学习所在学校之间的关系，运动员和篮球运动员之间的关系是概念和子概念之间的关系等。谷歌于 2012 年 5 月推出谷歌知识图谱，并利用其在搜索引擎中增强搜索结果，标志着大规模知识图谱在互联网语义搜索中的成功应用。

知识图谱以结构化的形式描述客观世界中概念、实体间的复杂关系，将互联网的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、管理和理解互联网海量信息的能力。知识图谱给互联网语义搜索带来了活力，同时也在智能问答中显示出强大威力，已经成为了互联网智能服务的基础设施。知识图谱与大数据和深度学习一起，已经成为推动人工智能发展的核心驱动力之一。

知识图谱技术是指在建立知识图谱中使用的技术，是融合认知计算、知识表示与推理、信息检索与抽取、自然语言处理与语义 Web、数据挖掘与机器学习等的交叉研究。知识图谱研究一方面探索从互联网语言资源中获取知识的理论和方法，另一方面促进知识驱动的语言理解研究。特别是，随着大数据时代的到来，研究从大数据中挖掘隐含的知识理论与方法，将大数据转化为知识，增强对互联网资源的内容理解，将促进当代信息处理技术从信息服务向知识服务转变。知识图谱在下面应用中已经凸显出越来越重要的应用价值：

知识融合：当前互联网大数据具有分布异构的特点，通过知识图谱可以对这些信息资源进行语义标注和链接，建立以知识为中心的资源语义集成服务；

语义搜索：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信息内容，而不是互联网网页；

问答系统：基于知识的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查询，直接得到用户关心问题的答案；

大数据分析与决策：知识图谱通过语义链接可以帮助理解大数据，获得对大数据的洞察，提供决策支持。

2.研究内容和关键科学问题

知识图谱技术通常包括知识表示、知识图谱构建和知识图谱应用三个方面的研究内容。知识表示研究客观世界的知识如何在计算机里表示和处理，知识图谱构建解决如何建立计算机的算法从客观世界或者互联网的各种数据资源中获取客观世界知识，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。可以看出，知识图谱表示、构建和应用是一项综合性的复杂技术。知识图谱技术既涉及自然语言处理中的各项技术，在资源内容的表示上可以使用从浅层的文本向量表示、到句法和语义结构表示，从自然语言处理技术上会使用到分词和词性标注、命名实体识别、句法语义结构分析、指代分析等。知识图谱反过来可以促进自然语言处理技术的研究，建立知识驱动的自然语言处理技术如基于知识图谱的词义排岐和语义依存关系分析等。

2.1 知识表示

知识表示对客观世界知识进行建模，表示客观世界知识中所蕴涵的语义内容以及关联，以便于机器识别和理解。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，知识表示理论是智能系统的基础性关键科学问题。

知识表示研究利用认知科学和心理学的研究成果，首先要了解人类本身是如何表示知识并利用他们解决问题的，然后将其形式化表示成计算机可以推理和计算的表达形式，建立基于知识的系统，提供智能知识服务。同时，知识表示也需要结合计算机对符号表示、处理和计算的能力。知识表示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的知识；2）建立什么样的知识表示可以具备语义表示能力；3）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。

现有的主要知识表示技术可以分成符号主义和联结主义。符号主义知识表示的基础是纽威尔和西蒙提出的物理符号系统假设，认为人类认知和思维的基本单元是符号，而认知过程就是在符号表示上的运算。联结主义认为人的认知就是相互联系的具有一定活性值的神经单元所形成网络的整体活动，知识信息不存在于特定的地点，而是在神经网络的联结或者权重中。

知识表示应该反应人类对客观世界的认知，并能够从不同层次和粒度表达客观世界所呈现的语义。本体这个概念在哲学中表示世界的本质，在计算机领域则表示计算机对客观世界或者感兴趣领域的概念化描述，通常表示对客观世界中概念、实体、事件及其关系的描述。

当前主要知识表示方法可以分成传统人工智能中基于符号逻辑的知识表示，如：产生式系统、谓词逻辑、框架表示、语义网等；互联网资源的开放知识表示方法，如 XML、RDF 和 OWL 等；基于知识图谱的表示学习通过深度学习可以将知识表示成低维连续实值稠密的实值向量空间，有助于实现高效的知识计算。

2.2 知识图谱构建

知识图谱构建是根据特定知识表示模型，从分布异构的海量互联网资源中采用机器学习和信息抽取等技术，建立大规模知识图谱的过程。知识图谱构建是知识图谱技术最为关键的技术之一，信息抽取和语义集成是知识图谱构建的核心技术问题。

知识图谱构建方法主要由三方面因素确定，其一是从什么样的数据资源中学习知识，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）和非结构化资源（如文本数据等）对象。维基类百科资源是利用群体智能建立的大规模供人阅读理解的知识资源，其中蕴含了大量的高质量的结构化知识，也是知识图谱构建时使用的重要资源。其二学习什么类型的知识，主要包括概念层次结构、事实知识、事件知识等。其三是使用什么样的学习方法获得知识，主要方法有有监督学习、半有监督学习和无监督学习方法。

此外，互联网上已经存在大量的结构化知识资源（如 Freebase，YAGO 等），这些知识资源之间互为关联，相互补充，很多知识计算任务需要联合多个知识资源给出结果。因此，异构知识资源的语义链接和集成也是知识图谱的一项核心技术，需要首先研究异构数据资源的关联方法，将其转化成为具有丰富链接关系的知识网络，进一步研究跨知识库的语义计算方法。因此，多源异构知识库的链接是一个亟需解决的问题。目前语义集成主要从语义网和自然语言处理两个方面分别进行。语义网领域的相关研究是数据链接，自然语言处理领域对应于实体链接。

2.3 知识图谱应用技术

知识图谱应用的任务是利用知识图谱，建立基于知识的系统并提供智能的知识服务，是知识图谱建立的终极目标。主要包括：基于知识的互联网资源的信息融合、语义搜索、基于知识的问答系统和基于知识的大数据分析和挖掘。

知识图谱不仅提供计算机更好的理解互联网资源的知识内容，同时也提供给计算机更好地组织和管理海量数据资源的结构。以下分别介绍一些核心的知识图谱应用技术：

- **基于知识图谱的大数据融合技术**研究语义标注或者实体链接技术，实现不同资源类型、不同媒体类型的互联网资源的融合、管理与服务。
- **基于知识图谱的语义搜索**实现当前从基于关键词搜索到基于语义的实体和关系搜索，可以直接得到用户感兴趣的客观世界的实体和实体关系信息，而不只是包含关键词的网页文档。其中对于实体类型匹配和实体链接、以及基于实体和关系的排序是核心技术。
- **基于知识图谱的问答系统**通过将用户的提问转换成对结构化知识图谱的查询可以直接得到用户的答案，其中问题理解和基于推理的知识匹配是核心技术。

知识图谱为更好的理解大数据提供了基础设施，通过基于知识图谱的融合技术可以更好的组织和管理大数据的同时，也为大数据分析和挖掘提供的丰富的语义信息，更好地理解大数据的语义，帮助人们制定决策。

3. 技术方法和研究现状

知识图谱的关键技术涉及自然语言处理、数据挖掘和信息检索等多个领域，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。研究方法主要可分为知识驱动和数据驱动两类：知识驱动的方法就是以领域专家的知识与经验为基础，构建能够媲美人类专家知识和问题解决能力的领域知识体系，并通过积累扩充至开放领域；数据驱动的方法则是数理统计为理论基础，以大规模的数据为驱动，通过机器学习和数据挖掘技术自动获取知识，构建大规模的知识图谱。

以下分别从表示、构建和应用三方面介绍相关技术方法和研究现状。

3.1 知识表示

基于符号逻辑的知识表示：是基于符号逻辑的知识表示方法，主要包括逻辑表示法（如一阶逻辑、描述逻辑），产生式表示法和框架表示等。逻辑表示与人类的自然语言比较接近，因此它也是最早使用的一种知识表示方法。基于符号逻辑的知识表示技术虽然可以很好地描述逻辑推理，但是由于在推理中机器生成规则的能力很弱，推理规则的获取需要大量的人力，并且对数据的质量要求较高。在目前大规模数据时代，基于符号逻辑的知识表示已经不能很好地解决知识表示的问题。

万维网内容的知识表示：Tim Berners-Lee 在其著作《Waving the Web》中提出了语义网（Semantic Web）的概念。在语义网中，网络内容都应该有确定的意义，而且可以很容易地被计算机理解、获取和集成。万维网内容知识表示包括半结构基于标记的置标语言 XML[2]、基于 RDF[3]万维网资源语义元数据描述框架和基于描述逻辑的 OWL[4]本体描述语言等；以及当前在工业界得到大规模应用的基于三元组的知识图谱知识表示方法。XML 将网页样式与内容分离，通过为内容置标，便于数据交换；RDF 通过三元组（主体，谓词，客体）描述互联网资源之间的语义关系；互联网语义资源的 OWL 构建在 RDF 之上，是具有更强表达及其解释能力的语言。这些技术使我们可以将机器理解和处理的语义信息发布在万维网上。

表示学习：表示学习的目标是通过机器学习或深度学习将研究对象的语义信息表示为稠密低维的向量。对不同粒度知识单元进行隐式的向量化表示，以支持大数据环境下知识的快速计算。表示学习主要包括张量重构和势能函数的方法：张量重构综合整个知识库的信息，但在大数据环境下张量维度很高，重构的计算量较大；势能函数方法认为关系是头实体向尾实体的一种翻译操作，Bordes 等人提出的 TransE 模型是翻译模型的代表。之后有大量的工

[2] https://www.w3.org/XML/
[3] https://www.w3.org/RDF/
[4] https://www.w3.org/OWL/

作对 TransE 进行扩展和应用，如通过优化向量化表示模型、结合文本等外部信息、应用逻辑推理规则等方法，这些方法进一步提升了表示学习效果。

相比传统的知识表示方法，知识表示学习方法可以显著提升计算效率，有效缓解数据稀疏性，更容易实现不同来源的异质信息融合。因此，表示学习对于知识库的构建、推理和应用具有重要意义。

3.2 知识图谱构建

知识图谱中知识的来源有两类，一类是互联网上分布、异构的海量资源，一类是已有的异构结构化语义资源。从第一类资源中构建知识图谱的方法根据获取知识的类型分为概念层次学习、事实学习、事件学习等，而第二类资源进行的知识图谱构建工作是语义集成。

概念层次学习：概念是人们理解客观世界的线索，不同粒度的概念能够给予知识不同层次的精确程度，概念层次是知识图谱的“骨骼”。概念层次学习就是通过合理的技术抽取知识表示中的概念并确定其上下位关系。概念层次学习多采用基于启发式规则的方法，其基本思路是根据上下位概念的陈述模式从大规模资源中找出可能具有上下位关系的概念对，并对上下位关系进行归纳。另一类是基于统计的概念层次学习方法，假设相同概念出现的上下文也相似，利用词语或实体分布的相似性，通过定义计算特征学习概率模型来得到概念结构。

事实学习：知识图谱中事实以三元组的形式表示，一个知识图谱中事实的数量决定了知识图谱的丰富程度。据不完全统计，Google 知识图谱到目前为止包含了 5 亿个实体和 35 亿条事实。按照知识图谱构建时采用的机器学习方法，事实学习方法可以分为有监督、半有监督及无监督方法。

有监督的事实知识获取方法使用已标注文档作为训练集，可以分为基于规则学习、基于分类标注和基于序列标注方法等。基于规则学习的语义标注方法从带语义标注的语料中自动学习标注规则，利用规则对数据资源进行语义标志，适合对具有比较规范出现的资源的知识获取；基于分类的知识获取方法将知识获取方法转化为分类算法，根据确定的标注特征从标注预料中学习标注模型；基于序列模式标注的方法同时考虑多个语义标志之间的关系，可以提高标注的准确率。还包括其他如考虑层次关系的语义标注的方法等。

半有监督的知识获取方法主要包括自扩展方法(bootstrapping)、弱有监督方法(distant supervision)和开放信息抽取方法(open information extraction)。自扩展方法需要初始的种子实体对，根据这些种子实体对，发现新的语义模板，再对语料进行迭代抽取以发现新的实体对，这种方法的主要问题是语义漂移，代表工作有 Mutual exclusive Bootstrapping, Coupled trainin 和 Co-Bootstrapping。弱监督方法使用知识库中的关系启发式地标注文本，它的问题主要在于训练实例中本身带有大量噪音。开放信息抽取法主要使用自然语言处理方法，无需预先给定要抽取的关系类别，自动将自然语言句子转换为命题。这种方法的主要缺点是在处理复杂句子时效果会受到影响。

无监督知识获取的代表性系统有 KnowItAll, 这套系统具有领域无关特性，可以使用自扩展的方式从大规模互联网信息中抽取语义信息，同时可以自动地评估所抽取信息的可信程度。

语义集成：互联网上已有许多大规模知识库，其中比较著名的有 DBPedia、YAGO 等。然而知识库之间的异构性，对知识在整个语义网上的共享造成了阻碍。语义集成就是通过发现异构知识库中实体间的等价关系，从而实现知识共享的技术。由于知识库多以本体的形式描述，因此语义集成中的主要环节是本体映射。本体匹配的方法主要包括：

- **基于文本信息的方法：**这种方法主要利用本体中实体的文本信息，例如实体的标签和摘要信息。通过计算两个实体字符串之间的相似度(常用的有编辑距离相似度，Jaccard 相似度)，来确定实体之间是否具有匹配关系。
- **基于结构的方法：**这种方法主要利用本体的图结构信息来对本体进行匹配。其中较为代表性的方法有 SimRank 和相似度传播, 这些方法利用本体的图结构，对实体间的相似度进行传播，从而提高对齐的效果。
- **基于背景知识的方法：**这种方法一般使用 DBPedia 或 WordNet 等已有的大规模领域无关知识库作为背景知识来提高匹配效果。例如，Aleksovski 等人利用 DICE 本体

(医学领域的本体)来匹配结构信息缺失的两个与医学相关的本体。

- **基于机器学习的方法：**这种方法将本体匹配问题视为一个机器学习中的分类或优化问题，采取机器学习方法获得匹配结果。例如将本体匹配视为一个贝叶斯决策问题。Niepert 等人将本体匹配问题使用马尔可夫逻辑网络(Markov Logic Network)建模，将本体中的各种信息转化为各种约束条件，并求出最优解.

3.3 知识图谱应用

Google 最初提出知识图谱是为了增强搜索结果，改善用户搜索体验。然而知识图谱的应用远不止这些，基于知识图谱的服务和应用是当前的一大研究热点。按照应用方式可以分为：语义搜索、基于知识图谱问答系统知识问答以及应用平台的研究等。

语义搜索：利用具有良好语义定义的形式，以有向图的方式提供满足用户需求的结构化语义内容。主要包括 RDF 和 OWL 的语义搜索引擎和基于链接数据的搜索等。语义搜索利用建立大规模知识库对用户搜索关键词和文档内容进行语义标注，改善搜索结果，典型的应用包括谷歌的 Knowledge Graph 和国内的百度知心、搜狗的知立方等。

基于知识图谱的问答技术：基于知识库的问答通过对问句的语义分析，将非结构化问句解析成结构化的查询语句，在已有结构化的知识库上查询答案。这类方法依赖于语义解析器的性能，受制于词、短语、从句等不同颗粒度下文本内容歧义、结构歧义的影响，在面对大规模、开放域知识库时，往往性能很低。近两年很多研究者开始研究基于深度学习的知识库问答方法，这类方法更具鲁棒性。但是目前这类方法还只能处理简单、单关系的问题，对于复杂问句的处理效果还是很差，特别是缺乏对于问句的情景感知能力，缺乏对于问句语义细致、个性化的分析。

知识图谱平台技术：近几年，国际很多研究团队投入到知识图谱应用平台的研究中，W3C 倡导的 Linked Open Data 将由互联文档组成的万维网扩展成为由互联数据组成的全球数据及知识共享平台，欧盟第七合作框架下的 LarKC、LOD2、Xlike 项目分别支持建立大规模知识获取和推理、互联数据生成与链接，以及跨语言知识抽取的平台，在包括政府开放数据、智慧医疗、智慧城市在内的很多应用领域获得了成功应用。相比之下国内在知识工程领域起步较晚，目前大多数的知识处理平台还多是数据挖掘或者语义分析的功能，并没有实现支撑建立知识图谱开发平台。

4. 总结及展望

知识图谱技术是知识表示和知识库在互联网环境下的大规模应用，显示出知识在智能系统中重要性，是实现智能系统的基础知识资源。纵观知识图谱研究发展的相关研究现状，以下研究方向或问题将成为未来知识图谱必须攻克的堡垒：

- **融合符号逻辑和表示学习的知识表示：**基于符号逻辑的知识表示使知识具有现实的语义定义，但存在数据稀疏问题，难以实现大规模的知识图谱应用。基于深度学习的知识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表示，方便知识计算。因此，研究融合表示学习与符号逻辑的知识表示理论，使知识既具有显式的语义定义，又便于大数据下的知识计算与推理是知识图谱知识表示一个有前景的研究问题；
- **高精确度大规模知识图谱构建：**人工构造知识图谱存在需要花费大量人工，建立的知识图谱覆盖度不足，而自动的知识图谱构建技术存在知识质量难以应用的问题。随着大数据时代的到来，如何从分布、异构、有噪音、碎片化的大数据中获得高质量的大规模知识图谱，为知识图谱构建带来了机遇，同时也成为一个研究热点。如何构建融合符号逻辑和深度计算的知识获取和推理技术是其中一个有前景的研究问题；
- **知识图谱平台技术：**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，如何建立知识图谱构建的平台，提供知识图谱的构建的管道技术，是未来知识图谱一个研发方向；

- **基于知识图谱的应用研究：**知识图谱虽然已经在海量信息资源的基于知识的语义集成、语义搜索、问题回答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，如何进一步应用知识图谱建立知识驱动的自然语言处理研究方法，基于知识的大数据分析和挖掘是非常值得研究的方向。

综合上述分析，我们有理由相信，随着机器学习、语义分析和篇章理解等相关技术的快速进展，这一人工智能中最具挑战的问题将在可预见的未来得到相当程度的解决，知识图谱的产业化应用前景将更加广阔。

第七章 文本分类与聚类研究进展、现状及趋势

1.任务定义和研究意义

现实世界中人们获取的大部分信息以文本的形式存在，例如书籍、报刊、电子邮件和Web页面等。随着互联网的高速发展，海量文本数据不断产生，这些数据中蕴含大量有用信息。因此，针对这些文本信息的文本挖掘（Text Mining）技术受到人们的广泛关注。文本挖掘是指从这些非结构或半结构化的文本数据中获取高质量的结构化信息的过程。换言之，文本挖掘的目的是从未经处理的文本数据中获取有用知识或信息。典型的文本挖掘任务包括文本分类、文本聚类、概念/实体抽取、情感分析、文档摘要等。

文本分类和聚类是文本挖掘的核心任务，一直以来倍受学术界和工业界的关注。文本分类（Text Classification）任务是根据给定文档的内容或主题，自动分配预先定义的类别标签。文本聚类（Text Clustering）任务则是根据文档之间的内容或主题相似度，将文档集合划分成若干个子集，每个子集内部的文档相似度较高，而子集之间的相似度较低。

文本分类和聚类技术在智能信息处理服务中有着广泛的应用。例如，大部分在线新闻门户网站（如新浪、搜狐、腾讯等）每天都会产生大量新闻文章，如果对这些新闻进行人工整理非常耗时耗力，而自动对这些新闻进行分类或聚类，将为新闻归类以及后续的个性化推荐等都提供巨大帮助。互联网还有大量网页、论文、专利和电子图书等文本数据，对其中文本内容进行分类聚类，是实现对这些内容快速浏览与检索的重要基础。此外，许多自然语言分析任务如观点挖掘、垃圾邮件检测等，也都可以看作文本分类或聚类技术的具体应用。

接下来，本文将着重介绍文本分类和聚类的关键科学问题，具体研究内容，截至目前的研究进展，以及未来的发展趋势。

2.研究内容和关键科学问题

对文档进行分类或聚类，一般需要经过两个步骤：（1）文本表示，以及（2）学习分类、聚类。文本表示是指将无结构化的文本内容转化成结构化的特征向量形式，作为分类或聚类模型的输入。在得到文本对应的特征向量后，就可以采用各种分类或聚类模型，根据特征向量训练分类器或进行聚类。因此，文本分类或聚类的主要研究任务和相应关键科学问题如下。

2.1 构建文本特征向量

构建文本特征向量的目的是将计算机无法处理的无结构文本内容转换为计算机能够处理的特征向量形式。文本内容特征向量构建是决定文本分类和聚类性能的重要环节。

为了根据文本内容生成特征向量，需要首先建立特征空间。其中典型代表是文本词袋（Bag of Words）模型，每个文档被表示为一个特征向量，其特征向量每一维代表一个词项。所有词项构成的向量长度一般可以达到几万甚至几百万的量级。这样高维的特征向量表示如果包含大量冗余噪音，会影响后续分类聚类模型的计算效率和效果。因此，我们往往需要进行特征选择（Feature Selection）与特征提取(Feature Extraction)，选取最具有区分性和表达能力的特征建立特征空间，实现特征空间降维；或者，进行特征转换(Feature Transformation)，将高维特征向量映射到低维向量空间。特征选择、提取或转换是构建有效文本特征向量的关键问题。

2.2 建立分类或聚类模型

在得到文本特征向量后，我们需要构建分类或聚类模型，根据文本特征向量进行分类或聚类。其中，分类模型旨在学习特征向量与分类标签之间的关联关系，获得最佳的分类效果；而聚类模型旨在根据特征向量计算文本之间语义相似度，将文本集合划分为若干子集。

分类和聚类是机器学习领域的经典研究问题。我们一般可以直接使用经典的模型或算法解决文本分类或聚类问题。例如，对于文本分类，我们可以选用朴素贝叶斯、决策树、k-NN、逻辑回归(Logistic Regression)、支持向量机(Support Vector Machine, SVM)等分类模型。对于文本聚类，我们可以选用 k-means、层次聚类或谱聚类(spectral clustering)等聚类算法。

这些模型算法适用于不同类型的数据而不仅限于文本数据。但是，文本分类或聚类会面临许多独特的问题，例如，如何充分利用大量无标注的文本数据，如何实现面向文本的在线分类或聚类模型，如何应对短文本带来的表示稀疏问题，如何实现大规模带层次分类体系的分类功能，如何充分利用文本的序列信息和句法语义信息，如何充分利用外部语言知识库信息，等等。这些问题都是构建文本分类和聚类模型所面临的关键问题。

3. 技术方法及研究现状

接下来，我们将从文本表示和分类聚类模型两个方面，分别介绍主要的技术方法和研究现状。

3.1 文本表示

自然语言文本数据是由词构成的序列。文本的词序列中蕴含了复杂的结构信息和丰富的语义信息。经典的文本分类和聚类模型为了简化文本表示，提出词袋模型（Bag of Words Model）假设，将句子看做词的集合，而忽略了词与词之间的序列信息以及句子结构信息。在词袋模型假设的基础上，向量空间模型(Vector Space Model)成为文本的主要表示方法，向量空间的每一维代表一个词项(词语或 N-Gram)，然后通过 TF-IDF 等方式就可以计算得到文本在向量空间中的表示。

大规模文本中可能出现的词项非常多，但并不是所有词项都可以作为文本特征。为了选取有效文本特征，降低特征空间维度，提高分类聚类的效果与效率，以特征选择(Feature Selection)、特征转换(Feature Transformation)和话题分析(Topic Analysis)为代表的特征降维方法被广泛研究与使用。下面我们将分别介绍这两类典型的特征降维方法。

3.1.1 特征选择

特征选择旨在从已有候选特征中，选取最有代表性和最具区分能力的特征。特征选择需要构造面向特征的评分函数，对候选特征进行评估，然后保留评分值最高的特征。下面是文本分类或聚类中常用的特征评分函数：

文档频率(Document Frequency, DF)是指在整个文本集合中，出现某个特征的文档的频率。其基本思想是，DF 值低于某个阈值的低频特征通常为噪音特征或者信息量较小不具有代表性。因此，我们一般手工确定某个阈值，将低频特征移除，从而有效地降低特征维度，提高分类或聚类效果。作为一种简单高效的特征选择方案，DF 值广泛应用于大规模语料的特征降维。

如果给定事先标注了类别标签的文本集合，我们可以用以下方法计算不同特征的类别区分度。

信息增益(Information Gain)计算新增某个特征后信息熵的变化情况，用以衡量特征的信息量。在计算出每个特征的信息增益后，就可以移除那些信息量较低的特征。

互信息(Mutual information)根据特征与类别的共现情况来计算特征与类别的相关度，具体来说，词项 t 与类别 c 之间的互信息定义如下：

$$I(t,c) = \log \frac{P(t,c)}{P(t)*P(c)} = \log \frac{P(t \wedge c)}{P(t)*P(c)}$$

如果词项与类别没有关联关系，那么两者同时发生的概率P(t,c)接近两者独立发生概率的乘积P(t)＊P(c)，此时互信息值趋近 0；若两者有关联关系，那么两者的联合概率会远大于独立概率的乘积，此时互信息远大于 0。因此，特征的互信息值越高，说明该特征与某个类别的关联程度更紧密，用来进行分类的话区分效果就更好。

卡方统计（χ^2 Statistics）是另一种计算特征与类别关联关系的方法。它定义了一系列词项t与类别c之间共现或不同现的统计量（A、B、C、D），该词项在该类别下的卡方统计值计算公式如下：

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B) + (C + D)}$$

定量研究表明，与 DF 相比，基于标注数据集合选取的特征更具区分性，对文本分类效果提升显著，其中以卡方统计的表现最佳。

3.1.2 特征转换

特征选择通过对所有特征进行重要性排序，选取最重要的特征集合，从而实现特征空间的降维。特征转换则是将高维的特征空间映射或者转换到低维特征空间，从而实现特征的降维。比较典型的特征转换方法包括：

主成分分析(Principal Component Analysis, PCA)也是一种常见的降维方法。PCA 首先计算特征变量之间的协方差矩阵，然后选择协方差矩阵特征值最大的若干个特征向量作为主成分。然后利用这些特征向量，通过线性映射就可以将高维特征映射到低维空间中。

线性判别分析(Linear Discriminant Analysis, LDA)是模式识别领域的经典特征转换方法。它通过将高维特征向量映射到具有最佳区分度的低维空间，来达到压缩特征维度的效果。这种方法能够保证转换后的表示具有最大的类间间距和最小的类内间距，这也意味着新的低维特征空间具有最佳的判别性。

3.1.3 话题分析

文档一般有若干个话题组成，也有研究通过分析文档话题作为文档特征表示。话题分析是文本挖掘领域的重要任务。它假设文档与词语之间存在潜在的语义关系，即话题。话题分析技术往往将文档看成不同话题上的分布，而将每个话题看成不同词语上的分布。话题分析的目标是，利用大规模文档集合，自动学习话题表示，构建“文档-话题”以及“话题-词语”之间的关系。话题分析的代表技术包括：

潜在语义分析(Latent Semantic Analysis, LSA)通过矩阵奇异值分解(Singular Value Decomposition, SVD)对文档-词语的同现矩阵进行分解，得到“文档-话题”矩阵以及“话题-词语”矩阵。由于 LSA 并没有对两个目标矩阵中的取值范围设定限制，不具备概率分布的良好属性。因此，后续提出了基于概率的潜在语义分析方法。

基于概率的潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)是由 Hofmann 等人于 1999 年提出。通过引入概率统计的思想，PLSA 学习得到的“文档-话题”矩阵以及“话题-词语”矩阵具有较好的概率分布属性，能够更直观地计算文档-话题以及话题-词语之间的语义关系，同时也避免了 LSA 中 SVD 的复杂计算过程。由于 PLSA 无法较好对新文档估计话题分布，Blei 等人于 2003 年提出了著名的产生式概率模型隐狄利克雷分布。

隐狄利克雷分布(Latent Dirichlet Allocation, LDA)是一个层次化的贝叶斯模型，通过为文档的话题分布、话题的词语分布分别设置基于 Dirichlet 的先验概率分布，从而使模型具有较好的泛化推理能力，可以为新文档自动估计话题分布。与 PLSA 利用 EM 算法进行参数估计不同，LDA 可以采用更高效的 Gibbs 抽样法和变分推断法来进行参数估计。人们基于 LDA 提出很多新的主题分析模型，例如考虑文档之间关系的 RTM(Relational Topic Model)，考虑主题之间相关性的的 CTM(Correlated Topic Model)、考虑话题随时间演变的 DTM(Dynamic Topic Model)，以及考虑文档作者信息的 Author-Topic Model，等，均得到较为较为广泛的关注与应用。

值得一提的是，以上方法进行话题分析的结果，既可以作为文档特征进行文本分类或聚类，也可以用来分析大规模文档集合中的话题分布与演化情况。这方面的重要应用是**话题检测与跟踪(Topic Detection and Tracking, TDT)**，一般面向新闻媒体，进行新话题发现和已知话题跟踪。以上主题模型均可用来进行有效的话题检测与抽取，而 DTM 等动态主题模型也可以得到同一主题在不同时期的变化情况。

3.2 文本分类模型

近年来，文本分类模型研究层出不穷，特别是随着深度学习的发展，深度神经网络模型也在文本分类任务上取得了巨大进展。我们将文本分类模型划分为以下三类：

3.2.1 基于规则的分类模型

基于规则的分类模型旨在建立一个规则集合来对数据类别进行判断。这些规则可以从训练样本里自动产生，也可以人工定义。给定一个测试样例，我们可以通过判断它是否满足某些规则的条件，来决定其是否属于该条规则对应的类别。

典型的基于规则的分类模型包括决策树(Decision Tree)、随机森林(Random Forest)、RIPPER 算法等。

3.2.2 基于机器学习的分类模型

典型的机器学习分类模型包括贝叶斯分类器(Naïve Bayes)、线性分类器（逻辑回归)、支持向量机(Support Vector Machine, SVM)、最大熵分类器等。

SVM 是这些分类模型中比较有效、使用较为广泛的分类模型。它能够有效克服样本分布不均匀、特征冗余以及过拟合等问题，被广泛应用于不同的分类任务与场景。通过引入核函数，SVM 还能够解决原始特征空间线性不可分的问题。

除了上述单分类模型，以 Boosting 为代表的分类模型组合方法能够有效地综合多个弱分类模型的分类能力。在给定训练数据集合上同时训练这些弱分类模型，然后通过投票等机制综合多个分类器的预测结果，能够为测试样例预测更准确的类别标签。

3.2.3 基于神经网络的方法

以人工神经网络为代表的深度学习技术已经在计算机视觉、语音识别等领域取得了巨大成功，在自然语言处理领域，利用神经网络对自然语言文本信息进行特征学习和文本分类，也成为文本分类的前沿技术。

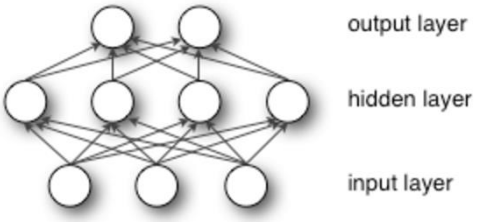


图 1. 多层感知机模型

前向神经网络：多层感知机（Multilayer Perceptron, MLP）是一种典型的前向神经网络。如图 1 所示，它能够自动学习多层神经网络，将输入特征向量映射到对应的类别标签上。通过引入非线性激活层，该模型能够实现非线性的分类判别式。包括多层感知机在内的文本分类模型均使用了词袋模型假设，忽略了文本中词序和结构化信息。对于多层感知机模型来说，高质量的初始特征表示是实现有效分类模型的必要条件。

为了更加充分地考虑文本词序信息，利用神经网络自动特征学习的特点，研究者后续提出了卷积神经网络（Convolutional Neural Network, CNN）和循环神经网络(Recurrent Neural Network, RNN)进行文本分类。基于 CNN 和 RNN 的文本分类模型输入均为原始的词序列，输出为该文本在所有类别上的概率分布。这里，词序列中的每个词项均以词向量的形式作为输入。

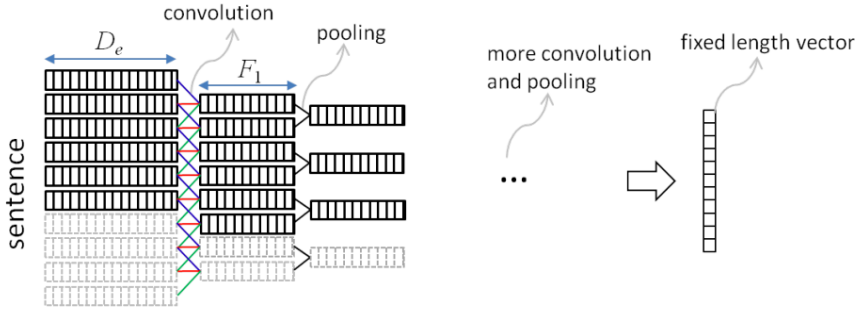


图 2. 卷积神经网络模型

卷积神经网络(CNN)：卷积神经网络文本分类模型的主要思想是，对词向量形式的文本输入进行卷积操作。CNN 最初被用于处理图像数据。如图 2 所示，与图像处理中选取二维域进行卷积操作不同，面向文本的卷积操作是针对固定滑动窗口内的词项进行的。经过卷积层、池化层和非线性转换层后，CNN 可以得到文本特征向量用于分类学习。CNN 的优势在于在计算文本特征向量过程中有效保留有用的词序信息。针对 CNN 文本分类模型有许多改进工作，如基于字符级 CNN 的文本分类模型、将词位置信息加入到词向量。

循环神经网络(RNN)：循环神经网络将文本作为字符或词语序列$\{x_0, \ldots, x_N\}$，对于第t时刻输入的字符或词语x_t，都会对应产生新的低维特征向量s_t。如图 3 所示，s_t的取值会受到x_t和上个时刻特征向量s_{t-1}的共同影响，s_t包含了文本序列从x_0到x_t的语义信息。因此，我们可以利用s_N作为该文本序列的特征向量，进行文本分类学习。与 CNN 相比，RNN 能够更自然地考虑文本的词序信息，是近年来进行文本表示最流行的方案之一。

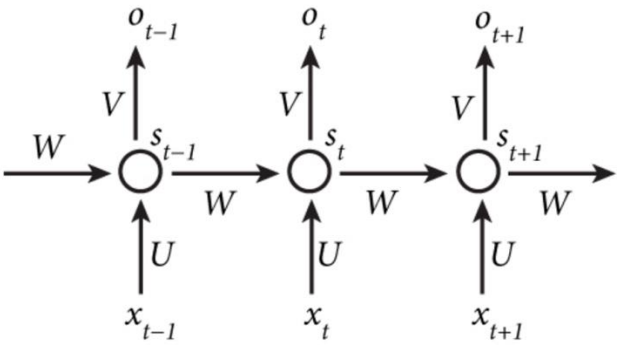


图 3. 循环神经网络模型

为了提升 RNN 对文本序列的语义表示能力，研究者提出很多扩展模型。例如，长短时记忆网络（LSTM）提出记忆单元结构，能够更好地处理文本序列中的长程依赖，克服循环神经网络梯度消失问题。如图 4 是 LSTM 单元示意图，其中引引入了三个门（input gate, output gate, forget gate）来控制是否输入输出以及记忆单元更新。

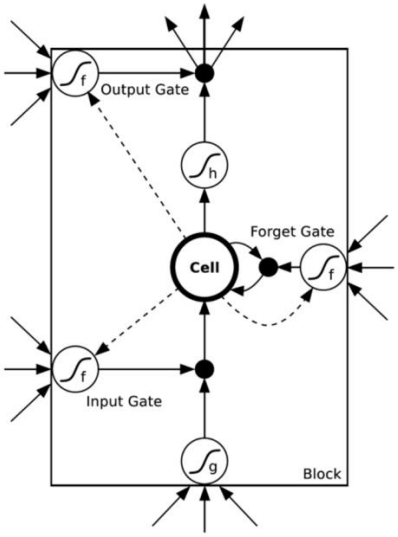


图 4. LSTM 单元示意图

提升 RNN 对文本序列的语义表示能力的另外一种重要方案是引入选择注意力机制(Selective Attention)，可以让模型根据具体任务需求对文本序列中的词语给予不同的关注度。

3.3 文本聚类方法

文本聚类是典型的无监督学习任务，文本聚类的代表算法包括如下两类：

3.3.1 基于距离的聚类算法

基于距离的聚类算法的基本思想是，首先通过相似度函数计算文本间的语义关联度，较为常见的是余弦相似度；然后根据文本间的语义相似度进行聚类。典型的基于距离的聚类算法包括层次法和划分法。层次法通过对原始数据按照不同簇之间的距离进行层次化分解，得到最终的聚类结果，代表算法包括 BIRCH、CURE 等。划分法则是通过对初始的 K 个分组进行迭代更新，直至达到较优的划分为止，K-means 算法就是划分聚类的代表方法。

3.3.2 基于概率模型的聚类方法

主题模型（Topic Model）是典型的基于概率的文本聚类方法。主题建模的思想是对文本集合学习概率生成模型。与基于距离的聚类方法不同，这种基于概率模型的聚类方法假设每篇文章是所有主题（聚集）上的概率分布，而不是仅属于一个聚集。典型的主题模型包括 PLSA 和 LDA 等等。

4. 总结及展望

文本分类与聚类是文本挖掘领域的重要文本分析技术。近些年来随着机器学习特别是深度学习技术的发展，文本分类与聚类也不断得到突破和进展。我们认为，文本分类和聚类领域的研究趋势如下：

- **面向互联网文本的分类聚类：**传统的文本分类与聚类任务更多聚焦在文本自身的分析上，在封闭小数据集合上优化分类模型。随着互联网的发展，许多社交媒体、移动客户端等产生了海量异构的复杂文本信息。这些文本信息包含大量噪音，也具有丰富的结构化或半结构化信息（如网页标签等）。同时，互联网文本信息往往也随着大量的图片、视频等与文本内容相关的多媒体信息。如何恰当地利用这些异构信息，构建高效的适用于互联网文本的分类与聚类模型，是文本挖掘领域面临的重要挑战。
- **神经网络文本分类模型优化：**由于深度神经网络能够更好地处理序列数据，并由分布式特征表示带来强大的表示能力，基于神经网络的文本分类模型取得了目前文本分类的最高水平。然而，神经网络文本分类模型仍有诸多不足。例如，文本表示向量的可解释性差，无法帮助人们有效定位文本类别的相关因素及干扰因素。而且，神经网络模型复杂度较高，需要大量的标注数据才能充分学习，而且需要大量的计算资源才能完成训练。因此，如何建立可解释的神经网络分类模型，如何降低模型学习复杂度，以及如何利用有限的标注样例取得更好分类效果，是深度神经网络分类模型亟待解决的问题。
- **基于神经网络的文本聚类模型：**神经网络模型已经在文本分类上取得了巨大成功。而由于文本聚类一般为无监督的方法，不能直接为神经网络的训练提供有效的监督信号，所以神经网络模型没有在文本聚类任务上取得有效进展。如何充分利用深度神经网络的强大语义表示能力，设计有效的目标函数，建立基于神经网络的文本聚类模型，是文本聚类所面临的挑战。

综上所述，文本分类聚类与机器学习与深度学习技术密切相关。随着相关技术的不断发展，文本分类与聚类效果将能够取得显著提升，并将进一步在互联网智能信息服务中得到广泛应用。

第八章 信息抽取研究进展、现状及趋势

1.任务定义、目标和研究意义

信息抽取（Information Extraction）是指从非结构化/半结构化文本（如网页、新闻、论文文献、微博等）中提取指定类型的信息（如实体、属性、关系、事件、商品记录等），并通过信息归并、冗余消除和冲突消解等手段将非结构化文本转换为结构化信息的一项综合技术。例如，从相关新闻报道中抽取出恐怖事件信息：时间、地点、袭击者、受害人、袭击目标、后果等；从体育新闻中抽取体育赛事信息：主队、客队、赛场、比分等；从论文和医疗文献中抽取疾病信息：病因、病原、症状、药物等。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。

信息抽取是组织、管理和分析海量文本信息的核心技术和重要手段，是大数据时代的使能技术,具有重要的经济和应用意义。随着计算机的普及以及互联网的迅猛发展，大量的信息以数字化文档的形式被存储在计算机里。这些数据与自然资源、人力资源一样，是重要的战略资源，隐含着巨大的经济价值。如何充分组织、管理和利用 Web 发展带来的海量数据，有效解决信息爆炸带来的严重挑战，已经成为了信息科学的核心问题。通过将文本所表述的信息结构化和语义化，信息抽取技术给我们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。目前信息抽取已被广泛应用于舆情监控、网络搜索、智能问答等多个重要领域。

与此同时，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。一直以来，人工智能的关键核心部件之一是构建可支撑类人推理和自然语言理解的大规模常识知识库。然而，由于人类知识的复杂性、开放性、多样性和巨大的规模，目前仍然无法构建满足上述需求的大规模知识库。信息抽取技术通过结构化自然语言表述的语义知识，并整合来自海量文本中的不同语义知识，是构建大规模知识库最有效的技术之一。每一段文本内所包含的寓意可以描述为其中的一组实体以及这些实体相互之间的关联和交互，因此抽取文本中的实体和它们之间的语义关系也就成为了理解文本意义的基础。信息抽取可以通过抽取实体和实体之间的语义关系，表示这些语义关系承载的信息，并基于这些信息进行计算和推理来有效的理解一段文本所承载的语义。

2.研究内容和关键科学问题

信息抽取系统处理各种非结构化/半结构化的文本输入（如新闻网页、商品页面、微博、论坛页面等），使用多种技术（如规则方法、统计方法、知识挖掘方法），提取各种指定的结构化信息（如实体、关系、商品记录、列表、属性等），并将这些信息在不同的层面进行集成（知识去重、知识链接、知识系统构建等）。根据提取的信息类别，目前信息抽取的核心研究内容可以划分为命名实体识别(Named Entity Recognition, NER)、关系抽取(Relation Extraction)、事件抽取和信息集成（Information Integration）。以下分别介绍具体的研究内容。

命名实体识别。命名实体识别的目的是识别文本中指定类别的实体，主要包括人名、地名、机构名、专有名词等的任务。例如，识别“2016 年 6 月 20 日，骑士队在奥克兰击败勇士队获得 NBA 冠军”这句中的地名（奥克兰）、时间（2016 年 6 月 20 日）、球队（骑士队、勇士队）和机构（NBA）。命名实体识别系统通常包含两个部分：实体边界识别和实体分类，其中实体边界识别判断一个字符串是否是一个实体，而实体分类将识别出的实体划分到预先

给定的不同类别中去。命名实体识别是一项极具实用价值的技术，目前中英文上通用命名实体识别（人名、地名、机构名）的 F1 值都能达到 90%以上。命名实体识别的主要难点在于表达不规律、且缺乏训练语料的开放域命名实体类别（如电影、歌曲名）等。

关系抽取。关系抽取指的是检测和识别文本中实体之间的语义关系，并将表示同一语义关系的提及（mention）链接起来的任务。关系抽取的输出通常是一个三元组（实体 1，关系类别，实体 2），表示实体 1 和实体 2 之间存在特定类别的语义关系。例如，句子“北京是中国的首都、政治中心和文化中心”中表述的关系可以表示为（中国，首都，北京），（中国，政治中心，北京）和（中国，文化中心，北京）。语义关系类别可以预先给定（如 ACE 评测中的七大类关系），也可以按需自动发现（开放域信息抽取）。关系抽取通常包含两个核心模块：关系检测和关系分类，其中关系检测判断两个实体之间是否存在语义关系，而关系分类将存在语义关系的实体对划分到预先指定的类别中。在某些场景和任务下，关系抽取系统也可能包含关系发现模块，其主要目的是发现实体和实体之间存在的语义关系类别。例如，发现人物和公司之间存在雇员、CEO、CTO、创始人、董事长等关系类别。

事件抽取。事件抽取指的是从非结构化文本中抽取事件信息，并将其以结构化形式呈现出来的任务。例如，从“毛泽东 1893 年出生于湖南湘潭”这句话中抽取事件{类型：出生，人物：毛泽东，时间：1893 年，出生地：湖南湘潭}。事件抽取任务通常包含事件类型识别和事件元素填充两个子任务。事件类型识别判断一句话是否表达了特定类型的事件。事件类型决定了事件表示的模板，不同类型的事件具有不同的模板。例如出生事件的模板是{人物，时间，出生地}，而恐怖袭击事件的模板是{地点，时间，袭击者，受害者，受伤人数,…}。事件元素指组成事件的关键元素，事件元素识别指的是根据所属的事件模板，抽取相应的元素，并为其标上正确元素标签的任务。

信息集成。实体、关系和事件分别表示了单篇文本中不同粒度的信息。在很多应用中，需要将来自不同数据源、不同文本的信息综合起来进行决策，这就需要研究信息集成技术。目前，信息抽取研究中的信息集成技术主要包括共指消解技术和实体链接技术。共指消解指的是检测同一实体/关系/事件的不同提及，并将其链接在一起的任务，例如，识别“乔布斯是苹果的创始人之一，他经历了苹果公司几十年的起落与兴衰”这句话中的“乔布斯”和“他”指的是同一实体。实体链接的目的是确定实体名所指向的真实世界实体。例如识别上一句话中的“苹果”和“乔布斯”分别指向真实世界中的苹果公司和其 CEO 史蒂夫·乔布斯。

概括说来，信息抽取目前主要面临如下三个关键科学问题：

- 自然语言表达的多样性、歧义性和结构性

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。然而，自然语言表达具有多样性、歧义性和结构性，导致信息抽取任务极具挑战性。自然语言表达的多样性指的是同一种意思可以有多种表达方式，例如“总部位置”这个语义关系可以用“X 的总部位于 Y”，“X 总部坐落于 Y”，“作为 X 的总部所在地，Y…”等等不同的文本表达方式。自然语言表达的歧义性指的是同一自然语言表达在不同上下文中可以表示不同的意思，例如“苹果”这个词在“我买了两斤苹果”和“苹果收购了 Beats”这两句话中指向不同的真实世界实体。自然语言表达的结构性指的是自然语言具有内在结构，例如“我从北京飞到了上海”和“我从上海飞到了北京”虽然使用了相同的词语，但是由于结构不同导致表达的语义不同。因此，如何有效处理自然语言表达的多样性、歧义性和结构性，建立从自然语言文本到无歧义、语义一致且结构明确的目标知识表示的映射，是信息抽取的第一个关键科学问题。

- 目标知识的复杂性、开放性和巨大规模

信息抽取的目标是将文本表达的信息转换为可供计算机处理的知识。然而，人类的知识具有复杂性、开放性以及规模巨大的特点。人类知识的复杂性指的是人类知识多种多样（实体、关系、事件等等），知识和知识之间相互关联、相互交互（关系论元的类别选择约束，事件的模板结构等等），且具有多种不同的结构关系（Taxonomy 结构、Part-of 结构、因果关系网络等等）。人类知识的开放性指的是知识并不是一个封闭的集合，而是随着时间增加、演化和失效。最后，人类知识规模巨大。上述知识的复杂性、开放性和巨大规模给信息抽取带来了巨大的挑战：人类知识的复杂性使得简单模型无法解决信息抽取问题；人类知识的开放性使得现有的有监督方法无法适应开放知识的抽取；人类知识的巨大规模使得无法使用枚举或者人工编写的方式来处理信息抽取。基于上述讨论，我们认为，构建可表示、建模并处理知识复杂性、开放性和巨大规模的技术，是信息抽取的关键科学问题。

- 多源异构信息的融合与验证

在 Web 中，同一条知识往往在不同地方、被不同人、使用不同的方式进行表达。传统的信息抽取研究往往关注单独文本中特定信息的提取，对如何融合来自不同数据源、以不同方式表达、且质量参差不齐的信息缺乏相应的研究。与此同时，互联网上的信息良莠不齐，有很多虚假、错误或者相互冲突的信息，如何验证信息的真假和质量将对后续的应用有极大的影响。因此，如何融合来自多源异构的信息，并验证信息的质量和真假，将碎片化的知识组装成完整的知识模型，是提升信息抽取系统性能和信息抽取实用化的一个重要科学问题。

3.技术方法和研究现状

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。现有信息抽取方法可以从不同的维度进行划分。例如，根据模型的不同，信息抽取方法可以分为基于规则的方法、基于统计模型的方法和基于文本挖掘的方法；根据对监督知识的依赖，信息抽取方法可以划分为无监督方法、弱监督方法、知识监督方法和有监督方法；根据抽取对象的不同，可以划分为实体识别方法、关系抽取方法、事件抽取方法，等等。以下按照模型的维度介绍目前的技术方法和研究现状。

3.1 基于规则的抽取方法

有许多现实生活中的信息抽取任务可以通过一系列抽取规则来进行处理。一个基于规则的抽取系统通常包括一个规则集合和规则执行引擎（负责规则的应用、冲突消解、优先级排序和结果归并）。规则系统在抽取可控且表达规范的信息时非常有效，如文本中的时间、电话号码、邮件地址，以及机器生成页面的结构化信息（如商品页面中的商品记录）。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。

信息抽取系统的规则可以有多种不同的表现形式，如正则表达式、词汇-语法规则、面向 HTML 页面抽取的 Dom Tree 规则等等。抽取规则可以通过人工编写得到或者使用学习方法自动学习得到。为了方便规则的编写，目前已有许多抽取规则开发平台被开放出来，如由 Apache 基金会推出的 UIMA Ruta 系统。与此同时，规则的自动学习也一直是研究界的关注所在，已经有许多自动规则学习方法被提出。

为抽取一类特定信息，通常需要一系列相关的抽取规则。在实际情况中，通常会存在规则相互冲突或规则不一致的情况。因此，抽取规则的管理、冲突消解和优先级排序也是基于规则的信息抽取研究内容。

由于基于规则的方法在扩展性、表达性、组合性和调试性上都具有良好的表现，目前基于规则的方法仍然被广泛使用。如何构建更高效的规则执行引擎、更方便的规则开发平台、更具表达能力的规则表示语言是当前规则抽取系统的研究重点。同时，如何学习更精准的抽取规则、如何消除抽取规则的歧义、如何自动评估规则的效果也一直是基于规则的信息抽取系统的研究难点所在（如 Bootstrapping 系统通常会遇到的语义漂移问题）。

3.2 基于统计模型的抽取方法

自 90 年代以来，统计模型一直是信息抽取的主流方法，有非常多的统计方法被用来抽取文本中的目标信息，如最大熵分类模型、基于树核的 SVM 分类模型、隐马尔可夫模型、条件随机场模型等等。基于统计模型的方法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。例如，条件随机场模型（CRF）是实体识别的代表性统计模型，它将实体识别问题转化为序列标注问题；基于树核的关系抽取系统则将关系抽取任务形式化为结构化表示的分类问题。

近年来，随着深度学习的引入，已有许多深度学习模型被用来进行信息抽取，如卷积神经网络、时序神经网络和递归神经网络。相比传统的统计信息抽取模型，这些深度学习模型无需人工定义的特征模板，能够自动的学习出信息抽取的有效特征；同时神经网络的深度结构使得深度学习模型具有更好的表达能力。因此在标注语料充分的情况下，深度学习模型往往能够取得比传统方法更好的性能。

统计模型往往需要大量的标注语料来学习，这导致构建开放域或 Web 环境下的信息抽取系统时往往会遇到标注语料瓶颈。为解决上述问题，近年来已经开始研究高效的弱监督或无监督策略，如半监督算法、远距离监督算法、基于海量数据冗余性的自学习方法等。

3.3 基于文本挖掘的抽取方法

除非结构化文本之外，Web 中往往还存在大量的半结构的高质量数据源，如维基百科、网页中的表格、列表、搜索引擎的查询日志等等。这些结构往往蕴含有丰富的语义信息。因此，半结构 Web 数据源上的语义知识获取（knowledge harvesting），如大规模知识共享社区（如百度百科、互动百科、维基百科）上的语义知识抽取，往往采用文本挖掘的方法。代表性文本挖掘抽取系统包括 DBPedia、Yago、BabelNet、NELL 和 Kylin 等等。文本挖掘方法的核心是构建从特定结构（如列表、Infobox）到目标语义知识（实体、关系、事件）的映射规则。由于映射规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤。

文本挖掘方法只从容易获取且具有明确结构的语料中抽取知识，因此抽取出来的知识质量往往较高。然而，仅仅依靠结构化数据挖掘无法覆盖人类的大部分语义知识：首先，绝大部分结构化数据源中的知识都是流行度高的知识，对长尾知识的覆盖不足；此外，人们发现现有结构化数据源只能覆盖有限类别的语义知识，相比人类的知识仍远远不够。因此，如何结合文本挖掘方法（面向半结构化数据，抽取出的知识质量高但覆盖度低）和文本抽取方法（面向非结构化数据，抽取出的知识相比文本挖掘方法质量低但覆盖度高）的优点，融合来自不同数据源的知识，并将其与现有大规模知识库集成，是文本挖掘方法的研究方向之一。

3.4 技术现状

经过了三十多年的发展，信息抽取技术已经得到了长足的发展。得益于研究界和工业界的广泛关注，信息抽取技术层出不穷。目前，许多信息抽取技术都已经达到了实用水平。例如，中英文常用命名实体识别技术（主要包括人名、地名、机构名）都达到了 0.90 以上的 F1 值；在 ACE 关系类别上的英文关系抽取技术达到了 0.77 以上的 F1 值。现有的信息抽取技术也已经被广泛应用于知识库构建、问答系统、文本分析、舆情分析等实际应用中。

尽管得到了长足的发展，现有信息抽取技术仍有很长的路要走。在构建成本上，现有高质量抽取系统往往依赖于标注语料，构建成本较高。在构建方式上，现有信息抽取系统依赖于许多预处理模块（如分词、词性标注、句法分析等），缺乏端到端的自动构建方式（随着深度神经网络的使用，已经有所改善），同时也容易受预处理模块性能的影响。在自适应性上，现有抽取系统的自适应性不强，往往在更换语料、更换领域、更换知识类别时会有一个大幅度的性能下降。在系统的性能上，现有信息抽取技术在抽取复杂结构（如事件、Taxonomy）时性能仍然离实用有一定距离。

4. 总结及展望

从 20 世纪 80 年代末以来，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，我们认为信息抽取的发展方向如下：

4.1 面向开放域的可扩展信息抽取技术

目前，绝大部分的信息抽取研究集中在构建更精准的抽取模型和方法，这些方法通常面向预先定义好的语义知识类别，使用标注语料训练模型参数。然而，在构建真实环境下的信息抽取系统时，这些有监督方法往往具有如下不足：1）现有监督模型在更换语料类型之后，往往会有一个大幅度的性能下降；2）现有监督模型无法抽取目标类别之外的语义知识；3）现有监督模型依赖于大规模的训练语料来提升模型性能；4）现有监督模型往往依赖于高复杂度的自然语言处理应用，如句法分析。我们认为现有监督抽取模型无法处理海量异质数据源上开放性和复杂知识的抽取。

基于上述观察，我们认为信息抽取的发展方向之一是构建面向开放域的可扩展抽取技术。具体包括：1）数据规模上的可扩展性：信息抽取系统需要能够高效的处理海量规模的待抽取数据；2）数据源类型上的可扩展性：信息抽取系统需要能够在面对不同类型数据源时取得鲁棒的性能；3）领域的可扩展性：信息抽取系统需要能够方便的从一个领域迁移到另一个领域；4）低构建成本：信息抽取系统的构建不能依赖于基于大量标注语料的有监督技术，而是需要需要基于无监督技术、弱监督技术、知识监督技术等低成本构建技术。

4.2 自学习、自适应和自演化的信息抽取系统

信息抽取数据源具备极强的异构性（来自不同数据源、数据结构具有差异性、不同语言、不同媒体、来自不同的领域等等），同时抽取任务也具备多样性（从实体抽取到知识库生成），这使得未来信息抽取的重点研究方向之一是信息抽取技术的自适应研究。为应对数据源的异构性，需要构建增量式的信息抽取系统，需要研究检测抽取技术是否在当前数据上失效，并根据当前数据源的特点，自学习的构建高性能的抽取系统。同时还要研究数据的变化对抽取系统性能的影响，构建能够自演化的信息抽取系统。为应对抽取任务的多样性，需研究高性能的信息抽取系统管理技术，对各种抽取系统进行高效的管理，根据任务的不同自动组合现有技术，实现高度的模块可重用性。最后，由于信息抽取系统往往达不到 100%的精度，不能达到完全替代人，因此需要研究融合人、信息、和计算机的信息抽取技术平台，充分利用人、计算机各自的优势，大幅提高抽取结果的可用性。

基于上述观察，需要研究面向开放域的数据源，研究自学习的信息抽取技术，在极少人工干预下构建高性能的终生学习信息抽取系统（Never End Learning System）；需要面向演化数据源，研究增量式的信息抽取技术，实现信息抽取系统的性能自检测和自动领域适应；需要研究信息抽取多任务管理技术，面向不同数据源、不同任务，自动的重用之前的信息抽取模块，并利用自学习技术构建高性能的抽取系统。

4.3 面向多源异构数据的信息融合技术

人类的知识是一个巨大的有机整体，知识和知识之间相互关联，相互约束。然而，目前大部分信息抽取系统抽取结果都是碎片化、分散和不一致的，很难构建一个完整的、可解释的复杂知识系统模型。同时，Web 文本规模巨大，质量参差不齐，导致信息抽取的结果存在冗余、冲突和错误，并存在一定程度的不确定性。

基于上述观察，我们认为信息抽取的一项重要研究方向是信息融合技术。信息融合技术一方面可以去除信息抽取结果的冗余、冲突和错误，并减少信息抽取结果的不确定性，使传

统的信息抽取能够到达一个更高层次的性能和结果可用性。同时通过将抽取出来的知识碎片组装成一个完整的全局系统，信息融合技术可以帮助我们构建一个完整的、解释性的知识系统，进而支撑更高层的智能应用，如医学药物分析、经济系统分析等等。

具体地，我们认为需要研究包括跨文档、跨语言和跨媒体三个层次上的融合技术，包括信息置信度衡量、冗余信息去除、解决信息之间的冲突、减少抽取信息的不确定性，并构建自动的缺失信息检测和补全技术。同时需要研究信息融合的全局机制，探索基于信息融合的复杂知识模型构建，如基于本体关系的知识图谱，基于因果关系的复杂因果网络，等等。

最后，纵观 30 余年来信息抽取的现状和发展趋势，我们有理由相信，随着海量数据资源（如 Web）、大规模深度机器学习技术（如深度学习）和大规模知识资源（如知识图谱）的蓬勃发展，信息抽取这一极具挑战性同时也极具实用性的问题将会得到相当程度的解决。同时，随着低成本、高适应性、高可扩展性、可处理开放域的信息抽取研究的推进，信息抽取技术的实用化和产业化将在现有的良好基础之上取得进一步的长足发展。

第九章 情感分析的研究进展、现状及趋势

1. 任务定义、目标和研究意义

狭义的情感分析(sentiment analysis)是指利用计算机实现对文本数据的观点、情感、态度、情绪等的分析挖掘。广义的情感分析则包括对图像视频、语音、文本等多模态信息的情感计算。简单地讲，情感分析研究的目标是建立一个有效的分析方法、模型和系统，对输入信息中某个对象分析其持有的情感信息，例如观点倾向、态度、主观观点或喜怒哀乐等情绪表达。

情感分析是一个典型的交叉学科问题，因此这项工作的开展具有重要的理论与实际意义。从社会学的角度，情感已经成为影响我们行为、人类互相交流的一个重要因素，深入分析情感信息的关键因素、社会影响力、传播模式对于理解情感信息非常重要；从计算科学的角度，如何理解和分析情感信息的表达方式对于提高人机交互、自然语言理解等人工智能任务的能力具有重要意义。两者结合，对情感分析研究的推动与发展，不仅有利于推动相关学科的发展进步，从更深层次上理解和处理情感信息，也能很大程度促进人工智能水平的提高。

随着我国综合国力不断提高，包括互联网、物联网等第三产业成为经济发展的最新引擎。这些应用领域，情感分析是推动其发展进步的重要力量之一，尤其是在舆情管理、商业决策、大数据分析等任务中具有举足轻重的作用。例如，在互联网舆情分析领域，利用情感分析技术可以获知广大网民对于特定事件的意见与观点，及时了解民众的舆论趋势，正确采取引导行动，实现有效有序的社会管理。在反恐领域，通过对社交媒体上极端情感的分析，可以发现潜在的恐怖分子。在商业决策领域，通过对海量用户评论的情感分析与观点挖掘，能够获取可靠的用户反馈信息，了解产品优缺点，同时深刻理解用户的真实需求，实现精准营销。此外，情感分析还被成功应用于股市预测、票房预测、选举结果预测等场景中，充分体现了情感分析在各行各业的巨大作用。

当然，现在的情感分析技术还不完善，仍然面临很多困难和问题。本文对情感分析研究的主要内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势进行简要的介绍。为了使得内容更加聚焦，本文所讨论的情感分析特指狭义的情感分析，即文本情感分析。

2. 研究内容和关键科学问题

情感分析属于自然语言处理的一个子课题，其本质科学问题还是语言文字的语义理解问题，但又具有情感情绪表达的领域特殊性。情感分析包括很多细分任务和应用，在这里我们仅列举主要的研究内容和其中所蕴含的关键科学问题：

2.1 情感资源构建

情感资源在情感分析中扮演非常重要和关键的作用，许多情感分析方法都是高度依赖情感资源的。情感资源通常体现为一些带有情感倾向标注的词或短语，这些资源成为各种情感分析任务的重要资源支撑。

情感资源的研究内容包括如下几个方面：1）类别体系的研究。即从情感倾向、情感表达强弱等方面对情感表达进行区分的类别体系，最常见的包括正、负倾向、主客观，以及细粒度的表达情感强度的强弱区分。2）不同粒度的情感资源研究：从资源词条的文本粒度来说，有词汇级别、短语级别和属性级别，而往往更细的粒度需要的领域知识更多，难度更大。

3）构建方法的研究：从构建方法上来说，一般有手工构建、基于词典扩展和基于语料库构建的方法。

2.2 情感信息的质量分析

由于情感、观点信息一般都发布在社交媒体上，社交媒体由于天然的开放性、虚拟性、随意性等特点，情感信息的质量分析就成为一个很重要的研究课题。质量分析包括几个方面：1)对信息内容本身的判别，包括评论内容可信度分析(Credibility)、垃圾评论识别(Spam)、评论内容的可用性（helpfulness）分析等；2）对信息内容提供者的判别。由于社交媒体传播和口碑对于营销、广告等具有巨大的商业利益，网络水军会被雇佣发布大量虚假性的内容。甄别这些虚假的用户是一个非常具有研究价值的课题。

然而这一任务也具有极大的挑战：首先，对于质量和发布者的判定缺少真实答案，很难从外部信息判断一个内容的虚假性或一个发布者是否具有恶意；其次，对这些问题的建模需要大量的数据作为支撑，而这些数据一般都属于社交媒体的内部数据，这些数据的大量获取往往非常困难。

2.3 情感分类

情感分类是情感分析中非常常见且基础的任务。情感分类通常定义为：对给定的信息内容，依据情感类别体系进行分类或评级。情感分类通常都被当做一个文本分类任务来考虑，而针对情感评分评级问题时，则一般被当做一个序回归任务，有研究表明，情感分析的任务，从分类和分级是有本质不同的。

从输入文本的粒度来看，可以分为篇章级、句子级、短语级、对象和属性级；从所采用的方法来看，可以分成无监督学习、半监督学习、有监督学习方法；从任务的定义上，可以分成主客观分类，情感倾向极性分类，以及情感倾向强度评级（例如 1~5 分，或 1~10 分）。

2.4 情感信息抽取

情感信息抽取是情感分析中的细粒度任务，其核心的目标是抽取观点对象、评价表达、对象和评价之间的搭配等。在观点对象抽取方面，通常有关于观点持有人、观点所针对的目标、对象的细粒度属性等不同层次的情感识别与抽取。在评价表达方面，通常是从输入内容中抽取情感词、情感表达式等内容，包括隐性表达（即通过事实类描述或其它隐晦描述）和显性表达（即具有明显的观点描述）。而在对象和评价之间的搭配抽取方面，则不仅要识别观点对象或属性及针对其的情感评价。

2.5 多模态情感分析

传统的情感分析任务大多是在文本信息上进行的。多模态的情感分析是指从图像、视频、语音、文字等多模态的数据中分析情感、情绪的表达。主要研究内容：1）单模态数据的情感分析，例如针对语音数据、面部视觉信息进行情感情绪识别；2）多模态融合的情感分析，例如从语音+视觉的数据中分析情绪表达，从图像+文字的数据中分析情感表达，从语音+文字的数据中分析观点表达等。

3. 技术方法和研究现状

3.1 规则为主的情感分析方法

规则为主的情感分析方法在早期的情感分析方法中占据了主要的地位，在情感资源构建、情感分类、情感抽取等各种任务中都有以规则为主的方法。在情感资源构建中，一般需要从一个通用的词典或种子词集合开始，例如 WordNet。最常见的是采用自举（Bootstrapping）的方法，从少数已知极性的情感词汇出发，利用同义词、反义词，或其它更精细复杂的句法规则，并结合某些统计量，扩展新的情感词到情感字典中，并进行多次迭代以保证覆盖率，最后结合手工校验形成最后可用的情感词汇。

采用基于规则的方法进行情感分类是一类较为早期和简单的方法。这类方法通常利用情感资源中的词条，结合否定、转折、递进等句法规则，对文本的情感极性进行判别。基于规则的情感分类方法虽然简单，但其在情感资源丰富的一些特定领域上表现得很好。

在情感抽取方面，基于规则的方法也是最常使用的一类方法。用户在评价某一对象时通常还会使用一个与之搭配的情感词对其进行描述。基于这一观测，已有研究者提出了利用句法依存关系来识别观点对象和评价词对。同时，也有研究者使用了短语级的句法依存工具来对句子进行分析，从而能够得到短语级别的观点对象和评价词。 通常，这类方法通过句法分析获得评论文本的句法树，在此基础上人工设定 3 类句法依存关系模板，包括情感词与情感词之间的模板、情感词和观点对象之间的模板以及观点对象和观点对象之间的模板。基于这些模板，这类方法采用迭代抽取的方法，从事先给定的种子词出发，不断寻找更多匹配的观点对象和情感词。

基于规则的方法其特点是简单，但往往也需要比较多的资源，例如情感词汇资源，词性、句法、语法规则，需要耗费大量的劳动力从数据中总结和挖掘规则，其中不可避免需要介入手工检查的工作，属于劳动密集型的方法。

3.2 传统机器学习的情感分析方法

最早采用机器学习进行情感分类的工作采用 unigram 为基本特征，SVM 或朴素贝叶斯作为分类器。而作为机器学习建模的关键步骤之一，各种特征如词性、情感词汇、句法依赖、情感变换词等被广泛地研究和使用。从方法上而言，除了常见的 SVM、朴素贝叶斯等模型外，有图分割方法，以及非负矩阵分解方法等。从特征的使用上，句法可以与传统特征结合在一起提高分类性能；而上下文的显著性和情感转换词（类似“not，no，never，neither”之类的词）等特征对情感分类也很有裨益。另外，语言学的知识库资源或情感资源也有助于进行情感分类。从少量的标注数据出发，结合大量的未标注数据，半监督学习或主动学习方法在情感分析中也被广泛采纳。

情感分类根据所处理文本的粒度不同，可以区分为文档级、句子级、属性级等。有研究者提出层次化的序列标注学习模型（类似 CRF），进行句子级和文档级的情感极性分类。进一步，句子之间的对比、转折、递进等上下文关系（discourse）可以在情感分类中得到充分利用。已有研究者采用后验正则化学习框架，将类似的 discourse 关系建模为约束，对用户评论中的句子进行情感分类取得很好的效果。在属性级别的情感分类中，通常方法多采用有监督学习方法和基于字典的无监督学习方法。在有监督学习方法中，句子级别或子句级别的情感分类方法都是可用的，一个关键的不同是属性级情感分类任务需要识别一个情感表达式所作用的范围。基于字典的分类方法在属性级的情感分类中表现得很好，可以避免大量的数据标注。简单利用一个情感字典、复合表达式、情感规则、依存句法树等作为特征，同时考虑情感转化词、转折连接词等特殊的句法现象，利用简单的机器学习算法就可以取得不错的结果。

在情感信息抽取方面，除了许多基于句法、语法规则的情感抽取工作外，最常见的做法

是把情感信息抽取当做一个序列标注问题来处理。这一类方法通常基于监督学习算法，因此需要人工标注的数据来训练模型。目前应用最为广泛的序列标注算法是隐马尔科夫模型(HMM)和条件随机场模型(CRF)。其中所用到的特征包括：词形、词性、句法依存关系、词距离、是否为主观句等。

近年来主题模型也成为情感抽取的一类重要方法。在这一类方法中观点对象和情感词都被当作是主题信息。一个主题中往往包含了数个概率较高的词，因此这类方法在抽取的同时也完成了词的聚类。最具代表性的方法是基于 pLSA 的特征-情感混合模型。这一方法同时包含了特征主题模型和情感分类模型。也有研究工作通过改进传统的 LDA（Latent Dirichlet Allocation）模型给出一种多粒度的主题模型。这一算法利用全局主题来识别出评论中的实体，利用局部主题来识别出实体中的特征。另外，也有研究者采用半监督的联合主题模型，它能够允许用户提供一些种子词作为输入，从而抽取出满足用户意图的特征和评价词。

3.3 基于深度学习的情感分析方法

神经网络模型的复兴使得深度学习在语音、图像、文本处理中获得了广泛的应用。在文本方面，针对不同的文本粒度，也提出一系列的深度学习模型，包括词向量表示学习、句子级表示学习（循环神经网络、递归神经网络、卷积神经网络等）、篇章级表示学习等（具体请参见《语言表示与深度学习研究进展、现状与未来发展趋势》一章内容）。在情感分析任务上也不例外，现有的深度学习方法得到了广泛的应用。

首先是词向量的表示。已有研究者在词向量的表示学习基础上，加入情感相关的目标函数，进行联合训练，以期望得到与情感信息相关的词向量表示。另外，在情感分类任务中，通常形容词扮演更重要的角色，已有研究者提出了根据词性选择合成函数，以及学习一个词性的嵌入向量，根据子节点向量、词性向量合成父节点的向量。

其次，采用自动编码器进行文本的表示学习有两种常见的类型。第一种简单的编码器，将文本的词袋表示（词表上的稀疏向量表示）转成隐藏层上的表示，学习的目标是最小化原始输入和重构表示之间（隐藏层表示经过非线性变换得到）的误差。这种方法广泛地应用于领域自适应、跨语言的表示或跨模态的数据表示。面对情感分析任务，现有研究者已经把情感分类和领域分类的监督信息加入到优化目标函数中，使得所得到的表示具有一定的情感表达的特点。

第三，面对句子级情感分析任务。Socher 等人提出了一种在句法成分树上进行递归编码的深度学习模型，通过在每个内节点上加入情感标注的监督信息，和重构误差一起进行优化，在句子级别的情感分类上较传统词袋模型获得了大幅提高。此外，卷积神经网络（CNN）是文本处理中较为广泛使用的深度学习模型，通常用来学习句子级别或更长粒度的表示。这些方法在处理文本的情感分类中，在标准评测中已经被证明可以获得了很好的性能。循环神经网络（Recurrent NN，ReNN）和长短期记忆模型（Long-short term memory）由于可以刻画序列单元之间的依赖和影响，因而具有很好的序列建模能力。虽然一般情况下 LSTM 描述常见序列，但这个模型同样可以应用到树的结构上，在情感分类上的性能比递归神经网络模型（Recursive NN）也更好。

总体来说，目前，基于深度学习的模型几乎占领了自然语言处理的各个领域和任务，但是有针对性考虑情感表达的工作还并不多，说明已有研究的重心还停留在通用的文本语义表示和学习上。绝大多数方法把深度学习模型当做一个黑盒子，把学习看成是端到端的处理过程，至于情感表达的内部机制和原理，以及情感表达与一般语义表达的不同则在很大程度上被忽视了，而这正是可以深入开展研究的方向。

4. 技术展望与发展趋势

情感分析经过十多年的发展，在某些领域上（例如产品评论、影评、宾馆、餐馆等）已经取得了相对成熟的发展和应用，在某些领域上达到了可完全实用的水准，但从一般意义上

来说，情感分析还需要进行长期研究和探索，其最本质的难题还是语言文字的理解问题，依然存在非常多的挑战和待解的问题。

4.1 面向社交媒体开放域文本的情感分析

针对社交媒体开放域文本的情感分析任务仍极具挑战性。由于微博、微信等社交媒体上用户生成文本的开放性、自由性和不规范性，现有方法针对社交媒体开放域文本的情感分析效果并不理想。在近几年 SemEval 所组织的面向 Twitter 的情感褒贬分类评测中，表现最好的参赛队伍也只取得了不到 70%的 F 值，远低于人们预期。针对社交媒体的开放话题进行情感分析的任务具有几个特点：评论对象或属性更加难以抽取，表达更加隐晦，甚至不存在明显属性描述词；观点表达更加多样，许多话题不存在明显的观点评价词；理解情感表达需要更多的上下文，例如评论、转发、反讽中需要通过上下文才能对内容进行充分理解。这些问题还都没有得到很好的解决，值得深入探索。

4.2 基于上下文感知的情感分析

上下文感知的情感分析要求在理解当前内容时候，考虑各种形式的上下文，例如句子之间 discourse 关系，微博转发内容的父亲微博，甚至基于社交媒体的上下文（如用户之间的关系、用户的背景信息等）。有许多重要的任务还需要深入研究：1）基于上下文感知的情感资源构建方法；2）基于上下文相关的情感分类，包括篇章级、句子级、对象级、对象属性级、社交媒体的上下文。

4.3 跨领域跨语言情感分析

跨领域跨语言的情感语义计算问题没有得到很好的解决。情感语义计算极大依赖于情感资源（包括情感词典与标注语料），而情感资源又通常跟领域、语言密切相关。尽管业界针对特定领域特定语言下的集中研究能够为该领域该语言积累丰富的情感资源，但是由于社交媒体上用户生成文本涉及众多的不同领域，以及不同的语种（例如中文、英文、日文，以及少数民族语言等），对于多数领域或语言而言均缺乏高质量的情感资源，这严重阻碍了情感语义计算在这些领域或语言内的研究进展与应用。亟待提出崭新的跨领域跨语言文本情感计算理论与方法，破除领域或语言壁垒。

4.4 基于深度学习的端到端情感分析

所谓端到端的方法，就是把数据输入作为一端，监督信息（数据标注信息）作为另一端，而把学习系统作为黑盒子不考虑中间的细节。但我们认为，对于语言文字这种高度概括和抽象数据的处理，简单的端到端并不能完全解决问题，还需要考虑更多语言学知识。包括：1）情感词典如何有效地利用在端到端学习方法中。2）句法（例如词性）、语法、语义信息如何有效地结合在端到端的深度学习方法中。因此，将语言学知识和深度学习方法进行深度结合，有可能形成情感分析方法新的性能突破。

4.5 新的情感分析任务

随着情感分析的发展，一些新的情感分析任务也不断涌现。情感解释分析、反讽和立场分析就是近些年研究的一些热点方向。

- **情感解释：** 情感解释就是挖掘与分析观点情感的原因。导致情感出现的原因非常复杂，用户对情感原因的解释也千差万别，可以为一个短语、一个句子或一段文本，甚至隐式的提及和表达。特别是在社交媒体上，面对热门事件或开放性话题，如何分析群体情感的演变模式和原因分析是一个难点问题。
- **反讽分析：**反讽是社交媒体上一类特殊的语言现象，网民有时候会利用反讽来表达与文本字面相反的语义或情感倾向。反讽的分析和检测具有非常高的挑战性，仅从字面理解内容会得到完全相反的分析结果。如何探究反讽表达的语言学机制和特点，从反讽文本中提炼有用特征，包括词汇、句法、情感特征等，成为一个前沿的研究课题。
- **立场分析：**立场分析目标是识别出讨论或辩论双方的所持立场，这一任务与情感分析密切相关，但相比情感分析，立场分析更具有挑战性。目前关于立场分析的研究主要针对辩论网站内容，而对微博等社交媒体文本的立场分析研究刚开始起步。由于微博文本长度短、语言表达自由，且回复关系非常稀疏，这些因素共同导致了针对微博的立场分析效果较差。特别地，面向中文的立场分析几乎处于空白，因此业界亟需为中文立场分析任务构建数据资源、举办评测活动，推动立场分析基础理论与方法的研究与应用。

第十章 自动文摘研究进展、现状及趋势

1.任务定义、目标和研究意义

随着互联网与社交媒体的迅猛发展和广泛普及，我们进入了一个信息爆炸的时代。网络上包括新闻、书籍、学术文献、微博、微信、博客、评论等在内的各类型文本数据剧增，给用户带来了海量信息，也带来了信息过载的问题。用户通过谷歌、必应、百度等搜索引擎或推荐系统能获得大量的相关文档，但用户通常需要花费较长时间进行阅读才能对一个事件或对象进行比较全面的了解。如何将用户从长篇累牍的文字阅读中解放出来是大数据时代面临的一个挑战，自动文摘技术则是应对该项挑战的一件利器。

具体来说，自动文摘（又称自动文档摘要）是指通过自动分析给定的一篇文档或多篇文档，提炼、总结其中的要点信息，最终输出一篇长度较短、可读性良好的摘要（通常包含几句话或数百字），该摘要中的句子可直接出自原文，也可重新撰写所得。简言之，文摘的目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的文字描述。用户可以通过阅读简短的摘要而知晓原文中所表达的主要内容，从而大幅节省阅读时间。

自动文摘研究的目标是建立有效的自动文摘方法与模型，实现高性能的自动文摘系统。近二十年来，业界提出了各类自动文摘方法与模型，用于解决各类自动摘要问题，在部分自动摘要问题的研究上取得了明显的进展，并成功将自动文摘技术应用于搜索引擎、新闻阅读等产品与服务中。例如谷歌、百度等搜索引擎均会为每项检索结果提供一个短摘要，方便用户判断检索结果相关性。在新闻阅读软件中，为新闻事件提供摘要也能够方便用户快速了解该事件。2013 年雅虎耗资 3000 万美元收购了一项自动新闻摘要应用 Summly，则标志着自动文摘技术的应用走向成熟。

但是，自动文摘技术还远远谈不上完美，在多文档摘要、综述自动生成等任务上还面临相当多的挑战和难题，需要广大科研工作者继续努力探索。本文将对自动文摘研究的主要内容、面临的关键科学问题，以及当前采用的主要技术、现状以及未来的发展趋势逐一进行介绍。

2.研究内容和关键科学问题

自动文摘可看作是一个信息压缩过程，将输入的一篇或多篇文档压缩为一篇简短的摘要，涉及到对输入文档的理解、要点的筛选，以及文摘合成这三个主要步骤。本报告主要关注如下的研究内容和涉及的关键科学问题：

2.1 要点筛选

文档中的重要信息可以通过要点来体现，如何从冗杂的文本信息中筛选出要点，是自动文摘系统能否成功的先决条件。

要点筛选本身又牵涉到如下科学问题：

- **如何表达要点信息？**目前各类文摘系统中采用了不同粒度的信息单元来表示要点信息，例如词汇、短语、依存关系、句子、甚至语义图等。不同的选择会影响要点筛选的可靠性，同时会影响后续的文摘合成步骤，但目前为止上述信息单元并没有绝对的优劣之分。
- **如何评估信息单元的重要性？**输入文档中通常包含大量的信息单元，无论是词汇、短语还是句子，我们需要找到一种合适的评估方法对每个信息单元进行重要性评估，

从大量信息单元中发现最重要的若干个，为后续文摘合成提供输入。

2.2 文摘合成

自动文摘系统会根据要点筛选的结果进行摘要的合成，产生最终的摘要。文摘合成步骤需要保证摘要具有良好的要点覆盖性与可读性，且满足摘要长度的限制。

文摘合成本身则牵涉到如下科学问题：

- **采用抽取式还是生成式方法？**抽取式方法基于原文中已有的句子进行文摘合成，所产生的摘要语句通顺，这也是目前大多数自动文摘系统所采用的方法。生成式方法则直接生成摘要语句，能够得到更加凝练的语句，但语句通顺性不能得到保障。还有一些方法允许对原文语句进行一定的压缩或融合，可以看作是一种混合方法。
- **如何评估摘要的可读性？**摘要可读性是衡量摘要质量的一个重要性质，能够严重影响读者对摘要的主观感受，而摘要通常由多个句子所组成，摘要的可读性不仅依赖于每个句子的通顺性，还依赖于多个句子之间的连贯性，这两部分目前均难以准确建模与评估。
- **如何同时满足摘要的多种性质要求？** 如前所述，一篇高质量的摘要需要满足多种性质与约束，早期的自动文摘系统采用贪心的处理方式，分步骤逐一考虑摘要的不同性质，而最新的自动文摘系统则力图在统一的优化框架下同时考虑多种性质，从而获得更优的摘要结果。

3. 技术方法和研究现状

自动文摘的研究在图书馆领域和自然语言处理领域一直都很活跃，最早的应用需求来自于图书馆。图书馆需要为大量文献书籍生成摘要，而人工摘要的效率很低，因此亟需自动摘要方法取代人工高效地完成文献摘要任务。随着信息检索技术的发展，自动文摘在信息检索系统中的重要性越来越大，逐渐成为研究热点之一。经过数十年的发展，同时在 DUC 与 TAC 等自动文摘国际评测的推动下，文本摘要技术已经取得长足的进步。国际上自动文摘方面比较著名的几个系统包括 ISI 的 NeATS 系统，哥伦比亚大学的 NewsBlaster 系统，密歇根大学的 NewsInEssence 系统等。

自动文摘所采用的方法从实现上考虑可以分为抽取式摘要（extractive summarization）和生成式摘要（abstractive summarization）。抽取式方法相对比较简单，通常利用不同方法对文档结构单元（句子、段落等）进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成摘要。而生成式方法通常需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，利用自然语言生成技术生成新的摘要句子。目前的自动文摘方法主要基于句子抽取，也就是以原文中的句子作为单位进行评估与选取。抽取式方法的好处是易于实现，能保证摘要中的每个句子具有良好的可读性。

为解决如前所述的要点筛选和文摘合成这两个关键科学问题，目前主流自动文摘研究工作大致遵循如下技术框架：

内容表示 → 权重计算 → 内容选择 → 内容组织

首先将原始文本表示为便于后续处理的表达方式，然后由模型对不同的句法或语义单元进行重要性计算，再根据重要性权重选取一部分单元，经过内容上的组织形成最后的摘要。

3.1 内容表示与权重计算

原文档中的每个句子由多个词汇或单元构成，后续处理过程中也以词汇等元素为基本单位，对所在句子给出综合评价分数。以基于句子选取的抽取式方法为例，句子的重要性得分由其组成部分的重要性衡量。由于词汇在文档中的出现频次可以在一定程度上反映其重要性，

我们可以使用每个句子中出现某词的概率作为该词的得分，通过将所有包含词的概率求和得到句子得分。也有一些工作考虑更多细节，利用扩展性较强的贝叶斯话题模型，对词汇本身的话题相关性概率进行建模。

一些方法将每个句子表示为向量，维数为总词表大小。通常使用加权频数作为句子向量相应维上的取值。加权频数的定义可以有多种，如信息检索中常用的词频-逆文档频率（TF-IDF）权重。也有研究工作考虑利用隐语义分析或其他矩阵分解技术，得到低维隐含语义表示并加以利用。得到向量表示后计算两两之间的某种相似度（例如余弦相似度）。随后根据计算出的相似度构建带权图，图中每个节点对应每个句子。在多文档摘要任务中，重要的句子可能和更多其他句子较为相似，所以可以用相似度作为节点之间的边权，通过迭代求解基于图的排序算法来得到句子的重要性得分。

也有很多工作尝试捕捉每个句子中所描述的概念，例如句子中所包含的命名实体或动词。出于简化考虑，现有工作中更多将二元词（bigram）作为概念。近期则有工作提出利用频繁图挖掘算法从文档集中挖掘得到深层依存子结构作为语义表示单元。

另一方面，很多摘要任务已经具备一定数量的公开数据集，可用于训练有监督打分模型。例如对于抽取式摘要，我们可以将人工撰写的摘要贪心匹配原文档中的句子或概念，从而得到不同单元是否应当被选作摘要句的数据。然后对各单元人工抽取若干特征，利用回归模型或排序学习模型进行有监督学习，得到句子或概念对应的得分。文档内容描述具有结构性，因此也有利用隐马尔科夫模型（HMM）、条件随机场（CRF）、结构化支持向量机（Structural SVM）等常见序列标注或一般结构预测模型进行抽取式摘要有监督训练的工作。所提取的特征包括所在位置、包含词汇、与邻句的相似度等等。对特定摘要任务一般也会引入与具体设定相关的特征，例如查询相关摘要任务中需要考虑与查询的匹配或相似程度。

3.2 内容选择

无论从效果评价还是从实用性的角度考虑，最终生成的摘要一般在长度上会有限制。在获取到句子或其他单元的重要性得分以后，需要考虑如何在尽可能短的长度里容纳尽可能多的重要信息，在此基础上对原文内容进行选取。

3.2.1 贪心选择

可以根据句子或其他单元的重要性得分进行贪心选择。选择过程中需要考虑各单元之间的相似性，尽量避免在最终的摘要中包含重复的信息。最为简单常用的去除冗余机制为最大边缘相关法，即在每次选取过程中，贪心选择与查询最相关或内容最重要、同时和已选择信息重叠性最小的结果。也有一些方法直接将内容选择的重要性和多样性同时考虑在同一个概率模型框架内，基于贪心选择近似优化似然函数，取得了不错的效果。

此后有离散优化方向的研究组介入自动文摘相关研究，指出包括最大边缘相关法在内的很多贪心选择目标函数都具有次模性。记内容选取目标函数为 $F(S)$，其自变量 S 为待选择单元的集合；次模函数要求对于 $\forall S \subseteq T \subseteq U \setminus u$，以及任意单元 u，都满足如下性质：

$$F(S \cup \{u\}) - F(S) \geq F(T \cup \{u\}) - F(T).$$

这个性质被称为回报递减效应（diminishing returns），很符合贪心选择摘要内容的直觉：由于每步选择的即时最优性，每次多选入一句话，信息的增加不会比上一步更多。使用特定的贪心法近似求解次模函数优化问题，一般具备最坏情况近似比的理论保证。而实际应用中研究发现，贪心法往往已经可以求得较为理想的解。由于贪心法易于实现、运行效率高，基于次模函数优化的内容选择在近年得到了很多扩展。多种次模函数优化或部分次模函数优化问题及相应的贪心解法被提出，用于具体语句或句法单元的选取。

3.2.2 全局优化

基于全局优化的内容选择方法同样以最大化摘要覆盖信息、最小化冗余等要素作为目标，同时可以在优化问题中考虑多种由任务和方法本身的性质所导出的约束条件。最为常用的形式化框架是基于 0-1 二值变量的整数线性规划。最后求解优化问题得到的结果中如果某变量取值为 1，则表示应当将该变量对应的单元选入最后的摘要中。

由于整数线性规划在计算复杂性上一般为 NP-难问题，此类方法的求解过程在实际应用中会表现较慢，并不适合实时性较高的应用场景。有研究工作将问题简化后使用动态规划策略设计更高效的近似解法。也有少量研究工作尝试在一部分特例下将问题转化为最小割问题快速求解，或利用对偶分解等技术将问题化为多个简单子问题尝试求得较好的近似解。更为通用的全局优化加速方案目前仍是一个开放问题。

3.3 内容组织

3.3.1 内容简化与整合

基于句子抽取得到的语句在表达上不够精练，需要通过语句压缩、简化、改写等技术克服这一问题。在这些技术中相对而言较为简单的语句压缩技术已经广泛被应用于摘要内容简化。现行主要做法基于句法规则或篇章规则，例如如果某重要性较高的短语需要被选择用于构成摘要，那么该短语所修饰的中心词也应当被选择，这样才能保证得到的结果符合语法。这些规则既可以直接用于后处理步骤衔接在内容选取之后进行，也可以用约束的形式施加在优化模型中，这样在求解优化问题完毕后就自然得到了符合规则的简化结果。局部规则很容易表达为变量之间的线性不等式约束，因此尤其适合在前面提到的整数线性规划框架中引入。另外，关于语句简化与改写方面目前也有相对独立的研究，主要利用机器翻译模型进行语句串或句法树的转写。由于训练代价高以及短语结构句法分析效率和性能等诸多方面原因，目前很少看到相关模块在摘要系统中的直接整合与应用。

一些非抽取式摘要方法则重点考虑对原句信息进行融合以生成新的摘要语句。基于句法分析和对齐技术，可以从合并后的词图直接产生最后的句子，或者以约束形式将合并信息引入优化模型等方式来实现。

国际上还有部分研究者尝试通过对原文档进行语义理解，将原文档表示为深层语义形式（例如深层语义图），然后分析获得摘要的深层语义表示（例如深层语义子图），最后由摘要的深层语义表示生成摘要文本。最近的一个尝试为基于抽象意义表示（Abstract Meaning Representation, AMR）进行生成式摘要。这类方法所得到的摘要句子并不是基于原文句子所得，而是利用自然语言生成技术从语义表达直接生成而得。这类方法相对比较复杂，而且由于自然语言理解与自然语言生成本身都没有得到很好的解决，因此目前生成式摘要方法仍属于探索阶段，其性能还不尽如人意。

3.3.2 内容排序

关于对所选取内容的排序，相关研究尚处于较为初级的阶段。对于单文档摘要任务而言，所选取内容在原文档中的表述顺序基本可以反映这些内容之间正确的组织顺序，因此通常直接保持所选取内容在原文中的顺序。而对于多文档摘要任务，选取内容来自不同文档，所以更需要考虑内容之间的衔接性与连贯性。早期基于实体的方法通过对实体描述转移的概率建模计算语句之间的连贯性，据此找到一组最优排序的问题很容易规约到复杂性为 NP-完全的旅行商问题，精确求解十分困难，因此多种近似算法已经被应用于内容排序。未来随着篇章分析、指代消解技术的不断进步，多文档摘要中的语句排序问题也有机会随之产生更好的解决方案。

3.4 端到端摘要

随着深度学习技术在分布式语义、语言模型、机器翻译等任务上取得了一系列突破性成果，相关方法在文摘任务上的应用研究也受到广泛关注。基于编码器-解码器（encoder-decoder）架构的序列到序列学习模型（sequence-to-sequence learning）目前最为流行，因为可以避免繁琐的人工特征提取，也避开了重要性评估、内容选择等技术点的模块化，只需要足够的输入输出即可开始训练。但这些方法因为参数规模较大或结构复杂，需要比传统方法规模远远更大的训练语料，加上当前主流的循环神经网络（recurrent neural networks, RNN）框架并不能够有效对长文档进行语义编码，因此目前的相关研究大多本质上仅仅是在做语句级的简化和标题生成任务。极少数近期工作同时在同一个神经网络框架里开始考虑句子选取和摘要生成，尝试对语句层次进行编码并在此基础上引入注意机制，但效果尚未能明显改善传统方法已经能够取得的性能。

3.5 技术现状

相比机器翻译、自动问答、知识图谱、情感分析等热门领域，自动文摘在国内并没有受到足够的重视。国内早期的基础资源与评测举办过中文单文档摘要的评测任务，但测试集规模比较小，而且没有提供自动化评价工具。2015 年 CCF 中文信息技术专委会组织了 NLPCC 评测，其中包括了面向中文微博的新闻摘要任务，提供了规模相对较大的样例数据和测试数据，并采用自动评价方法，吸引了多支队伍参加评测，目前这些数据可以公开获得。但上述中文摘要评测任务均针对单文档摘要任务，目前还没有业界认可的中文多文档摘要数据，这在事实上阻碍了中文自动摘要技术的发展。

近些年，市面上出现了一些文本挖掘产品，能够提供中文文档摘要功能（尤其是单文档摘要），例如方正智思、拓尔思(TRS)，海量科技等公司的产品。百度等搜索引擎也能为检索到的文档提供简单的单文档摘要。这些文档摘要功能均被看作是系统的附属功能，其实现方法均比较简单。

4. 总结与展望

自动文摘是自然语言处理领域的一个重要研究方向，近 60 年持续性的研究已经在部分自动文摘任务上取得了明显进展，但仍需突破很多关键技术，才能提高其应用价值、扩大其应用范围。

展望未来，以下研究方向或问题值得业界关注：

- **多语言自动文摘资源建设：**目前的自动文摘资源总体上偏少，无论是数据，还是工具与系统。DUC 和 TAC 提供的英文评测数据规模普遍较小，一方面会影响评测结果的准确性，另一方面也无法为统计学习方法尤其是深度学习方法提供充足的训练数据。对于包括中文在内的其他语言，自动文摘资源更是匮乏，严重影响了这些语言中自动文摘技术的发展。业界需要投入更多的人力物力来建设多语言自动文摘资源，这对自动文摘的研究将起到重大的推动作用。
- **自动文摘评价方法的完善：**目前的自动文摘评价方法需要进一步完善，尤其是自动评价方法。基于词汇重叠程度的 ROUGE 评价方法虽然被广泛采用，但质疑声不断，业界需要提出更加合理的自动评价准则，能综合评估摘要的多种性质，将有利于推动自动文摘的研究进展。
- **基于自然语言生成的自动文摘：**生成式摘要方法更符合人类撰写摘要的习惯，但自然语言生成技术的复杂性和不成熟阻碍了生成式摘要方法的研究进展，而随着深度学习技术在自然语言生成问题上的逐步应用，自然语言生成技术也不再高高在上，这给生成式摘要带来了希望和机遇，未来几年将会有越来越多的研究者基于深度学

习技术从事生成式摘要方法的研究，也有望取得重要进展。

- **篇章信息和语义信息的有效利用：**现有方法利用的信息主要基于由统计频数或出现位置所反映的重要性度量，一般比较表层，而忽视了对文档篇章信息与语义信息的利用。文档本身的语义表达具备很强的结构性，各语义单元之间存在紧密联系，这一点在目前提出的结构预测模型中也几乎没有考虑。另一方面，应尽可能保证最后抽取或生成的摘要在描述上前后一致、表达连贯。因此，对文档篇章与语义信息的有效利用将有可能大大改善自动文摘系统的性能。
- **综述自动生成：**综述自动生成是一类特殊的自动文摘任务，具有广泛的应用价值，可帮助自动撰写新闻事件深度报道、学术文献综述、舆情报告等。与传统自动文摘任务不同，综述一般较长，可以长达数千字，牵涉到篇章的整体逻辑性与局部连贯性，因此更具有挑战性。目前业界仅仅对学术文献自动综述进行了简单了尝试，效果差强人意，未来几年期待业界研究者在更多综述自动生成任务上进行有益的尝试，并在特殊应用场景下实现风格相对固定的综述文章自动撰写。
- **跨语言自动文摘：**为了方便进行多语言信息获取，跨语言自动文摘提供了一种有效和新颖的方式。具体而言，跨语言自动文摘的目的在于为源语言 A 中的文档以目标语言 B 的形式产生摘要，从而方便了解语言 B 的读者快速了解原文档信息。并不完善的机器翻译性能是跨语言自动文摘的最大障碍。目前业界仅对跨语言自动文摘进行了初步的尝试，后续预计将有更多的研究工作对该任务进行研究。另一种可能的方式为领域自适应或语言迁移，尝试将大规模英语摘要语料上的性质迁移到小规模的目标语言文摘语料，可以降低对于机器翻译的依赖，有机会在数据资源较少的语种下形成可读性更好的摘要。
- **多模态摘要：**新闻或社交媒体语料除了文本描述以外，往往也包含图像、音频、视频等多媒体数据。鉴于目前深度学习最大的成功和突破主要出现在语音处理、图像处理、计算机视觉等领域，有理由认为综合考虑多模态数据会对自动文摘技术有所帮助。最近已经有学者进行了一些初步探索，未来也有机会看到这一方向更多的研究成果不断出现。
- **面向复杂问题回答的自动摘要：**基于关键词检索的搜索引擎正在逐步向基于自然语言检索的问答引擎过渡。而对于很多类型的问题（例如为什么、怎么样等问题），并不适合使用简单的一个短语或一句话作答。相对完整地回答非事实型问题需要对单个文档甚至多个文档中的相关内容进行提取与聚合。由于非事实型问答固有的困难性，相关学术研究进展缓慢，期待未来有更多的研究者敢于迎接此项挑战。

除了上述研究方向与问题之外，未来自动文摘将会越来越多地与其他技术（例如情感分析、人机对话等）相结合，面向全新的应用需求，形成更具特色的自动文摘任务，该领域的研究也将更加多样化。

最后，我们有理由相信，随着语义分析、篇章理解、深度学习等技术的快速发展，自动文摘这一重要且有挑战性的自然语言处理问题在可预见的未来能够取得显著的研究进展，并且更多地应用于互联网产品与服务，从而体现自身的价值。

第十一章 信息检索研究进展、现状及趋势

1.任务定义、目标和研究意义

信息检索（Information Retrieval, IR）是指将信息按一定的方式加以组织，并通过信息查找满足用户的信息需求的过程和技术。1951 年，Calvin Mooers 首次提出了“信息检索”的概念，并给出了信息检索的主要任务：协助信息的潜在用户将信息需求转换为一张文献来源列表，而这些文献包含有对其有用的信息。

信息检索学科真正取得长足发展是在计算机诞生并得到广泛应用之后，文献数字化使得信息的大规模共享及保存成为现实，而检索就成为了信息管理与应用中必不可少的环节。互联网的出现和计算机硬件水平的提高使得人们存储和处理信息的能力得到巨大的提高，从而加速了信息检索研究的进步，并使其研究对象从图书资料和商用数据扩展到人们生活的方方面面。伴随着互联网及网络信息环境的迅速发展，以网络信息资源为主要组织对象的信息检索系统：搜索引擎应运而生，成为了信息化社会重要的基础设施。

互联网搜索引擎为人们提供了访问海量网络信息的高效便捷渠道，从而深刻的改变了人们的认知过程和信息获取方式。2011 年，Sparrow 等人在 Science 杂志上发表论文指出：包括搜索引擎和众多数据库在内的互联网信息环境，已经成为一种人们可以随时访问的外部记忆源，对人类原有的记忆过程产生了显著影响。2014 年，图灵奖获得者 Vinton Cerf 指出：搜索引擎已经成为人类记忆的代替与延伸，成为我们大脑对新事物认知过程中不可或缺的重要组成部分。2016 年初，中文搜索引擎用户数达到 5.66 亿人，这充分说明搜索引擎在应用层次取得的巨大成功，也使得信息检索，尤其是网络搜索技术的研究具有了重要的政治、经济和社会价值。

“知觉就是现实”，现代认知心理学普遍认为，对于环境的认知是人类知识产生和思维存在的基础，而环境的改变也必然对认知行为产生重要的影响。人类在网络搜索环境下是如何产生信息需求、如何在搜索引擎协助下获取信息资源、又是如何在这一过程中完善自身知识结构并满足其信息需求的？搜索引擎在这个过程中发挥了怎样的作用，我们又能够通过怎样的方式改进搜索引擎，使之在人类认知世界、学习知识、改善生活的过程中发挥更重要的作用？这些问题的回答，不仅有助于我们改进搜索引擎的算法和交互机制，对于我们深入掌握乃至进一步影响网络环境下人类的感知决策过程也是至关重要的。

2.研究内容和关键科学问题

检索用户、信息资源和检索系统三个主要环节组成了信息检索应用环境下知识获取与信息传递的完整结构，而当前影响信息获取效率的因素也主要体现在这几个环节，即：检索用户的意图表达、信息资源（尤其是网络信息资源）的质量度量、需求与资源的合理匹配。具体而言，用户有限的认知能力导致其知识结构相对大数据时代的信息环境而言往往存在缺陷，进而影响信息需求的合理组织和清晰表述；数据资源的规模繁杂而缺乏管理，在互联网“注意力经济”盛行的环境下，不可避免的存在欺诈作弊行为，导致检索系统难以准确感知其质量；用户与资源提供者的知识结构与背景不同，对于相同或者相似事物的描述往往存在较大差异，使得检索系统传统的内容匹配技术难以很好应对，无法准确度量资源与需求的匹配程度。上述技术挑战互相交织，本质上反映了用户个体有限的认知能力与包含近乎无限信息的数据资源空间之间的不匹配问题。

概括地讲，当前信息检索的研究包括如下四个方面的研究内容及相应的关键科学问题：

2.1 信息需求理解

面对复杂的泛在网络空间，用户有可能无法准确表达搜索意图；即使能够准确表达，搜索引擎也可能难以正确理解；即使能够正确理解，也难以与恰当的网络资源进行匹配。这使得信息需求理解成为了影响检索性能提高的制约因素，也构成了检索技术发展面临的第一个关键问题。

2.2 资源质量度量

资源质量管理与度量在传统信息检索研究中并非处于首要的位置，但随着互联网信息资源逐渐成为检索系统的主要查找对象，网络资源特有的缺乏编审过程、内容重复度高、质量参差不齐等问题成为了影响检索质量的重要因素。目前，搜索引擎仍旧面临着如何进行有效的资源质量度量的挑战，这构成了当前信息检索技术发展面临的第二个关键问题。

2.3 结果匹配排序

近年来，随着网络技术的进步，信息检索系统（尤其是搜索引擎）涉及的数据对象相应的变得多样化、异质化，这也造成了传统的以文本内容匹配为主要手段的结果排序方法面临着巨大的挑战。高度动态繁杂的泛在网络内容使得文本相似度计算方法无法适用；整合复杂异构网络资源作为结果使得基于同质性假设构建的用户行为模型难以应对；多模态的交互方式则使得传统的基于单一维度的结果分布规律的用户行为假设大量失效。因此，在大数据时代信息进一步多样化、异质化的背景下，迫切需要构建适应现代信息资源环境的检索结果匹配排序方法，这是当前信息检索技术发展面临的第三个关键问题。

2.4 信息检索评价

信息检索评价是信息检索和信息获取领域研究的核心问题之一。信息检索和信息获取系统核心的目标是帮助用户获取到满足他们需求的信息，而评价系统的作用是帮助和监督研究开发人员向这一核心目标前进，以逐步开发出更好的系统，进而缩小系统反馈和用户需求之间的差距，提高用户满意度。因此，如何设计合理的评价框架、评价手段、评价指标，是当前信息检索技术发展面临的第四个关键问题。

3. 研究进展和现状

3.1 信息需求理解

用户信息需求理解是用户与搜索引擎交互过程的核心，针对搜索意图的研究一方面能够促进对搜索环境中用户的普遍认知行为规律的深入认识，有重要的学术意义；另一方面，正确理解用户的信息需求，有助于搜索引擎返回更加满足用户信息需求的结果，提高用户查找信息的效率。常见的分析方法包括：

3.1.1 基于用户行为的分析方法

由于用户的信息需求会影响用户提交查询、浏览结果页面、点击相关结果等行为，通过分析用户行为记录，我们将能够有效的检测到一些用户信息需求。例如利用信息需求和点击记录之间的关系——导航类查询倾向于只伴随一次点击，而信息类查询往往伴随多次点击。再比如，用户提交的查询和提交查询后点击的 URL 会构成一个“查询-点击二部图”。基于该二部图，可以计算一对查询相互之间的相似程度。近年来，眼动追踪技术（Eye tracking）被广泛应用于研究和分析用户与搜索引擎交互过程。Guo 等人研究表明用户浏览搜索引擎结果页面时的注视位置与用户信息需求相关。

3.1.2 基于伪相关反馈信息的分析方法

利用包括查询日志、点击日志、眼动信息和鼠标移动信息在内的用户行为信息能够有效的推测查询背后的用户信息需求。但有些时候，尤其是针对查询频度较低的长尾查询，我们无法获得足够多的用户行为记录，来有效地进行搜索意图分析。所以，研究者们也尝试基于查询和查询对应的伪相关反馈信息（如搜索引擎结果页的内容），进行搜索意图分析。

3.1.3 基于自然语言理解的分析方法

随着输入手段的不断出现以及语音输入准确性的不断提升，用户在搜索引擎中所输入的检索项越来越长，如何直接通过对用户输入的检索项进行分析，从而得到用户意图也是近年来的研究热点。研究者们针对特定领域检索，提出了结构化表示方法，并利用自然语言处理方法对用户检索项进行语义分析，从而对用户搜索意图进行分析。

3.1.4 垂直需求理解分析方法

现代搜索引擎往往不再仅仅返回与查询相匹配的网页作为查询结果，而是根据用户提交的查询，返回包括新闻、图片、视频、本地搜索、购物信息等垂直结果在内的异质化结果页面。与垂直需求理解分析关系最为密切的是垂直搜索资源选择问题。 大多数工作将垂直搜索资源选择问题当作一个有监督分类问题处理。利用查询字符串、垂直搜索引擎的查询日志、垂直搜索引擎、 用户的反馈等信息构建分类模型。

3.2 数据质量评估

数据质量评估问题的核心，是清除索引中的冗余、低质量、不可信和过时数据，而保证真正满足用户需求的数据能够得到检索系统排序算法的关注。具体的施行方案中，较多被考虑的则是链接结构分析方法与垃圾网页识别技术。

3.2.1 基于链接结构的质量评估

当前网页质量评估的主要工作集中在链接关系分析方面，相关工作大都集中在利用 PageRank 框架进行某些特定应用需求的改进，或对标准 PageRank 传统算法进行效率提升上。

尽管以 PageRank 为代表的链接结构分析算法在搜索引擎排序系统、索引系统等的设计中取得了很好的应用，但搜索引擎对于链接结构数据的依赖也客观上造成了此类数据本身质量堪忧的现象。针对这一问题，以谷歌公司 Henzinger 为代表的部分研究人员指出，应当

采用多种特征共同评价网页质量，设计更加全面合理的质量评估算法。此外，微软研究院的 Liu 等人采用了搜索引擎通过浏览器插件等收集的用户浏览行为数据建立用户浏览关系图替代网络结构图实施链接结构分析，以期取得更优的质量评估效果。

3.2.2 垃圾网页识别

垃圾网页是利用搜索引擎运行算法的缺陷，采取针对搜索引擎的作弊手段，使其获得高于其网络信息质量排名效果的网页。垃圾网页的作弊方式主要可以分为基于内容的作弊（Content Spamming）与基于链接关系的作弊（Link Spamming）两种类型，这是从影响搜索引擎检索结果排序的两个不同角度对作弊手段进行的分类。传统的垃圾网页识别方法，大都是针对特定的作弊手段设计有针对性的识别算法予以应对，如采用内容压缩比、可见内容比例等特征识别关键词堆砌类垃圾网页，采用脚本解析应对自动跳转类垃圾网页等。这些方法尽管对特定类型的垃圾网页具有很好的识别效果，但是缺乏对新出现垃圾网页的应对能力，也缺乏识别通用性。

近年来，人们在具有通用识别能力的垃圾网页识别方法方面开展了不少研究工作。Gyöngyi 等研究人员试图采用链接结构分析方法避免对垃圾网页作弊手段本身的关注，代表性算法包括 TrustRank 及其延伸算法 Anti-TrustRank、 GoodBadRank 等。为了避免链接结构分析算法本身面临的链接结构数据质量问题，Liu 等人利用用户与垃圾和正常网页的交互模式差异，从作弊目的而非手段的角度来识别垃圾网页。

3.3 检索结果排序

信息检索系统目前的主要交互方式，是依据用户提交的查询，按照内容相似程度、质量水平、用户偏好情况、竞价情况、时效性情况等因素将结果文档进行排序，并以有序列表的形式反馈给用户。在检索结果排序技术研究中，当前的主要研究集中在对传统内容检索模型的改进与完善、以及如何采用机器学习方法整合包括用户偏好、多样化需求等因素提升检索效果方面。

3.3.1 信息检索模型

信息检索模型指如何对查询和文档进行表示并进行相似度计算的框架和方法。信息检索模型是信息检索系统的核心内容之一。当前信息检索模型的分类如图所示，其中经常使用的是布尔模型、向量空间模型、语言模型和学习排序模型。

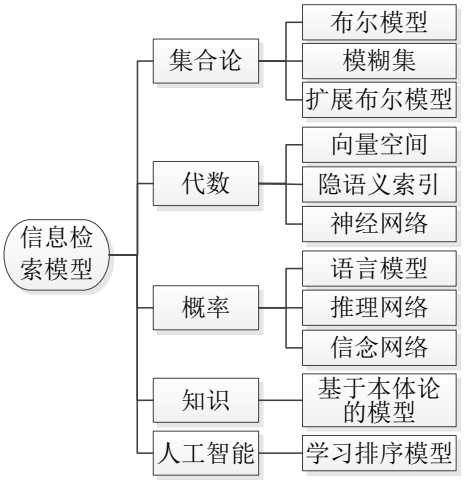


图 1. 信息检索排序模型

3.3.2 排序学习

实际搜索引擎中需要考虑的排序因素已经成百上千，单靠人工将它们整合到一个排序公式中已经不太现实。研究人员开始尝试使用排序学习方法，即从用户标注或者搜索日志数据中利用机器学习的方法训练排序模型（如下图所示）。与传统排序模型相比，排序学习的优势在于对大量的排序特征进行组合优化，自动进行参数的学习，最终得到一个高效精准的排序模型。

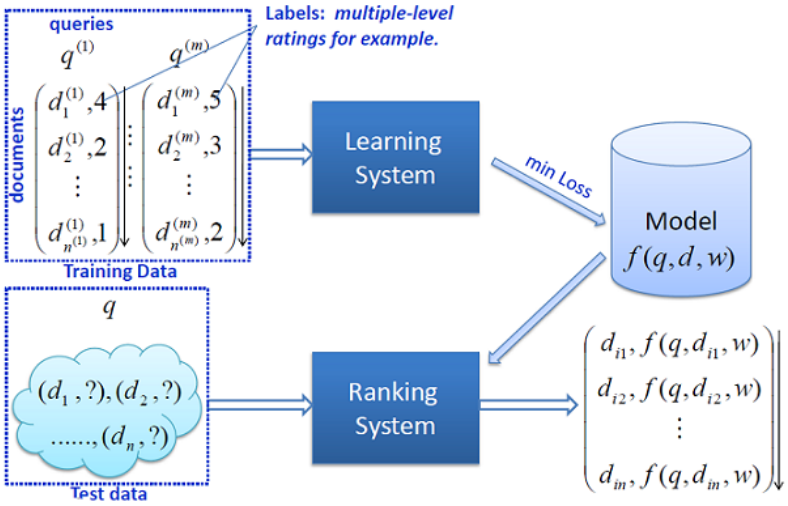


图 2. 排序学习过程示意（来自 Tie-Yan Liu 在 WWW2008 的 Tutorial）

在过去的十几年里，排序学习研究在学术界和工业界都获得积极的发展和推广，并取得了巨大的成功。在算法研究方面，针对不同的排序场景和模型假设研究人员提出了多种不同的排序学习算法，例如早期大多数研究人员将排序看做回归或者分类问题所提出的单点型排序学习算法(Pointwise 算法)，后来占据主导地位的点对型排序学习算法(Pairwise 算法)，其思想是将排序问题看做是同一查询下两个文档间的相对相关性关系建模，以及在排序学习高潮期的列表型排序学习算法（Listwise 算法），该方式直接建模一系列文档间的序列型关系，避免了前两类方法的近似误差；在产品实际应用方面，排序学习已经成为各个互联网搜索产品网页排序的核心技术，目前雅虎、必应、百度和搜狗等商业互联网搜索巨头均采用排序学习对网页进行排序。

3.3.3 多样化搜索

多样化搜索成为检索结果排序一个重要的研究方向源于实际应用中三个方面的原因：1. 网络数据存在大量的冗余信息；2. 很多用户查询具有歧义；3. 对同一个查询不同的用户有不同方面的信息需求。传统的仅仅考虑相关的排序显然无法解决上述三个问题。多样化搜索在相关性的基础上，进一步考虑结果之间的差异性（或者说结果的新颖性），从而达到去除冗余、覆盖不同信息需求的目的。

早期解决多样化搜索的方法是启发式的排序模型，主要包含隐式和显式两类方法。隐式的方法主要假设相似的文档覆盖的话题或者满足的信息需求相似，通过定义文档间的依赖关系来捕捉多样性；而显式的方法则是通过显式地定义或者挖掘查询的各个子话题，从而直接选择能够覆盖这些子话题的文档作为排序结果。近年来，越来越多的工作通过机器学习的方法进行结果的多样性排序。为了建模多样性，排序学习模型需要考虑文档间的关系，因此大多属于序列级（listwise）排序方法（即优化目标定义在整个序列上）。代表性工作包括结构化排序方法、关系排序学习以及在线学习等。优化目标从极大似然的目标发展为直接优化多样性评价指标。近期，深度学习的方法也被引入到多样性排序工作中来，以便解决传统机器学习方法中多样性特征难以定义的难题。

3.3.4 个性化搜索

现有的主要个性化搜索算法可分为基于内容分析的算法、基于链接分析的方法和基于协作过滤的算法。

基于内容的个性化搜索算法通过比较用户兴趣爱好和结果文档的内容相似性来对文档的用户相关性进行判断进而对搜索结果进行重排。用户模型一般表述为关键词或主题向量或层次的形式。个性化算法通过比较用户模型和文档的相似性，判断真实的搜索意图，并估计文档对用户需求的匹配程度。

基于链接分析的方法主要是利用互联网上网页之间的链接关系，并假设用户点击和访问过的网页为用户感兴趣的网页，通过链接分析算法进行迭代最终计算出用户对每个网页的喜好度。

基于协作过滤的个性化搜索算法主要借鉴了基于协作过滤的推荐系统的思想，这种方法考虑到能够收集到的用户的个人信息有限，因此它不仅仅利用用户个人的信息，还利用与用户相似的其它用户或群组的信息，并基于用户群组和相似用户的兴趣偏好来个性化当前用户的搜索结果。用户之间的相似性可以通过用户的兴趣爱好、历史查询、点击过的网页等内容计算得出。

3.3.5 排序点击模型

搜索引擎用户在与搜索引擎的交互过程中反映出的隐性反馈信息(主要是点击行为信息)是搜索引擎用来改进结果排序的重要影响因素。然而，由于结果位置、展现形式等各种因素的影响，将反馈信息直接应用于搜索排序任务往往难以取得较好的效果。针对这一问题，研究人员提出了构建描述用户点击行为的点击模型(Click Model)，并基于不同的点击模型估计用户对展现结果的浏览概率，进而尝试去除结果展现位置等因素对用户行为的偏置性影响，以达到更好利用隐性反馈信息的目的。作为一种用户交互信息的有效利用方法，点击模型在学术界得到了充分关注，并在工业界得到了广泛的应用。

传统的点击模型主要针对于传统同质化的搜索页面(搜索页面中的结果均采用相近的文本形式展现，结果之间除了文字内容不同外并没有明显的展现形式差异）进行设计。这方面的典型工作包括级联模型(Cascade model)，依赖点击模型(Dependent Click Model，DCM)，用户浏览模型(User Browsing Model，UBM)，动态贝叶斯网络模型(Dynamic Bayesian Network，DBN）等。上述模型主要依照结果位置因素提出用户检验行为假设，并进而对点击行为进行推演。

随着 Web2.0 时代的到来，富媒体展现形式被越来越多的应用于搜索交互界面，搜索结果也变得越来越异质化。因此研究人员开始提出针对于垂直搜索结果的点击模型以及针对非顺序检验行为的点击模型。针对垂直搜索结果方面，典型的点击模型包括 Chen 等人提出的联合点击模型（Federated Click Model，FCM）模型与 Wang 等人提出的垂直点击模型（Vertical Click Model，VCM）等。

图 3. 多媒体垂直结果对用户的前 2 秒视觉注视行为的影响（左侧为不含垂直结果的页面，右侧为包含多媒体垂直结果的页面）

在非顺序检验行为的建模方面，主要的工作包括 Xu 等人提出的用于描述广告搜索用户行为的时间点击模型（Temporal Click Model，TCM），Wang 等人提出的局部可观测马尔科夫模型（Partially Observable Markov Model，POM）以及 Wang 等人基于真实用户眼动行为实验提出的局部有序点击模型（Partially Sequential Click Model，PSCM）。

3.4 检索性能评价

对搜索系统的评价通常可以分为对检索结果质量（Effectiveness）、检索效率（Efficiency）、用户界面友好度及易用性（Interface）的评价。其中检索效率主要考虑检索的时间开销、空间开销和响应速度等。检索结果质量重点考虑检索结果是否满足用户的信息需求，如考虑返回的文档中有多少相关文档、所有相关文档中返回了多少、返回的是否靠前等，是评测的重点。

3.4.1 离线性能评价

检索系统的离线性能评价主要采用基于 Cranfield 范式的方法。它的主要特点是使用一套可重用的评测集来评价信息检索系统的好坏。整个 Cranfield 评测集通常包括一个文档集合，一个信息需求的集合，以及和信息需求集合匹配的标注集。不同的搜索系统通过在相同的文档集合需求集上生成结果并通过统一的标注集进行评测对比来比较彼此之间的优劣。常见的评测集合包括美国标准与技术研究院组织的 TREC 会议构建的评测集、日本国立情报学研究所组织的 NTCIR 会议构建的评测集、欧盟 CLEF 组织构建的跨语言检索评测集等。基于 Cranfield 的评测集合往往是可重用的，一旦被构建，就可以被用来评测新的搜索系统。不同的评测指标通常具有不同的表达能力和适用范围。由于评测集是静态的，并且对每个搜索系统都能通过评分指标获得一个绝对的分数，因此不同的搜索系统的优劣就可以通过分数数值来反映。因此可以用这种方法评测后续改进的评测算法。

3.4.2 在线性能评价

与离线性能评价不同，在线性能评价并不需要专业人员进行针对文档相关性的标注，而是依照用户在使用检索系统时的显式（Explicit）或隐式（Implicit）反馈信息对检索系统的性能进行评价。典型的用户显式反馈的信息包括满意度评价（Satisifaciton）、用户偏好（Preference）、信息需求完成情况（Search Outcome）等，而隐式反馈信息则往往包括用户点击（Click-through）、查询重组（Query reformulation）、停留时间（Dwell time）等交互行为信息。

除了上述采用用户隐式反馈信息与行为假设来解读在线性能评价结果的方法之外，也有

部分研究人员尝试利用机器学习方法对各类隐式反馈信息进行综合，并对满意度评价、用户偏好等显式反馈信息进行预测。通常使用的隐式反馈信息包括用户与搜索引擎交互过程中的各种粗粒度（Coarse grain，如查询修改、结果点击等）或细粒度（Fine grain，如鼠标滚轮行为，结果页面停留时间、鼠标移动行为模式等）信息。

4. 总结与展望

纵观信息检索研究发展的态势和技术现状，以下研究方向或问题将可能成为未来信息检索研究的潜在重要研究方向：

4.1 交互式搜索技术

随着移动计算设备的迅速普及和移动网络接入资费的下调，随时随地利用最自然便捷的方式获取信息不再是梦想。2015 年底，谷歌公司在英国计算机学会召开的 Search Solutions 会议上展望，能够支持人机自由对话的搜索引擎系统在 20 年之内将成为现实。近年来，包括 Apple 公司 Siri，微软公司 Cortana，谷歌公司 Google Now 等在内的深度整合交互式搜索功能的移动互联网新产品都是这一发展趋势的见证。

在近年来由深度学习、增强学习等技术推进的新一轮人工智能技术研究热潮中，交互式搜索及其代表性应用人机对话系统由于与图灵测试的密切关联受到学术界与工业界共同的密切关注。然而实现人类自由交谈、解决人类面临的日常生活中的各类问题的机器人，仍然面临极大的技术挑战，包括用户理解与建模、搜索资源整合和自然语言交互能力等方面。除此之外，缺乏对于自然界与人类社会中各类常识性知识的积累与理解，作为各类人工智能应用面临的普遍性问题，也是交互式搜索系统面临的重要技术挑战。

4.2 搜索意图理解技术

搜索引擎涉及的数据对象已经扩展到包括虚拟空间、物理世界、人类社会在内的泛在网络空间中。一方面，搜索引擎索引的内容因此逐渐多模态化，另一方面，随着用户的增长以及智能手机和平板等智能设备的普及，搜索引擎的使用场景更为多样化。

以上两方面的变化，使得用户的搜索意图相应的变得多样化、异质化，为传统的以用户群体行为分析为主要手段的意图分析方法带来巨大的挑战。搜索意图分析已经成为当前各类搜索引擎技术发展的核心环节之一。

4.3 语义搜索技术

随着互联网信息的爆炸式增长，传统的以关键字匹配为基础的搜索引擎，已越来越难以满足用户快速查找信息的需求。同时由于没有知识引导及对网页内容的深入整理，传统网页搜索返回的网页结果也不能精准给出所需信息。针对这些问题，以知识图谱为代表的语义搜索（Semantic Search）将语义 Web 技术和传统的搜索引擎技术结合，是一个很有研究价值但还处于初期阶段的课题。

在未来的一段时间，结合互联网应用需求的实际和技术、产品运营能力的实际发展水平，我们认为语义搜索技术的发展重点将有可能集中在以各种情境的垂直搜索资源为基础，知识化推理为检索运行方式，自然语言多媒体交互为手段的智能化搜索与推荐技术。首先将包括各类垂直搜索资源在内的深度万维网数据源整合成为提供搜索服务的资源池；随后利用广泛分布在公众终端计算设备上的浏览器作为客户端载体，通过构建的复杂情境知识库来开发多层次查询技术，并以此管理、调度、整合搜索云端的搜索服务资源，满足用户的多样化、多

模态查询需求；最后基于面向情境体验的用户行为模型构建，以多模态信息推荐的形式实现对用户信息需求的主动满足。

第十二章 信息推荐与过滤研究进展、现状及趋势

1.任务定义、目标和研究意义

信息推荐与过滤(Information Recommendation and Filtering)简称信息推荐，是指根据用户的习惯、偏好或兴趣，从不断到来的大规模信息中识别满足用户兴趣的信息的过程。信息推荐任务中的信息往往称为物品(Item)。根据具体应用背景的不同，这些物品可以是新闻、电影、音乐、广告、商品等各种对象。

简而言之，信息推荐研究的目标就是建立用户兴趣和物品之间的有效匹配算法、模型和系统，最终实现用户感兴趣物品的推荐，从而缓解用户在面对大量物品时的信息过载问题，提高物品信息的利用率。

从上面的定义可以看出，信息推荐和传统的信息检索(Information Retrieval)任务十分类似，后者也是从大量信息中返回满足用户需求信息的过程。实际上，广义的信息检索包括两类子任务，一类称为即兴搜索(Ad hoc Search)，另一类称为过滤(Filtering)。前者对应传统的检索任务，后者即对应信息推荐。两者都是从大量信息中返回满足用户所需信息的过程。但是，即兴搜索任务的用户需求相对动态，而信息库相对静态。搜索引擎就是一个典型的即兴搜索任务，用户不断输入代表不同需求的查询，系统从后台相对静态的信息库中返回匹配的信息。与即兴搜索任务不同，信息推荐任务的用户需求是用户的兴趣，该需求在一段时间内相对静态，而其面对的信息却在不断动态变化。网络新闻订阅就是一个典型的信息推荐任务，用户每天会获得与其兴趣相关的信息。正因为即兴搜索任务和信息推荐任务之间的相似性，因此即兴搜索中的很多方法和技术都可以应用于信息推荐任务。但是两者在具体任务上又有所不同。也因为这一点，有些学者干脆认为信息推荐是与传统信息检索完全不同的一个新的学科领域。本文不刻意纠结信息推荐到底是否属于信息检索领域，相信读者能够在各自上下文中理解当前任务的具体含义。

此外，除了上面提到的从学术角度提到的“过滤”含义，过滤通常还有应用层面的另一种含义，比如垃圾邮件过滤、不良网页过滤等应用中提到的“过滤”。这个过滤并不强调用户兴趣的静态性和信息的动态性，而往往强调的是任务的结果，即去掉不需要的信息。本文主要介绍学术意义上的信息“过滤”。

众所周知，人类社会进入了大数据时代，数据量极度膨胀，人们面临严重的信息过载(Information Overload)问题，从大量信息中获得满足用户需求的信息成为从现在到未来的永恒需求。信息推荐技术是实现这一需求的重要手段，具有重要的商业价值。电子商务网站是运用信息推荐技术的最典型应用。不论是亚马逊、京东还是淘宝网站，都大量运用信息推荐技术。这些网站根据用户的购买或浏览历史，向用户推荐用户最有可能购买的商品，从而在满足用户购买意图的同时大大提高商品的销售率。有数据表明，亚马逊推荐系统的上线将公司的利润提高了至少 30%。除了电子商务之外，信息推荐技术还在包括视频、广告、新闻、交友等在内的许多其他互联网应用中也起着非常重要的作用。例如在视频领域，YouTube、优酷等公司都推出了自己的视频推荐功能，强大的个性化推荐功能给用户带来了很大的方便，使得用户可以不经搜索就能自动观看到自己想看的视频，这大大提高了用户的体验，从而给信息提供商带来巨大的流量。此外，信息推荐技术对于政府、企业的信息化建设也具有重要的价值。以舆情分析为例，政府、企业关注某些特定的话题，这些话题可以看成某种稳定的用户兴趣，通过信息推荐技术可以从互联网中不断筛选出满足这些兴趣的内容，从而为政府、企业的决策提供有力的支撑。

由于信息推荐的重要应用价值，它一直受到学术界和工业界的广泛关注。美国计算机协会(ACM)也从 2007 年开始组织召开一年一度的推荐系统学习会议(简称 RecSys)，同时学术界和工业界也推出了许多信息推荐竞赛，这些活动大大推进了信息推荐技术的研究。但是，信息推荐技术的研究仍然面临很多问题和困难。本文对信息推荐研究的主要内容、面临的科

学问题和主要困难以及当前采用的主要技术、研究现状和未来发展的趋势作简要介绍。

2.研究内容和关键科学问题

一般而言，信息推荐包含用户兴趣的建模、物品的建模和用户兴趣-物品的匹配三个基本步骤。作为匹配的两方，用户兴趣和物品在计算机中通常都要进行形式化建模，即转化成计算机的某种内部表示。在此基础上，计算两者的匹配程度，匹配程度高的物品推荐给用户。上述三个步骤往往并非一遍就结束，系统会根据用户的反馈对用户的兴趣模型进行调整更新。从另外一个角度来看，信息推荐就是已知如下用户-物品评分矩阵中的部分值(用非 0 值表示，代表用户对物品的喜好程度)求其它值(矩阵中用 0 表示未知值)的过程。

表 1 用户-物品评分矩阵

	物品 1	物品 2	物品 3	物品 4	物品…	物品 n
用户 1	1	0	0	0	…	2
用户 2	0	4	0	0	…	5
用户 3	0	0	0	3	…	0
用户…	…	…	…	…	…	…
用户 m	0	0	2	0	0	0

总的来说，信息推荐研究主要面向如下三个关键科学问题：

2.1 用户兴趣的建模

信息推荐的目的就是向用户推荐其感兴趣的物品。因此，如何获取并刻画用户的兴趣是信息推荐所面临的首要科学问题。按照用户是否直接提供兴趣数据，用户兴趣的建模可以划分为显式方法和隐式方法。显式方法中，用户提供了自己的兴趣数据(比如填写兴趣表格或者对物品打分)，建模过程只需要从中构建相应的特征表示即可。隐式方法中，用户并没有明显表示出自己的兴趣，此时，可以根据用户的浏览、点击、收藏等行为数据来预测用户的兴趣。有证据表明，出于惰性或者隐私保护需要，用户并不愿意显式地提供兴趣数据，因此，隐式建模是目前更主流的用户兴趣建模方法。从用户兴趣建模的结果来分，可以分成基于浅层语义的方法和基于语义概念或知识的方法。基于浅层语义方法的典型代表是关键词表达式，它通过关键词组合来表示用户的兴趣。另一种基于浅层语义的方法是向量表示法，通过将用户的兴趣表示成向量，系统可以通过向量空间中的计算方法来进行推荐。基于语义概念或知识的方法将用户兴趣表示成某种语义网络、或者统计推理规则、或者关键词之间的关联关系来进行推荐。隐式建模中用户的行为数据往往带有噪音，如何从中选择高质量的数据是用户兴趣建模所需解决的问题之一。而用户的兴趣往往又十分广泛，因此建模时要考虑用户的兴趣多样性问题。用户的兴趣还会随着时间的推移发生变化，建模时也要及时对用户的兴趣进行更新。

2.2 物品的建模

信息推荐中的另一个重要对象就是物品。因此，如何刻画物品是信息推荐过程中的另一个重要问题。物品建模的目的是构建物品的某种形式化表示。最常见的一种做法就是将物品表示为其重要特征或属性表示的向量。这里涉及到重要特征的选择问题和特征的表示问题。进行特征选择时，可以借用传统文本分类中的特征选择方法。在进行特征表示时，则可以借用传统文本检索中的 TF-IDF 表示方法，即从属性在物品内的出现次数(比如文本中的某个词

语或图像中的某种颜色)和出现属性的物品数目(比如文本中出现某个词语或图像库中出现某种颜色的对象数目)两个方面来综合考虑属性的权重，前者刻画了属性的代表性，即该属性在物品内部出现越多，则意味着该属性的权重越大；后者刻画了属性的区别性，即该属性在所有物品中出现越多，则意味着其区别性不大，此时反而要降低权重。需要指出的是，物品和用户可以采用不同的表示方法，只要满足用户-物品相似度计算的输入要求即可。即使二者均采用向量表示方法，也可以处于完全不同的向量空间。物品的建模主要需要考虑物品关键特征的提取，要面向用户可能的兴趣，来抽取相应的本质特征并进行表示。

2.3 用户兴趣-物品的匹配度计算

要实现满足用户兴趣的物品推荐，最关键的一步是计算用户兴趣和物品之间的匹配度，匹配度越大，推荐的可能性也越大。匹配度计算主要考虑用户的满意度。用户兴趣-物品的匹配方法大体上可分为两种：基于统计的方法和基于规则的方法。基于统计的方法中，用户兴趣和物品往往表示成某种概率统计量(如向量或者属性的某种概率统计值)，在进行匹配时，可以基于这些概率统计量来计算两者的匹配度。此时可以采用传统中的文本检索模型(如向量空间模型、概率模型、统计语言模型)或者分类模型(如 k-近邻、朴素贝叶斯、支持向量机)来进行计算。为了克服原始向量空间匹配度计算的不足，一些方法将用户兴趣和物品通过矩阵分解或其他方法映射到某个隐性空间，然后再进行匹配度计算。基于规则的方法中，用户的兴趣往往表示成类似于“IF…THEN…ELSE…”之类的规则表达式，用户兴趣-物品的匹配就是规则匹配的过程。用户兴趣和匹配度的计算，要同时考虑效果和效率问题。在效果上，尽量推荐用户感兴趣的结果。用户感兴趣不一定是匹配度最大，有时候要考虑物品的新鲜度、多样性等因素。在效率上，如何在极大规模的数据条件下进行快速的推荐是一个十分重要的研究问题。

2.4 信息推荐的研究难点

信息推荐的研究面临诸多挑战，主要包括如下几个方面：

(1) 数据稀疏性(Data Sparsity)问题。简而言之，信息推荐可以认为是根据已有用户对物品的喜好情况(如通过评级或评分来表示)来预测未知的用户-物品喜好情况。但是，在大规模的推荐系统当中，用户数目和物品数目都非常大。而实际上用户表示过喜好的物品数目极小。也就是说，已知的喜好数据存在极大的稀疏性(表 1 所示的矩阵中的绝大部分元素都是 0)。数据稀疏性会带来计算过程中的偏差，比如在进行推荐时往往需要计算用户或物品间的相似度，但是数据稀疏时算出的用户或物品间的相似度可能很不准确。

(2) 冷启动(Cold Start)问题。在推荐系统中，冷启动问题是指新用户或者新物品面对的“推荐困难”问题。新用户由于没有或极少对物品进行过评分，所以很难分析得到他的喜好，从而无法对他进行有效的物品推荐。同样，新物品加入系统时，也由于还没有或只有极少数用户对其表示喜好程度，因此也无法将新物品推荐给用户。

(3) 大规模计算问题。在大型推荐系统中，用户和物品数目都十分巨大，而且还在不断增长。在如此巨大的规模条件下为大量的在线用户提供个性化快速推荐，是一个很大的挑战。此外，推荐系统的推荐精度和实时性有时是一对矛盾，大部分推荐技术为了保证实时性，是以牺牲推荐系统的推荐质量为代价的。在提供实时推荐服务的同时，如何有效提高推荐的推荐质量，有待进一步的研究。

除了上述几个主要的难点之外，信息推荐还面临很多挑战，比如：推荐系统的评价、推荐结果的可解释、推荐系统的多目标、推荐系统的被攻击等等问题。由于篇幅有限，这里不再一一介绍。

3.技术方法和研究现状

信息推荐主要包含两类方法：基于内容过滤(Content-based Filtering)的推荐方法和基于协同过滤(Collaborative Filtering)的推荐方法。基于内容过滤的方法往往也称为基于感知过滤(Cognitive Filtering)的方法，可以认为它是一种“直接”计算用户兴趣-物品的方法。这类方法通过直接计算用户兴趣和待推荐物品的匹配度进行推荐。基于协同过滤的方法也称为基于社会过滤(Sociological Filtering)的方法，可以认为它是一种“间接”计算用户兴趣和待推荐物品匹配度的方法。这类方法在计算用户兴趣-物品的匹配度时，往往通过计算其他用户兴趣和当前物品的匹配度或者当前用户兴趣和其他物品的匹配度来估计当前用户兴趣-当前物品的匹配度。

3.1 基于内容过滤的信息推荐

基于内容过滤的信息推荐是一种重要的信息推荐方法,其基本思想是给用户推荐与他们喜欢的物品在内容上比较相似的物品。例如，用户喜欢《机器学习》这本书，那么基于内容过滤的信息推荐系统可能会给他推荐《机器学习实战》、《机器学习导论》等书籍。这是因为用户的兴趣可以通过他喜欢的书籍表示出来，与该书籍相似的书籍会被推荐。因此，基于内容的过滤的信息推荐系统最主要的任务就是计算物品之间的相似度。对于计算物品相似度，最著名的模型是在信息检索领域经常使用的向量空间模型。前面也已经提到，该模型主要是对物品内容中关键特征进行抽取，接着利用诸如 TF-IDF 的权重模型计算这些关键特征的权重，然后通过夹角余弦、内积等方式计算物品之间的相似度，最后向用户推荐其没有表态是否喜欢的物品集合。计算物品相似度的关键环节与使用关键词特征对物品建模，这也是是很多基于内容过滤的推荐系统的重要部分。但是如何准确抽取出物品的关键特征是一个非常困难的问题，特别在视频、音频、图像等领域，很难根据内容信息抽取出关键特征。一种可能的做法是利用社会化标签来对待匹配的物品进行文本描述,但是大部分推荐物品都不存在对应的社会化标注,并且社会化标注中也存在大量噪音,因此,该方法仍有进一步提高的空间。

基于内容过滤的信息方法虽然经过机器学习、信息检索、自然语言处理等相关技术的发展，推荐效果得到了很大的改善。但是由于基于内容的过滤受到特征抽取方法的限制，不能充分考虑用户的个性化等缺点，因此推荐效果不能令人满意。基于内容过滤推荐方法的一个明显不足是推荐的同质化问题。用户的兴趣多种多样，其显式或隐式表示出的兴趣是十分有限的，而基于内容过滤的推荐方法只推荐与当前兴趣表示匹配的物品，因此，推荐的结果具有同质性，用户还没表示出来但是实际上感兴趣的物品难以得到推荐。

3.2 基于协同过滤的信息推荐

基于协同过滤的信息推荐是当前推荐研究领域最广泛的算法。该算法的基本思想十分直观，即 “物以类聚，人以群分”，也就是说，喜欢相似物品的用户兴趣也相似，或具有相似兴趣的用户喜欢的物品也相似。在向某用户推荐物品时，可以先找到与该用户兴趣相似的若干用户，然后基于这些用户的喜好来推荐物品。另一种推荐方法是先找到与物品相似的其它物品，然后根据当前用户对其它物品的喜好程度来判断其对当前物品的喜好程度。前一种推荐方法称为基于用户(User based)的协同过滤推荐，后者称为基于物品(Item based)的协同过滤推荐。由于上述两种方法将用户-物品的喜好信息存储在系统中并在推荐中直接使用，所以两者统称为基于内存或记忆(Memory based)的协同过滤方法。

基于内存的协同过滤方法是最早出现的协同过滤算法。GroupLens 早在 1994 年就发表了第一篇基于协同过滤的论文，论文的核心算法是基于用户的协同过滤算法。此后基于物品的协同过滤算法吸引了很多研究者的关注。2003 年，Amazon 介绍了其基于物品的协同过滤算法。这两种算法也是目前商用推荐系统中使用最多的算法。由于在一个推荐系统中的注册

用户一般远远大于网站中的物品数，因此，在计算物品相似度的时候，共同的用户比较多，所以基于物品协同过滤的推荐效果一般比基于用户的协同过滤的推荐效果要好。

基于物品的协同过滤方法的另外一个好处是，容易给推荐结果提供合理的解释，而合理的解释在推荐系统中有着非常重要的作用。比如，网站给用户推荐了一本名为《Introduction to Algorithms》的书，推荐理由主要是因为该用户曾经购买过《Pattern Classification》和《The Elements of Statistical Learning》，并且用户还喜欢《Probabilistic Graphical Models》这本书，这些书和《Introduction to Algorithms》这本书比较相似。通过对推荐结果的解释，用户更能相信推荐结果的合理性，从而能够提高用户购买物品的机率。

基于协同过滤的另外一种推荐算法是基于模型(Model based)的推荐算法。基于模型的推荐方法主要是通过设计机器学习、数据挖掘等模型使得系统能够学习在训练数据集中的复杂模式，然后基于学习到的模型对测试集合或者现实世界中的数据进行预测。基于模型的方法包括贝叶斯模型、聚类模型、隐语义模型等等。还有很多研究者提出利用数学上的矩阵分解(Matrix Factorization)模型来进行信息推荐，这些矩阵分解方法包括 SVD、RSVD、AsySVD、SVD++、NMF 等等。其基本思想就是，对表 1 所示的用户-物品矩阵进行分解，让该矩阵分解成多个小矩阵的乘积。该分解过程往往会被定义成一个优化问题来求解，通常一方面要保证分解后的多个小矩阵相乘之后尽量和原有矩阵中的已知值接近，另一方面，又要尽量使得分解的结果不至于过拟合。这种矩阵分解的框架有一个优点，可以将数据集的某种约束(比如一些物品属于同类)直接形式化加入到优化目标函数中进行求解。除此之外，基于模型的方法，还有受限玻尔兹曼机(RBM)以及基于图的模型、基于深度学习的方法等等。

当然上述两类方法可以混合起来使用，即使同一类方法中的不同做法也可以组合使用，通过综合使用多种方法能够提高推荐的效果。

在信息推荐时，有时候不仅仅只使用用户-物品矩阵，还可以引入一些外部资源。下面主要介绍几种引入外部资源的方法，包括使用人口统计学、社会化过滤、位置信息过滤等方法的推荐模型。下面分别对它们进行简单的介绍。

3.3 基于人口统计学的过滤方法

前面提到挖掘用户的兴趣是信息推荐中十分重要的问题。然而，对于新注册用户而言，由于还没有充分了解其喜好，因此无法对其进行有效推荐，这也是前面提到的所谓“冷启动”问题。该问题的一种解决方法是利用用户的人口统计学特征。每个用户都有自己的人口统计学(Demographic)特征，包括年龄、性别、职业、学历、居住地、国籍等。这些信息对预测用户的兴趣也起着重要的作用。例如，不同的年龄段喜欢观看的电视剧种类是不同的，根据不同年龄段的不同，我们可以推荐给他们该年龄段比较喜欢看的电视剧。例如推荐给儿童动画片，推荐给青少年男女偶像剧，推荐给老年人戏曲等等。

基于用户统计学特征进行过滤的一个最大优点是可以处理注册用户的冷启动问题。当一个新用户没有产生显式（评分/购买）或者隐式（观看/浏览）反馈等数据时，可以根据用户的年龄、性别、国籍、身份等人口统计学特征来预测该用户的兴趣。当然，基于人口统计特征的主要缺点是推荐粒度比较大，只区分了不同的群体，并没有真正实现用户的个性化。而且，很多用户在填写个人信息时，出于隐私考虑，可能不会提供真实的信息，使得基于人口统计学的推荐方法产生的结果误差较大。

3.4 基于社会化过滤的推荐方法

随着社交网站(例如 Twitter、新浪微博等网站)的兴起，大量用户之间具有社交关系。如何利用这些社交关系设计推荐系统是最近几年推荐系统领域热门的问题。基于社交网络的推荐算法被称为社会化过滤(Social Filtering)。在社会化过滤方面，最常见的做法是在利用传统用户-物品喜好信息的基础上，增加用户的之间的信任度信息，从而联合构建信息推荐模型。在利用用户之间的信任度时，还可以对社交网络中的社区进行挖掘，从而在进行推荐时同时考虑两两朋友之间的关系及用户组的兴趣模型。

基于社会化过滤的推荐方法的优点是:可以使用社会关系缓解在电子商务或者其他推荐系统中遇到的数据稀疏性问题;由于人们的选择常常受到社会关系的影响,因此引入社会化过滤可以推荐出新的物品,从而增加推荐结果的多样性(Diversity)和用户的惊喜度(Serendipity)。社会化过滤的主要缺点是,用户之间的社会关系形成原因很多,但是只有兴趣相近的关系对用户推荐有比较大的作用。因此,如何鉴别不同的社会关系对预测用户不同行为的作用是社会化过滤中的一个重要的研究方向。

3.5 基于位置的过滤

随着移动终端、无线网络的普及,在很多智能手机或者其他设备上都有 GPS 定位的功能,因此最近几年考虑位置的过滤(Location-based Filtering)也得到了人们的关注。 例如,通过用户的位置,进行对用户推荐在他附近的好友,以及在他附近的他可能喜欢的商场,饭馆等。可以利用不同的用户在不同时间下的活动信息,对用户进行推荐。另外,基于位置的过滤还被用在旅游领域主要是通过分析用户的位置,给用推荐他们好友在附近旅游的景点。在很多基于位置的推荐算法中,位置信息主要被认作为一种上下文信息。位置信息往往和传统用户-物品数据综合使用。

3.6 研究现状

前面提到,信息推荐面临很多难题。下面分别针对这些难题来介绍当前的大致研究现状。

针对数据稀疏性问题,一种做法就是对用户-物品矩阵进行填充。最简单的填充办法就是将用户对未评分项目的评分设为一个固定的缺省值,或者设为其他用户对该项目的平均评分。当然由于用户对未评分项目的评分不可能完全相同,上述办法不能从根本上解决稀疏性问题。很多研究人员采用预测评分的方法来填充用户-物品矩阵,能够产生较理想的推荐效果。典型的预测评分方法包括 BP 神经网络、朴素贝叶斯、矩阵分解等方法。另一种做法是传递法,首先构建用户图或物品图或用户-物品图,然后图上运行随机游走之类的算法来填充矩阵从而进行推荐。此外,前面提到的通过融合上下文(时间、位置、人口统计学信息、物品的标签信息)的做法也可以认为是某种程度上弥补了数据性稀疏性的不足。虽然这些方法往往并不直接对用户-物品矩阵填充,但是通过补充上下文信息,它们能够缓解数据稀疏性,从而提高信息推荐的效果。

信息推荐面临的一个重要问题是冷启动问题。赌博机(Bandit)算法常用于处理推荐中的冷启动问题。它的主要思想是通过多次尝试加上概率预估,来选择最有可能获得最大收益的用户兴趣来进行推荐。赌博机算法又有多种具体的执行策略,目前在信息推荐领域能取得一定的结果。另一种做法利用用户的描述信息(如人口统计学信息)或者行为信息进行推荐。从这些信息中猜测出用户的兴趣,然后进行推荐。

大规模信息推荐是当前信息推荐领域面对的另一个挑战。目前的一种做法是引入分布式计算框架(如 Hadoop),对问题的规模分而治之进行求解。另一种做法是减少匹配计算的次数,比如引入 Hash 算法,来剔除那些不需要进行匹配计算的用户-物品对。

此外,更多的研究和具体的领域相结合,产生了大量的推荐应用,这些应用中又会面临该领域具体的问题。比如,社交网络中的朋友推荐,往往就要考虑用户本身的信息以及用户的关注转发等关系,结合内容和结构两个方面的信息进行推荐。

需要一提的是深度学习和推荐的结合。和其他领域相比,信息推荐领域和深度学习的结合并没有那么风声水起。研究者已经将深度学习用于音乐推荐应用,取得了很好的效果。包括 Google、Amazon、Netflix 在内的著名公司都号称在推荐领域使用了深度学习,特别是 Google 在 2016 年 6 月发布了宽度&深度学习(Wide & Deep Learning)开源框架,并号称已经用于推荐系统。2016 年 9 月,第一届“基于深度学习的推荐系统”研讨会将和 ResSys2016 一起召开。由于深度学习的本质目标就是学习到对象的特征表示,而推荐系统的核心就是生成用户兴趣和物品的表示,因此,它们的目标是一致的。可以预见,深度学习也会在推荐领域得到重要应用。

4. 技术展望与发展趋势

从信息推荐技术的发展来看，与具体领域的结合是推荐领域最重要的研究话题。不论是电子商务网站的商品推荐，还是社交网络中的朋友推荐，都有其自身领域具体的研究问题。通用的推荐技术不充分与领域的具体问题相结合，难以发挥出最佳效果。因此，在推荐中充分考虑领域知识、领域数据的特点，是领域信息推荐的重要研究方法。

另一方面，如何深入理解用户的需求，从帮助用户完成特定任务的多个环节入手，打通不同平台与应用之间的鸿沟，扩展推荐目标对象的类型，融合多个领域、多种应用和平台、多种模态的数据，进行基于任务的跨领域异质信息的精准推荐，也是一个重要的研究课题。

信息推荐的精度仍然有大幅度的提高空间，面向群体和个人的用户画像、用户的兴趣表示和匹配仍然需要更深入的研究。如何挖掘并综合利用上下文信息来提高推荐的可靠性是长盛不衰的话题。特别是随着手机的大规模普及，融入位置信息的推荐是一个重要的研究方向。

传统的推荐方法基本是黑盒的，直接为用户输出推荐的结果，而没有给出充分的有说服力的理由。近年来，可解释的推荐，开始受到越来越多的关注。

大规模推荐技术是大数据时代的重要研究课题。如何在极大规模数据的条件下，进行卓有成效的实时推荐，是一个值得研究的问题。

另外，如何利用深度学习进行有效推荐、在推荐中如何保护用户的隐私、如何对抗对推荐系统的攻击、如何实现多种目标下的信息推荐，都是十分重要的研究课题，值得进一步研究。

第十三章 自动问答研究进展、现状及趋势

1.任务定义、目标和研究意义

自动问答（Question Answering, QA）是指利用计算机自动回答用户所提出的问题以满足用户知识需求的任务。不同于现有搜索引擎，问答系统是信息服务的一种高级形式，系统返回用户的不再是基于关键词匹配排序的文档列表，而是精准的自然语言答案。近年来，随着人工智能的飞速发展，自动问答已经成为倍受关注且发展前景广泛的研究方向。

自动问答的研究历史可以溯源到人工智能的原点。1950 年，人工智能之父阿兰图灵（Alan M. Turing）在《Mind》上发表文章《Computing Machinery and Intelligence》，文章开篇提出通过让机器参与一个模仿游戏（Imitation Game）来验证“机器”能否“思考”，进而提出了经典的图灵测试（Turing Test），用以检验机器是否具备智能。同样，在自然语言处理研究领域，问答系统被认为是验证机器是否具备自然语言理解能力的四个任务之一（其它三个是机器翻译、复述和文本摘要）。自动问答研究既有利于推动人工智能相关学科的发展，也具有非常重要的学术意义。

从应用上讲，现有基于关键词匹配和浅层语义分析的信息服务技术已经难以满足用户日益增长的精准化和智能化信息需求，已有的信息服务范式急需一场变革。2011 年，华盛顿大学图灵中心主任 Etzioni 在 Nature 上发表的《Search Needs a Shake-Up》中明确指出：在万维网诞生 20 周年之际，互联网搜索正处于从简单关键词搜索走向深度问答的深刻变革的风口浪尖上。以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态。同一年，以深度问答技术为核心的 IBM Watson 自动问答机器人在美国智力竞赛节目 Jeopardy 中战胜人类选手，引起了业内的巨大轰动。Watson 自动问答系统让人们看到已有信息服务模式被颠覆的可能性，成为了问答系统发展的一个里程碑。此外，随着移动互联网崛起与发展，以苹果公司 Siri、Google Now、微软 Cortana 等为代表的移动生活助手爆发式涌现，上述系统都把以自然语言为基本输入方式的问答系统看做是下一代信息服务的新形态和突破口，并均加大人员、资金的投入，试图在这一次人工智能浪潮中取得领先。

当然，现有自动问答技术还不完美，仍面临许多具体问题和困难。本文对自动问答的主要研究内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势，进行概要介绍。

2.研究内容和关键科学问题

自动问答系统在回答用户问题时，需要正确理解用户所提的自然语言问题，抽取其中的关键语义信息，然后在已有语料库、知识库或问答库中通过检索、匹配、推理的手段获取答案并返回给用户。上述过程涉及词法分析、句法分析、语义分析、信息检索、逻辑推理、知识工程、语言生成等多项关键技术。传统自动问答多集中在限定领域，针对限定类型的问题进行回答。伴随着互联网和大数据的飞速发展，现有研究趋向于开放域、面向开放类型问题的自动问答。概括地讲，自动问答的主要研究任务和相应关键科学问题如下。

2.1 问句理解

给定用户问题，自动问答首先需要理解用户所提问题。用户问句的语义理解包含词法分析、句法分析、语义分析等多项关键技术，需要从文本的多个维度理解其中包含的语义内容。

在词语层面，需要在开放域环境下，研究命名实体识别（Named Entity Recognition）、术语识别（Term Extraction）、词汇化答案类型词识别（Lexical Answer Type Recognition）、实体消歧（Entity Disambiguation）、关键词权重计算（Keyword Weight Estimation）、答案集中词识别（Focused Word Detection）等关键问题。在句法层面，需要解析句子中词与词之间、短语与短语之间的句法关系，分析句子句法结构。在语义层面，需要根据词语层面、句法层面的分析结果，将自然语言问句解析成可计算、结构化的逻辑表达形式（如一阶谓词逻辑表达式）。

2.2 文本信息抽取

给定问句语义分析结果，自动问答系统需要在已有语料库、知识库或问答库中匹配相关的信息，并抽取出相应的答案。传统答案抽取构建在浅层语义分析基础之上，采用关键词匹配策略，往往只能处理限定类型的答案，系统的准确率和效率都难以满足实际应用需求。为保证信息匹配以及答案抽取的准确度，需要分析语义单元之间的语义关系，抽取文本中的结构化知识。早期基于规则模板的知识抽取方法难以突破领域和问题类型的限制，远远不能满足开放领域自动问答的知识需求。为了适应互联网实际应用的需求，越来越多的研究者和开发者开始关注开放域知识抽取技术，其特点在于：1）文本领域开放：处理的文本是不限定领域的网络文本；2）内容单元类型开放：不限定所抽取的内容单元类型，而是自动地从网络中挖掘内容单元的类型，例如实体类型、事件类型和关系类型等。

2.3 知识推理

自动问答中，由于语料库、知识库和问答库本身的覆盖度有限，并不是所有问题都能直接找到答案。这就需要在已有的知识体系中，通过知识推理的手段获取这些隐含的答案。例如，知识库中可能包括了一个人的“出生地”信息，但是没包括这个人的“国籍”信息，因此无法直接回答诸如“某某人是哪国人?”这样的问题。但是一般情况下，一个人的“出生地”所属的国家就是他（她）的“国籍”。在自动问答中，就需要通过推理的方式学习到这样的模式。传统推理方法采用基于符号的知识表示形式，通过人工构建的推理规则得到答案。但是面对大规模、开放域的问答场景，如何自动进行规则学习，如何解决规则冲突仍然是亟待解决的难点问题。目前，基于分布式表示的知识表示学习方法能够将实体、概念以及它们之间的语义关系表示为低维空间中的对象（向量、矩阵等），并通过低维空间中的数值计算完成知识推理任务。虽然这类推理的效果离实用还有距离，但是我们认为这是值得探寻的方法，特别是如何将已有的基于符号表示的逻辑推理与基于分布式表示的数值推理相结合，研究融合符号逻辑和表示学习的知识推理技术，是知识推理任务中的关键科学问题。

3.技术方法和研究现状

根据目标数据源的不同，已有自动问答技术大致可以分为三类：1）检索式问答；2）社区问答以及 3）知识库问答。以下分别就这几个方面对研究现状进行简要阐述。

3.1 检索式问答

检索式问答研究伴随搜索引擎的发展不断推进。1999 年，随着 TREC QA 任务的发起，检索式问答系统迎来了真正的研究进展。TREC QA 的任务是给定特定 WEB 数据集，从中找到能够回答问题的答案。这类方法是以检索和答案抽取为基本过程的问答系统，具体过程包括问题分析、篇章检索和答案抽取。根据抽取方法的不同，已有检索式问答可以分为基于模式

匹配的问答方法和基于统计文本信息抽取的问答方法。

基于模式匹配的方法往往先离线地获得各类提问答案的模式。在运行阶段，系统首先判断当前提问属于哪一类，然后使用这类提问的模式来对抽取的候选答案进行验证。同时为了提高问答系统的性能，人们也引入自然语言处理技术。由于自然语言处理的技术还未成熟，现有大多数系统都基于浅层句子分析。

基于统计文本信息抽取的问答系统的典型代表是美国 Language Computer Corporation 公司的 LCC 系统。该系统使用词汇链和逻辑形式转换技术，把提问句和答案句转化成统一的逻辑形式（Logic Form），通过词汇链，实现答案的推理验证。LCC 系统在 TREC QA Track 2001 ~ 2004 连续三年的评测中以较大领先优势获得第一名的成绩。

2011 年，IBM 研发的问答机器 Watson5在美国智力竞赛节目《危险边缘 Jeopardy!》中战胜人类选手，成为问答系统发展的一个里程碑。Watson 的技术优势大致可以分为以下三个方面：（1）强大的硬件平台：包括 90 台 IBM 服务器，分布式计算环境；（2）强大的知识资源：存储了大约 2 亿页的图书、新闻、电影剧本、辞海、文选和《世界图书百科全书》等资料；（3）深层问答技术（DeepQA）：涉及统计机器学习、句法分析、主题分析、信息抽取、知识库集成和知识推理等深层技术。然而，Watson 并没有突破传统问答式检索系统的局限性，使用的技术主要还是检索和匹配，回答的问题类型大多是简单的实体或词语类问题，而推理能力不强。

3.2 社区问答

随着 Web2.0 的兴起，基于用户生成内容（User-Generated Content, UGC）的互联网服务越来越流行，社区问答系统应运而生，例如 Yahoo! Answers6、百度知道7等。问答社区的出现为问答技术的发展带来了新的机遇。据统计 2010 年 Yahoo! Answers 上已解决的问题量达到 10 亿，2011 年“百度知道”已解决的问题量达到 3 亿，这些社区问答数据覆盖了方方面面的用户知识和信息需求。此外，社区问答与传统自动问答的另一个显著区别是：社区问答系统有大量的用户参与，存在丰富的用户行为信息，例如用户投票信息、用户评价信息、回答者的问题采纳率、用户推荐次数、页面点击次数以及用户、问题、答案之间的相互关联信息等等，这些用户行为信息对于社区中问题和答案的文本内容分析具有重要的价值。

一般来讲，社区问答的核心问题是从大规模历史问答对数据中找出与用户提问问题语义相似的历史问题并将其答案返回提问用户。假设用户查询问题为q_0，用于检索的问答对数据为$S_{Q,A}=\{(q_1,a_1),(q_2,a_2)\},\ldots,(q_n,a_n)\}\}$，相似问答对检索的目标是从$S_{Q,A}$中检索出能够解答问题$q_0$的问答对$(q_i,a_i)$。

针对这一问题，传统的信息检索模型，如向量空间模型、语言模型等，都可以得到应用。但是，相对于传统的文档检索，社区问答的特点在于：用户问题和已有问句相对来说都非常短，用户问题和已有问句之间存在“词汇鸿沟”问题，基于关键词匹配的检索模型很难达到较好的问答准确度。目前，很多研究工作在已有检索框架中针对这一问题引入单语言翻译概率模型，通过 IBM 翻译模型，从海量单语问答语料中获得同种语言中两个不同词语之间的语义转换概率，从而在一定程度上解决词汇语义鸿沟问题。例如和“减肥”对应的概率高的相关词有“瘦身”、“跑步”、“饮食”、“健康”、“远动”等等。 除此之外，也有许多关于问句检索中词重要性的研究和基于句法结构的问题匹配研究。

3.3 知识库问答

检索式问答和社区问答尽管在某些特定领域或者商业领域有所应用，但是其核心还是关键词匹配和浅层语义分析技术，难以实现知识的深层逻辑推理，无法达到人工智能的高级目

[5] http://www.ibm.com/smarterplanet/us/en/ibmwatson/

[66] http://answers.yahoo.com/

[7] http://zhidao.baidu.com/

标。因此，近些年来，无论是学术界或工业界，研究者们逐步把注意力投向知识图谱或知识库(Knowledge Graph)。其目标是把互联网文本内容组织成为以实体为基本语义单元(节点)的图结构，其中图上的边表示实体之间语义关系。目前互联网中已有的大规模知识库包括DBpedia、Freebase、YAGO 等。这些知识库多是以“实体-关系-实体”三元组为基本单元所组成的图结构。基于这样的结构化知识，问答系统的任务就是要根据用户问题的语义直接在知识库上查找、推理出相匹配的答案，这一任务称为面向知识库的问答系统或知识库问答。

要完成在结构化数据上的查询、匹配、推理等操作，最有效的方式是利用结构化的查询语句，例如：SQL、SPARQL 等。然而，这些语句通常是由专家编写，普通用户很难掌握并正确运用。对普通用户来说，自然语言仍然是最自然的交互方式。因此，如何把用户的自然语言问句转化为结构化的查询语句是知识库问答的核心所在，其关键是对于自然语言问句进行语义理解（如图 1 所示）。目前，主流方法是通过语义分析，将用户的自然语言问句转化成结构化的语义表示，如λ范式和 DCS-Tree。相对应的语义解析语法或方法包括组合范畴语法（Category Compositional Grammar, CCG）以及依存组合语法（Dependency-based Compositional Semantics, DCS）等。

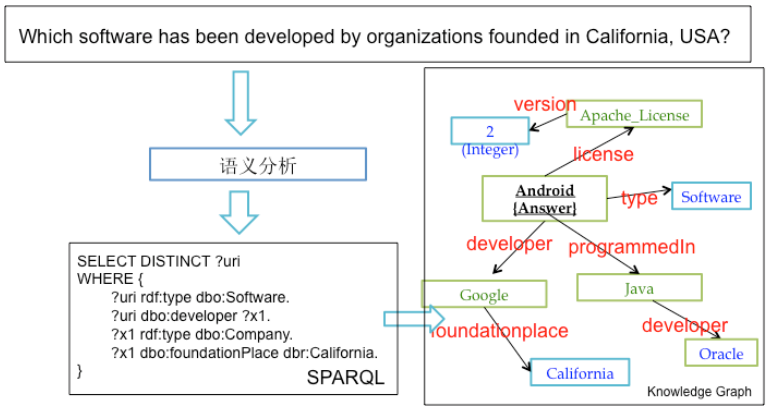


图 1. 知识库问答过程

尽管很多语义解析方法在限定领域内能达到很好的效果，在这些工作中，很多重要组成部分（比如 CCG 中的词汇表和规则集）都是人工编写的。上述方法当面对大规模知识库时会遇到困难，如词汇表问题（在面对一个陌生的知识库时，不可能事先或者用人工方法得到这个词汇表)。目前已有许多工作试图解决上述问题，如利用数据回标方法扩展 CCG 中的词典，挖掘事实库和知识库在实例级上的对应关系确定词汇语义形式。

但是，上述方法的处理范式仍然是基于符号逻辑的，缺乏灵活性，在分析问句语义过程中，易受到符号间语义鸿沟影响。同时从自然语言问句到结构化语义表达需要多步操作，多步间的误差传递对于问答的准确度也有很大的影响。近年来，深度学习技术以及相关研究飞速发展，在很多领域都取得了突破，例如图像、视频和语音等，在自然语言处理领域也逐步开始应用。其优势在于通过学习能够捕获文本（词、短语、句子、段落以及篇章）的语义信息，把目标文本投射到低维的语义空间中，这使得传统自然语言处理过程中很多语义鸿沟的现象通过低维空间中向量间数值计算得到一定程度的改善或解决。因此越来越多的研究者开始研究深度学习技术在自然语言处理问题中的应用，例如情感分析、机器翻译、句法分析等等，知识库问答系统也不例外。与传统基于符号的知识库问答方法相比，基于表示学习的知识库问答方法更具鲁棒性，其在效果上已经逐步超过传统方法，如图 2 所示。这些方法的基本假设是把知识库问答看做是一个语义匹配的过程。通过表示学习，我们能够把用户的自然语言问题转换为一个低维空间中的数值向量（分布式语义表示），同时知识库中的实体、概念、类别以及关系也能够表示成同一语义空间的数值向量。那么传统知识库问答任务就可以看成问句语义向量与知识库中实体、边的语义向量之间的相似度计算过程。

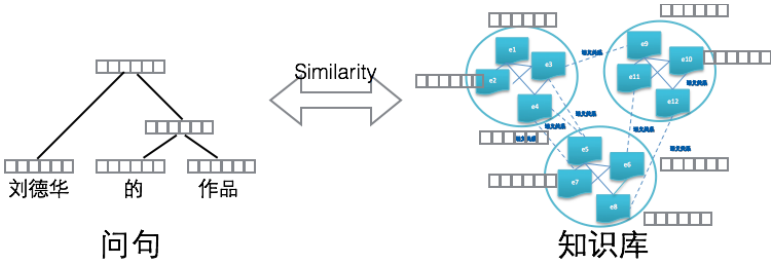


图 2. 基于表示学习的知识库问答示意图

3.4 技术现状

根据上面的阐述可以看到，根据不同的技术路线，检索式问答、社区问答以及知识库问答所采用的评测数据集也不尽相同。

在检索式问答方面，最权威的评测是美国国家标准技术研究所(NIST)推动的 TREC(Text Retrieval Evaluation Conference）于 1999 年开始组织的问答评测任务（QA Track）8 和 NTCIR（NII Testbeds and Community for Information access Research）组织的跨语言问答评测任务（CLQA）9。TREC QA 评测考察三类不同的问题：事实性（factoid)、列表类（list）和定义类（definition)。然后综合这三类问题的平均得分，对于参评系统进行评价。TREC QA 评测一直持续到 2007 年，该评测一直都是检索式问答领域最受关注、参加机构最多的 TREC 评测项目之一。根据 2007 年的评测结果来看，最好的评测系统 MRR（Mean Reciprocal Rank）可以达到 0.48 以上（接近 0.5 意味着评测系统对于所有的问题将在前两位返回结果中获得正确答案)。基于 TREC 评测系统，IBM 公司于 2011 年研发的 Watson 系统参加了美国 Jeopardy 知识比赛，并战胜了人类选手，可以看做是检索式问答系统的一个里程碑。但是不可忽略的是，Jeopardy 比赛还是一个限定问题类型、限定答案类型的知识比赛，面对开放式的场景和环境，已有检索式问答系统还有很长的路要走。

在社区问答方面，目前并没有权威的评测数据集，公认的数据集通常是由 Yahoo! Answers 社区问答系统上利用为研究人员提供的 API10接口下载的。目前，最好的检索系统在 Top 10 的准确率可以达到 40%。尽管社区问答系统相对于检索式问答和知识库问答技术简单，但是目前已经商业化，例如 Yahoo Answer11和百度知道12。

在知识库问答方面，已有的评测主要针对于一些限定领域的知识库进行问答。已有方法也取得了不错的结果。例如：在 Geoquery13（美国地理知识查询）数据集上（600 个训练样本，280 个测试样本）上，使用 CCG 和本体匹配的方法 F 值能达到 89.0%，使用 DCS 的方法 F 值能达到 91.1%；在求职（JOBS）数据集上（500 个训练样本，140 个测试样本)，使用 CCG 的方法 F 值能达到 79.3%,使用 DCS 的方法 F 值能达到 95%。在这一方面，QALD（Question Answering over Linked Data）评测的举办更是推动了这方面的研究。QALD 每年举办一届，目前已经举办到了第六届。每一次评测，组织者都会给出一些问题，要求参加评测系统在给定知识库的基础上，将所给问题转化为结构化的 SPARQL 查询语句，并在给定知识库上查询答案。但是，目前的研究趋势是从限定领域的知识库向大规模开放域甚至是多领域知识库进行扩展，例如 Freebase。与限定领域知识库相比，大规模开放知识库包含的资源和关系数量要大得多，比如 Geoquery 中只包含 8 个关系谓词，而 Freebase 包含上万个关系。因此开放知识库上的语义解析效果有明显下降。例如利用 Freebase 知识库，开放查询测试的最好效果只有 39.9%；而在 QALD 评测中，在 DBpedia 上、开放查询中，表现最好的问答系统的正确率只有 40%。下图给出在面对开放域知识库 Freebase 时，在公开问题库 WebQuestion

[8] http://trec.nist.gov/data/qamain.html
[9] http://www.slt.atr.jp/CLQA/
10 http://developer.yahoo.com/answers
11 https://answers.yahoo.com/
12 http://zhidao.baidu.com
13 http://www.cs.utexas.edu/users/ml/nldata/geoquery.html

上，已有系统能够达到的精度。

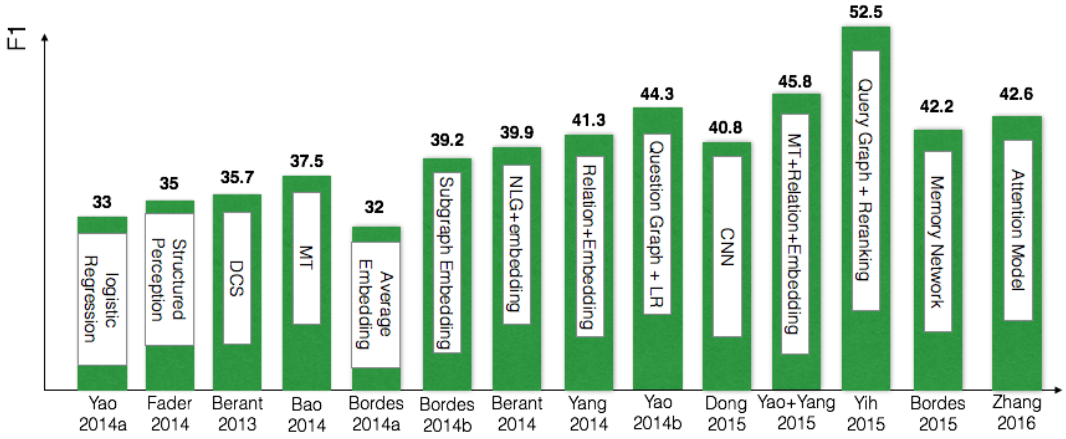


图 3. 已有知识库问答方法在 WebQuestion 问题集上的性能

4. 总结和展望

纵观自动问答研究的发展态势和技术现状，以下研究方向或问题将可能成为未来整个领域和行业重点关注的方向：

4.1 基于深度学习的端到端自动问答

目前，基于深度学习的问答系统试图通过高质量的问题-答案语料建立联合学习模型，同时学习语料库、知识库和问句的语义表示及它们相互之间的语义映射关系，试图通过向量间的数值运算对于复杂的问答过程进行建模。这类方法的优势在于把传统的问句语义解析、文本检索、答案抽取与生成的复杂步骤转变为一个可学习的过程，虽然取得了一定的效果，但是也存在很多问题。例如： 1）资源问题。深度学习的方法依赖大量的训练语料，而目前获取高质量的问题-答案对仍然是个瓶颈。Bordes 提出了一些模板利用已有三元组来生成问句，用较小的代价生成了大量的问题-答案对，但是相应的问句质量并不能保证，而且问句同质化严重。在训练资源上的提高空间仍然很大。2）已有的基于深度学习的问答方法多是针对简单问题（例如单关系问题）设计的，对于复杂问题的回答能力尚且不足。如何利用深度学习的方法解决复杂问题值得继续关注。

4.2 多领域、多语言的自动问答

开放域环境下，用户的问题复杂多样，很多场景下，单单只用一个数据源或单一语言的语料库或知识库的信息不能完全回答用户的问题，需要对于多个资源进行综合利用。然而，网络中不同来源的语料库和知识库在框架和内容层面均存在差异，同时也存在大量冗余。已有自动问答方法只能处理单一数据源的问答操作，尚缺乏对于多源异构知识库异构性和冗余性的有效处理机制。

4.3 面向问答的深度推理

尽管已有网络知识资源规模巨大，能够覆盖多个领域，但仍旧面临信息缺失问题，给知识库问答带来巨大挑战。这就需要研究面向问答的深度推理技术。传统基于符号逻辑的逻辑

推理方法基于严格的符号匹配，过分依赖于推理规则的生成，因此具有领域适应性差、无法进行大规模推理的缺点。而深度学习基于分布式语义表示，利用语义空间中的数值模糊计算替代传统问答过程中的符号严格匹配，为解决上述问题供了一种有效途径，但也存在推理结果准确度低、可解释性差的问题。因此，如何利用深度学习大规模、可学习的特点，在深度神经网络框架下，融入传统的逻辑推理规则，构建精准的大规模知识推理引擎是自动问答迫切需要解决的难点问题。

4.4 篇章阅读理解

机器阅读理解是近几年兴起的问答任务，类似于传统的问答任务，核心在于考察机器的文本理解和推理能力。从这个角度上说，我们可以把机器阅读理解任务看作是问答系统的延伸。但是，机器阅读和传统问答仍然存在区别，主要在于：传统问答任务往往要求系统根据用户所提的问题，在海量文本库或大规模结构化知识库中检索、抽取或推理出相应的答案，大多数情况下会利用海量数据的冗余特性对于答案进行检索和抽取。因此，传统问答任务多考察系统的文本匹配、信息抽取水平。而在阅读理解任务当中，系统被要求回答一些非事实性的、高度抽象的问题。同时，信息源被限定于给定的一篇文章，虽然可以利用一些已有背景知识，但是问题的答案往往来源于当前给定篇章中的文本。特别考察系统对于文本的细致化的自然语言理解能力以及已有知识的运用能力和推理能力。从这个角度上来说，相对于传统问答任务，机器阅读理解更具挑战。

4.5 对话

传统的自动问答都是采用一问一答的形式。然而在很多场景下，需要提问者和系统进行多轮对话交互，完成问答过程。针对这一问题，已有研究已经提出若干方法，但是由于场景的开放性以及用户问题的复杂度，这一问题一直难以很好解决。特别是在对话上下文建模与知识表示、对话策略学习以及对话准确性评价等方面亟待解决。

总之，自动问答作为人工智能技术的有效评价手段，已经研究了 60 余年。整体上，自动问答技术的发展趋势是从限定领域向开放领域、从单轮问答向多轮对话、从单个数据源向多个数据源、从浅层语义分析向深度逻辑推理不断推进。我们有理由相信，随着自然语言处理、深度学习、知识工程和知识推理等相关技术的飞速发展，自动问答在未来有可能得到相当程度的突破。伴随着 IBM Watson、Apple Siri 等实际应用的落地与演进，我们更有信心看到这一技术将在不远的未来得到更大、更广的应用。

第十四章 机器翻译研究进展、现状及趋势

1.任务定义、目标和研究意义

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，机器翻译研究的目标就是建立有效的自动翻译方法、模型和系统，打破语言壁垒，最终实现任意时间、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。

人们通常习惯于感知（听、看和读）自己母语的声音和文字，很多人甚至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任务，包括中文信息处理技术的快速发展。从应用上讲，无论是社会大众、政府企业还是国家机构，都迫切需要机器翻译技术。特别是在“互联网+”时代，以多语言多领域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，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、53 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技术。

当然，机器翻译技术还不完美，它仍面临很多具体问题和困难。本文对机器翻译研究的主要内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势，做简要介绍。

2.研究内容和关键科学问题

一般来说，一种自然语言被翻译成另外一种语言时，需要经过源语言的理解、源语言到目标语言的转换和目标语言生成这三个基本步骤。因此，机器翻译是一项综合性的复杂技术，既涉及源语言的处理问题，又涉及两种语言之间的转换和目标语言的生成问题，可以说几乎自然语言处理中的所有问题都会在机器翻译中出现，包括：词法分析（或称形态分析）与词语切分、命名实体识别、句法分析、词义消歧与句子语义表示、自然语言句子生成等，当然在其中最重要的是翻译模型构建问题。随着数据驱动方法出现，面向翻译的双语甚至多语对照语料的自动获取与语料库构建和翻译知识自动学习等都是机器翻译研究的内容。

概括地讲，机器翻译的主要研究任务和相应关键科学问题如下。

2.1 源语言语句的理解

语言翻译本质上是利用目标语言准确地表达源语言语句所呈现的语义，因此，对源语言

语句的语义理解成为机器翻译研究面临的首要科学问题。源语言理解包括词语、短语、句子和语篇（对话）等多层次语言单位的处理。在词语层面，需要研究如何界定和切分适合机器翻译的基本语言单元（如汉语、日语、越南语、泰国语等语言的分词问题）、多义词的歧义消解问题以及指代、省略等语言现象的歧义消除等问题。在短语和句子层面，需要解析句子中词与词之间、短语与短语之间的语义关系。语义理解得越深刻，越有利于机器翻译过程。例如，若能准确获得源语言语句的谓词-论元结构关系（反映句子的谓词与实施者、受事者和修饰成分之间的语义关系），就能够把握语句的主框架和转换核心。在语篇层面，需要分析句子之间的结构和语义关系，如句子之间的连贯性、衔接性等。

2.2 源语言到目标语言的转换

源语言翻译单元（如词、短语等）到目标语言的转换是机器翻译最关键的问题，也是大多数机器翻译方法关注的焦点。其中，转换规则（有时候我们称其为“翻译知识”）的表示和获取是两大核心问题。

翻译知识表示主要涉及翻译规则表达的知识层次和表示形式。在知识层次方面，一般有基于词、基于短语、基于句法子树和基于语义结构树等不同层次，而在表示形式方面，通常有基于离散符号的表示方法和基于连续实数空间的表示方法两种。

翻译知识获取方法主要有两种，一种是基于理性主义思路的专家经验总结和手工编写方式；另一种是基于经验主义思想的以数据驱动的自动获取方法。近年来，将翻译知识隐含在神经网络结构和参数中的方法实际上也是一种经验主义方法。

2.3 目标语言语句的生成

目标语言语句生成的目的是针对给定源语言句子，根据翻译知识生成（搜索出）一组最佳译文片段集合，使其能够按照目标语言的语法自然、流畅地表达源语言句子的含义。其中，译文片段的组合方式和目标语言句子的流畅性是目前译文生成中重点关注的两个问题。根据翻译系统所接受的输入类型和系统工作方式的不同，机器翻译又可分为文本机器翻译（text-to-text machine translation）、口语翻译（spoken language translation, SLT）和计算机辅助翻译（computer assisted machine translation）。默认情况下，一般指文本机器翻译或者泛指。不同机器翻译系统涉及的研究问题略有不同，如交互式人机辅助机器翻译系统主要涉及如何通过人机交互过程自动获取翻译知识和实时更新系统性能的问题；在口语翻译中涉及到如何处理含语音识别错误和噪声的非规范口语句子的问题等。在本文后面的部分里，我们专注于翻译方法的问题，基本在“泛指”的层面，而不专门针对某一种翻译系统阐述。

3. 技术方法和研究现状

自机器翻译诞生以来，其研究方法一直在理性主义方法和经验主义方法的交替中前进。理性主义方法就是指以语言学理论为基础，由语言学家手工编写翻译规则和词典，基于规则的翻译（rule-based machine translation）方法是其中的典型代表；经验主义方法则是以信息论和数理统计为理论基础，以大规模语料库为驱动，通过机器学习技术自动获取翻译知识，这种方法又被称为基于语料库的翻译方法（corpus-based machine translation），或者数据驱动的翻译方法（data-driven machine translation）。

3.1 基于规则的翻译方法

自上个世纪 50 年代后期到 90 年代初期，一直是基于规则的翻译方法占据主导地位。在基于规则的翻译方法中，源语言句子分析、源语言到目标语言的转换和目标语言的生成都是由基于规则的方法完成，而所有规则几乎都是由通晓双语的语言学专家总结、编纂获得的。

基于规则的翻译方法的效果与源语言分析和目标语言生成技术的水平以及转换规则的质量有着密切的关系。由于这种方法能够充分利用语言学家总结出来的语言规律，具有一定的通用性，因此，对于符合源语言语法规范的句子一旦翻译正确，往往能够获得较高质量的译文。美国 SYSTRAN 公司和我国中软公司、华健公司、格微公司等开发的机器翻译系统都是基于规则方法完成的。规则翻译方法存在一些难以突破的瓶颈问题，如规则一般只能处理规范的语言现象，获取规则的人工成本较高，而且维护大规模的规则库往往比较困难，新规则与已有规则易发兼容性问题等。

3.2 基于语料库的翻译方法

基于语料库的翻译方法主张从已知的翻译实例中自动学习两种语言之间的转换规则。随着互联网技术的快速发展和普及，获取大规模双语平行或可比语料的机会持续增加，机器学习技术和计算机运算能力不断增强，基于语料库的数据驱动方法自上世纪 90 年代以来成为机器翻译研究的主流技术。按照翻译方法提出的先后次序，语料库方法又可进一步划分为基于实例的翻译方法（example-based machine translation）、统计翻译方法（statistical machine translation）和基于深度学习的翻译方法（deep-learning based machine translation）。

3.2.1 基于实例的翻译方法

基于实例的机器翻译方法是由日本学者长尾真教授于 1980 年代提出来的。这种方法试图在事先构建的翻译实例库中找出与待翻译的源语言句子相似的实例（通常是句子），并根据待翻译句子的具体情况对实例对应的译文进行适当的替换、删除和插入等操作，实现翻译过程。

基于实例的翻译方法无需对源语言句子进行复杂的分析，可充分利用已经确认的翻译实例。但是，由于该方法采用的实例一般是句子，因此，如何从大规模实例库中快速找到相似度很高的实例，尤其是语义高度相似的实例，始终是该方法面临的挑战。

3.2.2 统计翻译方法

统计翻译方法是上个世纪 80 年代末期和 90 年代初期由 IBM 公司首先实现的基于噪声信道模型的翻译方法，随后研究者们推出了一系列改进的翻译模型，其基本思想是利用机器学习技术从大规模双语平行语料中自动获取翻译规则和概率参数，然后利用翻译规则对源语言句子进行解码。如果将源语言句子看作是噪声信道模型的输出信号（观察序列）S，目标译文看作是噪声信道模型的输入信号 T，那么，机器翻译可以简单地看作求解 $\arg\max_T P(T|S)$ 的过程。根据贝叶斯公式，可以得到如下变换：

$$\arg\max_T P(T|S) \propto \arg\max_T P(T)\times P(S|T)$$

因此，在统计翻译方法中有三个关键技术模块：语言模型（language model）、翻译模型（translation model）和解码器（decoder）。语言模型用于计算候选译文的句子概率，翻译模型用于计算给定候选译文时源语言句子的概率，解码器用于快速搜索语言模型概率与翻译模型概率相乘之后概率最大的候选译文。为了融入更多的翻译特征，噪声信道模型逐渐

被对数线性模型所取代。

在 20 多年的发展历程中，统计翻译方法经历了基于词、基于短语和基于句法树翻译模型的一系列转变。这些翻译模型可以简单归纳成如下金字塔图：

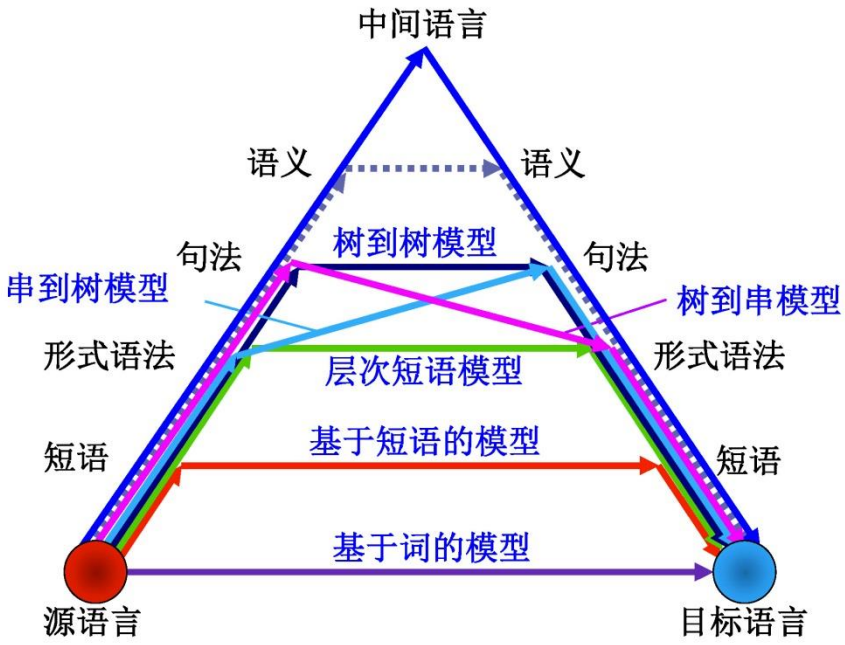


图 1. 机器翻译的金字塔

其中，基于短语的翻译模型是相对成熟的模型。这里所说的“短语”指连续同现的词串，并非语言学上定义的短语。该模型的基本思想是：在训练阶段从双语句子对齐的平行语料中自动抽取源语言短语到目标语言短语的翻译规则及其概率，在翻译阶段将源语言句子切分为短语序列，利用翻译规则得到目标语言句子的短语序列，然后借助短语重排序模型和语言模型对目标语言句子的短语序列进行排序，最终获得最佳的目标译文。很多后续工作主要目的是解决候选译文的消岐问题和目标译文短语的重排序问题。

层次化短语翻译模型通过在翻译规则中引入变量提高了翻译规则的表达能力和泛化能力，从而可获得较好的翻译性能。但由于仅采用唯一的非终结符 X 表示所有的变量，导致了过度泛化的问题。

之后，研究者们不断将语言学知识融入到翻译建模过程中，出现了一批基于句法的翻译模型，包括：树到树（tree-to-tree）的翻译模型、树到串（tree-to-string）的翻译模型和串到树（string-to-tree）的翻译模型等。最近几年，句法翻译模型主要针对其中的两个问题开展研究：（1）句法结构树与词语对齐不兼容；（2）双语的句法知识很难同时被有效地利用。

由于统计翻译方法具有很多规则方法所不具备的优势，如开发速度快、周期短、无需人工干预等，在特定领域训练数据充分的情况下译文虽不完美，但也能够达到可理解的水平，因此成为谷歌、微软、百度和有道等互联网公司在线翻译服务系统的核心技术。值得提及的是，在国家“十二五”计划“863”项目的支持下，由百度公司牵头，中科院自动化所、浙江大学、哈尔滨工业大学、中科院计算所和清华大学联合完成的基于互联网大数据的统计机器翻译产业化项目获得了成功应用，荣获 2015 年度国家科技进步奖二等奖，标志着统计翻译方法已经步入普惠大众、服务社会的实用阶段。

3.2.3 基于深度学习的翻译方法

2013 年以来，由于深度学习方法在特征表示和端到端建模方面具有独特的优势，基于深度学习的方法逐渐成为机器翻译领域的研究热点。2015 年之前，基于深度学习的机器翻

译主要以统计翻译为框架，旨在改进源语言句子解析、翻译转换和目标译文生成中的某些关键技术，如词语对齐、翻译概率估计、短语重排序和语言模型等。其核心思想在于利用深度学习方法的分布式表示，解决统计翻译方法对全局上下文和深层语义信息建模难的问题。在传统离散符号表示体系中，任意词语（短语或句子）都是独立的存在，词语（短语或句子）之间没有任何关系，一方面导致特征表示的数据稀疏问题，另一方面无法挖掘词语（短语或句子）间的语义相似性。在分布式表示体系中，词语（短语或句子）由低维、连续实数空间中的点表示，从而便利了深度特征组合与表示、以及语义计算。

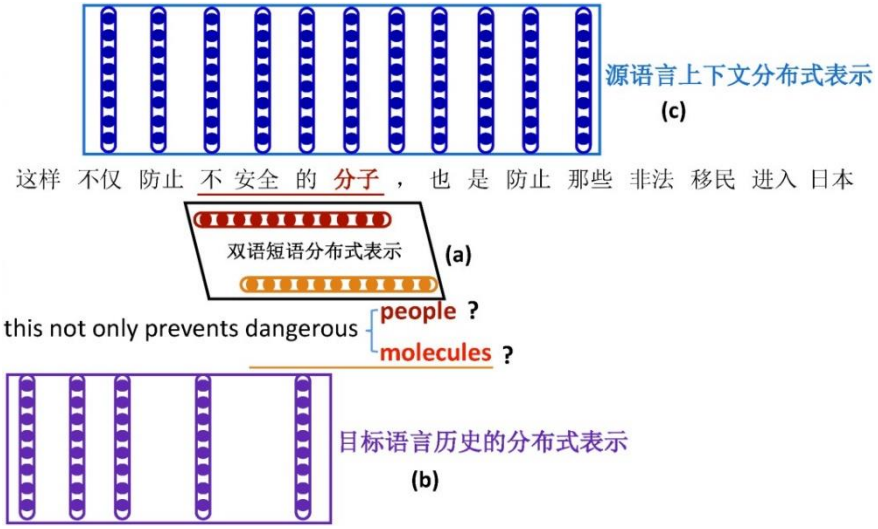


图 2. 深度学习在统计机器翻译中的应用：利用不同的方法对“分子”的译文进行消岐

以图 2 中对“分子”的目标译文进行消岐为例，(a) 利用深度学习获得翻译单元的深度语义表示，例如组合词语的表示得到双语短语分布式表示，在共享低维、实数向量空间中计算短语翻译规则“不 安全 的 分子 → dangerous people”和“不 安全 的 分子 → dangerous molecules”的互译置信度，从而选择语义距离最近的候选译文；(b) 采用深度学习对更远的目标语言历史信息进行建模，例如在语言模型中，循环神经网络可利用所有历史词语信息“this not only prevents dangerous”预测下一个词；(c) 利用深度学习以目标译文历史信息 n-1 个词的分布式表示以及源语言句子局部上下文的分布式表示作为神经网络输入，联合对源语言词语“分子”的译文进行预测。最后一种联合消岐方法由美国 BBN 公司的 J. Devlin 等人于 2014 年提出，通过分布式表示有效克服了语义鸿沟和数据稀疏的问题，显著改善了统计机器翻译的译文质量(在阿拉伯语到英语的翻译上获得了约 6 个 BLEU 值的提升)，也因此获得了 2014 年自然语言处理国际顶级会议 ACL 的最佳论文奖。

然而，尽管深度学习显著改善了统计机器翻译的效果，但是也明显增加了统计机器翻译系统的复杂度。

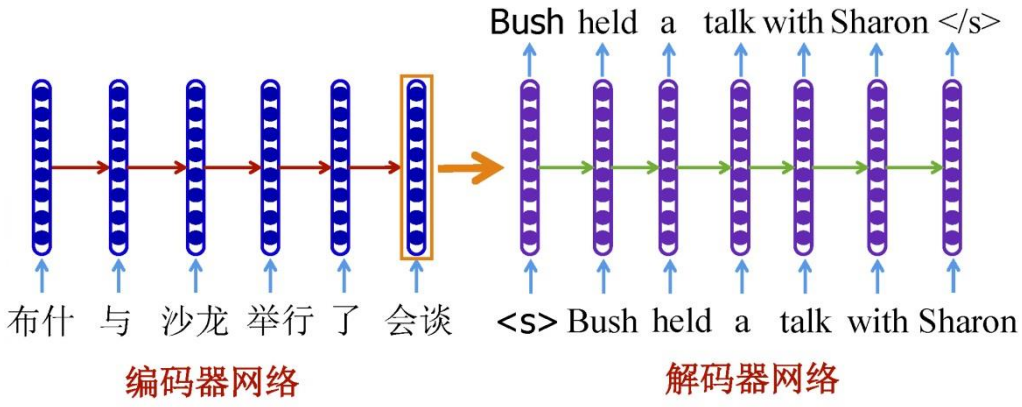


图 3. 基于编码-解码的端到端神经网络机器翻译

自 2014 年底以来，简单的端到端的神经网络翻译方法逐渐成为一种全新的机器翻译范

式，并迅速引起学术界和产业界的广泛关注和跟踪。不同于统计机器翻译中人工特征设计和流水线架构的实现方法（例如短语划分、翻译和重排序），端到端的神经网络翻译方法采用一个神经网络框架完成源语言文本到目标语言文本的直接转换。该思想由英国牛津大学的 Kalchbrenner 等人于 2013 年最先提出，后经 Sutskever 等人与 Bahdanau 等人分别在 2014 年和 2015 年进行了改进和扩展。

如图 3 所示，端到端的神经网络机器翻译方法仅包含两个神经网络：一个称为编码器（encoder），用于将源语言句子映射为一个（或一组）低维、连续的实数向量；另一个称为解码器（decoder），完成将源语言句子的向量表示转化为目标语言句子。在研究起初，无论源语言句子的长短，编码器仅将其映射到一个维数固定的实数向量，很难准确地表示源语言句子的完整语义。Bahdanau 等人将注意机制（attention）思想引入到了端到端的神经网络翻译模型：编码器生成并保留每个源语言词对应的上下文语义向量；解码器每次产生目标语言的单词时，首先利用注意机制模型计算当前译文应与源语言哪些位置的词语有关，然后加权得到源语言的上下文表示，最后用其预测当前译文的概率。目前，这种方法得到了越来越多研究者的青睐，使神经网络翻译方法在数据驱动的机器翻译领域异军突起。

据报道，在 2016 年针对欧洲语言的国际机器翻译评测中，基于端到端的神经网络翻译系统和统计翻译系统共同参与了 9 个评测任务，其中基于端到端的神经网络翻译系统在 7 个评测任务中以明显优势击败统计机器翻译系统。可见这种新的机器翻译范式所蕴含的强大发展潜力。目前，谷歌和百度公司都已在部分翻译语言对上采用端到端的神经网络翻译方法替代原来的统计翻译方法，取得了相对较好的翻译效果。

3.3 技术现状

根据上面的阐述可以看到，基于规则的机器翻译方法需要人工设计和编纂翻译规则，统计机器翻译方法能够自动获取翻译规则，但需要人工定义规则的形式，而端到端的神经网络机器翻译方法可以直接通过编码网络和解码网络自动学习语言之间的转换算法。从某种角度讲，其自动化程度和智能化程度在不断提升，机器翻译质量也得到了显著改善。

机器翻译技术的研究现状可从欧盟组织的国际机器翻译评测（WMT）的结果中窥得一斑。该评测主要针对欧洲语言之间的互译，2006 年至 2016 年每年举办一次。对比法语到英语历年的机器翻译评测结果可以发现，译文质量已经在自动评价指标 BLEU 值上从最初小于 0.3 到目前接近 0.4（大量的人工评测对比说明，BLEU 值接近 0.4 的译文能够达到人类基本可以理解的程度）。另外，中国中文信息学会组织的全国机器翻译评测（CWMT）每两年组织一次，除了英汉、日汉翻译评测以外，CWMT 还关注我国少数民族语言（藏、蒙、维）和汉语之间的翻译。相对而言，由于数据规模和语言复杂性的问题，少数民族与汉语之间的翻译性能要低于汉英、汉日之间的翻译性能。

虽然机器翻译系统评测的分值呈逐年增长的趋势，译文质量越来越好，但与专业译员的翻译结果相比，机器翻译还有很长的路要走，可以说，在奔向“信、达、雅”翻译目标的征程上，目前的机器翻译基本挣扎在“信”的阶段，很多理论和技术问题仍有待于更深入的研究和探索。

4. 总结和展望

纵观机器翻译研究发展的态势和技术现状，以下研究方向或问题将可能成为未来机器翻译研究必须攻克的堡垒：

端到端的神经网络机器翻译的优化：基于端到端的神经网络翻译方法具有较大发展潜力，但仍有诸多不足，如可解释性差导致译文错误无法追踪分析，计算复杂度高导致只有 GPU 服务器才能高效训练，网络结构简单导致难于融入先验信息。因此，增强模型的可解释性、降低神经网络的计算复杂度（使之能在 CPU 上高效训练）、以及设计更加合理的编码和解码网络都是未来广受关注的研究问题；

面向小数据的机器翻译：统计翻译方法和神经网络翻译方法都强烈依赖于大规模高质量的双语训练数据，而很多语言对（如藏语和汉语）或特定应用领域中往往没有足够多的双语平行语料，这就导致数据驱动的翻译方法无法取得理想的翻译效果。因此，如何基于深度学习强大的特征表示能力、采用半监督或弱监督的方法解决小数据机器翻译问题也将会成为未来的一个热点研究方向；

非规范文本的机器翻译：互联网是产生语言文本最多的地方，而互联网上使用的语言文本大多具有口语化、社交化等诸多新的特征，弱规范甚至不规范的现象比较严重，而目前的机器翻译系统几乎都是面向规范文本开发的，处理非规范语言现象的能力较弱。因此，如何提高非规范文本的处理能力和翻译效果也必将成为大家关注的一个研究方向；

篇章级机器翻译：无论是统计机器翻译还是神经网络机器翻译方法，都是以句子为基本翻译单位进行的，忽略了指代消解、省略和译文句子之间衔接性和连贯性等深层次的语义表达问题。如何以更大粒度的语言单位（如段落甚至篇章）为翻译单位或上下文背景，对译文的篇章信息进行建模，或将成为改善译文质量的一个值得关注的研究方向；

融合离散符号表示与连续向量表示的机器翻译：建立于离散符号表示的统计翻译方法与建立于连续向量表示的神经翻译方法都各有其优势，同时各有弊端，那么，如何设计新的方法融合两者的优势，将是未来研究的一个重要方向。

纵观 20 余年来机器翻译研发的趋势和现状，我们有理由相信，随着机器学习、语义分析和篇章理解等相关技术的快速进展，这一人工智能中最具挑战的问题将在可预见的未来得到相当程度的解决，机器翻译系统的产业化应用前景将更加广阔。

第十五章 社会媒体处理研究进展、现状及趋势

1.任务定义、目标和研究意义

社会媒体处理（Social Media Processing, SMP）是从社会媒体数据中挖掘、分析和表示有价值信息的过程。

简单来讲，社会媒体处理研究的目标就是通过挖掘社会媒体中用户生成内容和社交关系网络，来衡量用户之间的相互作用，进而发现这其中蕴含的特定模式来更好地理解人类行为特点。

社会媒体打破了现实世界与虚拟世界之间的边界，使得可以从海量的社会媒体数据中挖掘人类行为模式，进而对人类个体及群体进行全面的剖析和理解，这在社会媒体出现前是根本无法完成的任务。因此，社会媒体处理在现实生活和工作中具有重要的社会需求。从理论上讲，社会媒体处理涉及计算机科学、社会学、传播学、管理学、经济学、语言学、心理学等多个学科，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，揭示人类行为模式的本质，又有助于促进传统自然语言处理任务的快速发展。从应用上讲，无论是社会大众、政府企业还是国家机构，都迫切需要社会媒体处理技术。特别是在“互联网+”时代，国家倡导跨界融合、开放共生、以人为本、连接一切的互联网生态环境。通过对微博、微信、维基、社会标注系统等以人为本的社会媒体中的大数据进行分析处理，可以为用户获取新闻时事、人际交往、自我表达、社会分享及社会参与提供便利，并使人们的生活模式、产业运行模式与社会发展形态得到了全方位的提升，进而为社会管理提供重要的支撑作用。“互联网+” 的本质是传统产业在线化、数据化。随着互联网上用户产生的数据信息急剧增加与信息传播方式的巨大改变，使得企业与消费者间的关系趋向平等、互动和相互影响，因此亟需构建新型的营销策略。而企业通过对社会媒体大数据分析处理可以为用户提供专属性的个性化产品和服务，进而为企业竞争情报监测、品牌口碑监测等企业营销管理提供重要的技术保障。

当然，社会媒体处理技术还不完美，它仍面临很多具体问题和困难。本文对社会媒体处理研究的主要内容、面临的科学问题和主要困难，以及当前采用的主要技术、现状和未来发展的趋势，做简要介绍。

2.研究内容和关键科学问题

按照社会媒体的定义，社会媒体的主要呈现形式是用户创造和传播的信息流。因此，社会媒体处理研究的主要内容是挖掘和利用社会媒体产生的大量信息。概括来讲，社会媒体处理研究主要包含社会媒体信息的挖掘以及基于社会媒体的应用研究。其中，社会媒体的信息挖掘按照信息类型，又可分为客观信息挖掘和主观信息挖掘。社会媒体处理的主要研究内容和涉及到的关键科学问题如下：

2.1 社会媒体客观信息的挖掘

社会媒体客观信息挖掘主要从用户的属性信息、网络结构信息和行为信息等方面出发，对应的三个具体的研究内容及相关关键科学问题如下：

2.1.1 用户画像

用户画像（User Profile）是指利用社会媒体中用户的文本、图片、社交行为等数据构建用户的未知属性信息或未知标签。社会媒体拥有海量的用户信息，从中发掘用户的属性及兴趣等信息，是为用户提供更加个性化服务的基础。由于用户的隐私保护等因素，社交网络中存在用户个人信息不全或虚假的情况，这使得挖掘用户的属性信息变得愈发困难。

2.1.2 社交圈识别

除了用户属性，社会媒体还蕴含着用户之间的社交网络信息。社会媒体的出现使得对用户社区的研究可以更加微观和细致，因此社会媒体中局部社区结构的挖掘和用户关系分析成为社会媒体处理的重要科学问题。

2.1.3 信息传播分析

社会媒体的广泛流行不仅导致信息呈爆炸式增长，而且为互联网信息传播的方式带来了巨大的变革。人与人之间的互联、人与信息之间的互联高度融合，人人参与到信息的产生与传播过程，这种传播方式使得一条信息能够在短时间内传播到数百万计的用户。然而，大量的用户生成信息也带来了诸如信息过载、虚假信息泛滥等问题，对社会媒体信息传播的研究为解决这些问题提供了可能。

2.2 社会媒体主观信息的挖掘

社会媒体中的主观信息主要包括人们的观点、情感、意图、建议等。观点是指人们对某一人物或事物表现出支持、中立或反对的态度。情感是指人们在社会媒体中流露出的喜、怒、悲、恐、惊等情绪状态。不同的观点通常伴随着不同的情感。意图是指人们表现出想要做某事，是一种主观意愿。对应以上信息，社会媒体处理主要有如下两个具体的研究内容及相应的关键科学问题：

2.2.1 情感分析

文本情感分析旨在从无结构的文本中自动地抽取、分析和整理带有主观色彩的文本，对主观性文本进行归纳和整理。社会媒体中充斥着大量的带有主观情感的文本，正确的识别出社会媒体中用户的这些情感能够辅助政府决策，掌握产品口碑，同时为生产者和消费者之间的连接提供技术支持。

2.2.2 消费意图挖掘

消费意图在社会媒体用户意图中占有非常大的比重。用户在社会媒体中表达出对产品或服务的购买意愿，我们将其称之为消费意图。通过消费意图挖掘可以深入理解人类消费行为并进行精准的个性化产品推荐。因此，消费意图挖掘研究是社会媒体营销的基础，对准确预测用户消费行为、评估市场走势和精准广告投放等方面具有重要意义。

2.3 基于社会媒体的应用研究

随着社会媒体信息挖掘研究的日渐成熟，基于社会媒体的应用研究备受瞩目。推荐和预测是基于社会媒体的两大关键应用。

2.3.1 基于社会媒体的推荐

推荐是当前解决信息过载和实现个性化服务的重要技术。社会媒体涵盖了丰富的用户个人信息，包括用户发布内容，用户属性以及用户的社交网络等。在学习用户历史行为的同时，如何融合用户的发布内容、属性信息和社交关系等多维度信息帮助提升推荐算法的准确性是基于社会媒体的推荐研究的关键科学问题。

2.3.2 基于社会媒体的预测

基于社会媒体的预测是指通过对社会媒体数据的挖掘与分析，聚集大众的群体智慧，运用科学的知识、方法和手段，对事物未来发展趋势和状态做出科学的估计和评价。根据预测方法的不同可以进一步划分为基于相关关系的预测和基于因果关系的预测。基于相关关系的预测是指通过找到一个现象的良好的关联物来帮助了解现在和预测未来。对于某些事件来说没有过多的相关性数据可用时，因果是最有效的预测指南。例如稀有事件预测、新闻事件预测等。因此，当我们对于某一事物预测不准或者认识不准时，一个合理的做法是分析因果并使用因果进行再认识。

3. 技术方法和研究现状

社会媒体处理发展至今，其技术方法融合了计算机科学、社会学、传播学、经济学等多个交叉学科的研究方法。目前，社会媒体处理研究中涉及的技术方法包括用户画像，社会网络挖掘，社会媒体传播，社会媒体情感分析，消费意图挖掘，以及基于社会媒体的预测等。

3.1 用户画像

用户画像根据建模方法的不同，可以分为三类，即基于内容的用户建模，基于关系的用户建模以及基于行为的用户建模。

3.1.1 基于内容的用户建模

用户生成内容是社会媒体的主要信息载体之一，包括用户发布的内容、转载的文章、关注的问题等等。1958 年 John L Fischer 对社会中不同人群对词汇形态学和音韵学使用情况和相关关系进行了研究，证明了不同人群在用词上存在差异。在此理论基础上，早期用户画像研究主要基于用户文本内容进行建模研究。随着社交网络的兴起，基于短文本的用户画像识别得到了众多学者的关注。其中很大一部分基于短文本的用户画像研究工作将各个属性识别转化为分类问题（如男与女、青少年与中老年、南方人与北方人）并通过不同的分类器进行分类，也有部分工作将属性识别视为回归问题（如年龄）。研究的用户属性包含生物人属性（如性别、年龄、地域等）和个性化属性（如政治倾向、学历、个人标签等），其中，性别和年龄尤其受到关注。Delip Rao 等人基于 Twitter 上的文本信息对性别、年龄、地域、政治倾向四个属性进行识别，其研究结果表明，用户的词汇使用、标点符号使用、表情符号

使用对识别准确率有较大影响。

3.1.2 基于关系的用户建模

社会媒体的另一项主要信息载体是用户关系网络，在典型的社会媒体中，用户之间通过关注关系构成一张有向图，不同标签的用户在关注人、关注数量等方面体现出不同的变化。同时，用户之间的直接关系（如评论、提及）等也对用户画像任务起到了重要作用。Faiyaz Al Zamal 等人通过用户的朋友信息对性别、年龄和政治倾向性进行了识别，重点探讨了不同的朋友信息（如朋友数量、关系亲疏）对识别结果的影响。Aron Culotta 等人基于用户对 Twitter 上 150 个网站的公共主页的关注关系，利用逻辑回归的方法，对性别、年龄、收入、教育程度等属性进行了识别，并取得了很好的效果。

3.1.3 基于行为的用户建模

用户在社会媒体上的行为包括登录、转发、评论、点赞以及搜索等。不同标签的用户在日常活动上呈现出明显的差异性。Bi 等人利用用户的搜索记录对用户进行了用户画像建模，他们利用 Facebook 上的用户属性信息构建标签化数据，并通过 Open Directory Project 项目中的类别信息综合了 Facebook 上的点赞信息和在必应上的搜索记录信息，最终将通过 Facebook 数据训练的模型直接用于必应搜索记录数据中。

3.2 社会网络挖掘

社会网络挖掘中一项重要任务是对社会媒体用户的社交圈进行挖掘。社交圈在计算机科学领域是一个相对主观的概念。Facebook 和 Twitter 等很多社交网络中为用户提供了好友分组功能。这些分组与社交圈的概念很接近，但是由于很多用户把很多单向关注的好友(如明星、媒体等)进行了分组,所以这种用户分组跟社交圈还存在一定区别,社会网络中用户社交圈识别的相关研究,主要分为基于网络结构特征的识别和基于用户分享内容特征的识别两类。

3.2.1 基于网络结构特征识别

同一个社交圈内成员之间具有很强的拓扑结构同质性。也就是说同一社交圈内成员之间的关联度越高，用户之间的交互也更加密切。Zhao 根据关注列表分析了 Twitter 中用户关注网络的变化。Ferrara 用种子扩展的方式发现新浪微博中的用户社区。Xu 从隐私保护的角度在局部网络中发现用户所在社区。Hu 根据用户交互频度发现用户所在社区。在 Twitter 中用户可以根据类别不同对好友自定义列表，一些文献讨论了为列表自动推荐好友的算法。Ferrara 用标签扩散法(Label Propagation Algorithm)和快速网络社区识别法(Fast Network Community Algorithm)对 Facebook 中的用户社区进行识别。

3.2.2 基于用户分享内容的识别

同一社区的成员往往有着共同的特点和爱好，所以用户发布、分享的内容主题也是用户所在社交圈的重要特性之一。Enoki 结合用户网络结构相似矩阵和内容相似矩阵,对 Twitter 中的社区进行识别。Xiong 把新浪微博中的用户关系和内容融合在一起，从而构建一个新的用户网络,再在新的网络中进行用户社区识别。

3.3 社会媒体传播

由于社会媒体传播分析的最终目的是对信息传播进行预测。预测的方法整体可概括为以信息为中心、以用户为中心和以信息和用户为中心这三个方面的研究。

3.3.1 以信息为中心的预测研究

忽略个体的传播行为，只关注信息的整体传播趋势，如传播范围、传播周期等特性，从而为舆情监控提供了可能，其主要任务是对信息流行度进行预测。信息流行度是指信息在社会网络中最终的传播过程和结果，通常与信息的形式和传播方式有关。比如，视频分享网站上，视频信息的流行度通常由浏览数、分享数来衡量；新闻平台中的信息流行度则由新闻评论数来表示。从模型角度看，信息流行度预测的研究方法以基于传染病模型和分类或回归模型的预测方法为主。

传染病模型是对疾病在人群中的表现和分布形式进行数学建模的方法，由于信息传播与传染病传播类似，因此被应用于博客、微博等信息流行度预测研究之中。典型的传染病模型包括 SI （susceptible-infected)， SIS （susceptible-infected-susceptible)， SIR （susceptible-infected-recovered）以及 SIRS （susceptible-infected-recovered-susceptible)。其主要思想是：将人群中的个体按照其所处的状态进行分类，关注每类状态下个体数量比例的演化，处在每个状态的个体比例通过微分方程求解。

分类或回归模型主要从分析影响信息传播关键因素的角度出发，利用统计机器学习模型对信息流行度进行预测。将待预测的信息表示成一组基于影响因素的特征，把信息流行度预测问题转化为分类或者回归问题，通过大量的已知数据训练出机器学习模型对未知信息进行预测。

3.3.2 以用户为中心的预测研究

以用户的兴趣和行为建模为基础，分析用户是否会参与某信息的传播，从而为个性化推荐提供了可能，其主要任务是用户传播行为预测，预测对象是用户。用户传播行为预测是指通过一定的手段学习已知用户的兴趣和行为规律，从而对未知的用户传播行为进行预测。按照预测基本假设的不同，用户传播行为预测方法可分为基于用户过往行为的预测、基于用户文本兴趣的预测、基于用户所受群体影响的预测以及基于混合特征学习的预测。主要使用的模型包括：协同过滤模型、主题模型、因子图模型以及分类模型等。

基于用户过往行为的预测方法假设用户的传播行为反映用户的兴趣，依据用户在预测时间点前的过往行为预测用户未来的行为。这类方法主要使用的模型是协同过滤模型。

基于用户文本兴趣的预测方法假设用户对某信息的传播行为主要源于用户对社会媒体文本内容的兴趣，通过用户的过往文本信息对用户进行文本建模，从而预测用户对信息的传播行为。这类方法在用户拥有一定数量的社会媒体文本时效果较好；但对于文本内容较少的用户，很难学到其真正感兴趣的内容。

基于用户所受群体影响的预测方法假设用户传播行为的产生源于所受群体的影响，包括信息发布者的影响和其他信息转发者的影响。这类方法中较多使用因子图模型，除用户之间的相互影响外，因子图模型还可建模其他影响因素，如内容流行度的影响等。

基于混合特征学习的预测方法将传播行为预测视为二元分类问题，认为用户传播行为是多种因素作用的结果。分析影响用户传播行为的因素并将其表示为特征，然后选择适当的分类器训练分类模型。这种方法最为简单、直观，模型解释性弱，依赖于特征的选择与组合。

3.3.3 以信息和用户为中心的预测研究

其主要任务是通过分析个体的传播行为或传播概率预测信息的传播路径，预测对象是信息和用户。核心思想是：已知用户之间的网络结构，根据节点之间的关系推测信息传播的可能性。主要预测方法包括独立级联模型、线性阈值模型、分类模型以及博弈论模型。基于独立级联模型的预测方法假设信息传播的过程是依靠相邻节点之间的相互影响，个体的传播行为取决于某个相邻节点对它的激活概率。基于线性阈值模型的预测方法假设个体的传播行为取决于所有相邻节点对它的影响是否超过激活阈值。基于分类模型的预测方法将信息传播路径预测转化成用户传播行为的预测问题。基于博弈论的预测方法将每个用户视为一个智能体，假设每个用户在接收到某信息时会进行利益的博弈，采取令自己获利最多的策略。

3.4 社会媒体情感分析

在众多的大数据形式中，社会媒体数据，如微博和微信数据，是很好的一种洞察民情，观测大众行为的数据形式。微博大数据可以挖掘出民众普遍关注的话题类型、暴露出民众的整体情感趋势，供舆情部门监测。此外对事件和情感的深层分析和透视，理解事件的起因和发展，并基于此来指导社交治理方案的制定也是目前研究的重点和难点。

3.4.1 基于社会媒体的舆情监控

目前国内外已经有多项借助微博或 Twitter 来进行浅层社交治理和分析的技术和系统。Zhao 等人构建了一个名为 MoodLens 的中文微博情感分析系统，将微博的情感分为愤怒、厌恶、高兴和低落四类，进行异常或突发事件的监测。Wang 等人构建了一个实时的预测 2012 年美国大选结果的系统，该系统通过统计在 Twitter 上民众对于四位候选人的情感分布来进行结果预测。从这些研究中，我们发现在社会媒体上公众的互动具有情感演变和漂移等特点，随时间、环境的变化，公众立场将发生不断变化，兴趣点也在不断的演化。康奈尔大学的研究人员 Scott A. Golder 和 Michael W. Macy，在《Science》杂志，对全球 84 个国家超过 200 万个 Twitter 用户所发布的 5 亿多条内容进行了基于关键字的分析，得出人们在一天中的早餐时段会表现出最正面积极的情感；午餐过后人们的心情逐渐跌到谷底；临睡之前人们的情感又开始急剧回升。Twitter 成为人类情感的脉搏，人们的情感随着时间的改变而发生变化，很有必要通过细粒度的社会媒体处理分析来追溯产生情感迁移的原因。

3.4.2 情感分析的深层透视

目前大部分有代表性的系统和算法均是围绕微博或 Twitter 大数据中焦点事件抽取和情感分析这样两大项任务进行的。然而，传统的系统和研究往往只关注民众关心的焦点事件是什么，情感走向是什么，所分析出的结果可以为相关部门提供一定的预警信号。对于社会治理而言，相关部门更关心的是为何某一事件的发生会产生异常情感，什么样的人群会导致某些情感的产生等深入的原因剖析，基于此来指导社会治理方案的制定。例如：看到民众对于“汶川地震”事件的情感分布后，相关部门更想知道为何会有人喜悦，为何会有人愤怒等异常情感的形成原因。我们称这项任务为群体情感深度挖掘。此外，很多系统可以绘制出随着时间的变化各种情感的走向图，民众和相关部门更关心趋势图中的大趋势拐点产生的原因。我们称这项任务为群体情感演化分析。

3.5 消费意图挖掘

消费意图可以划分成“显式消费意图”和“隐式消费意图”两大类。显式消费意图是指在用户所发布微博文本当中显式的指出想要购买的商品。而隐式消费意图是指用户不会在所发布微博文本当中显式的指出想要购买的商品，需要阅读者通过对文本语义的理解和进一步推理才能够猜测到用户想要购买的商品，例如根据“孩子缺钙”，可以推理出用户可能要买钙片类产品为孩子补钙。

对于显式消费意图，很多学者通过模式匹配的方法进行识别。例如，在识别观影意图时，基于依存句法分析结果构建模板，识别带有某部电影观影意图的微博，其准确率可以达到80%左右。而隐式消费意图的识别则难得多，难点包括：（1）如何理解用户的语义文本，进而理解用户的消费意图；（2）用户消费意图挖掘任务是领域相关的，因此构建的模型需要具有领域自适应能力。

为了解决以上难点，哈工大提出了基于领域自适应卷积神经网络的社会媒体用户消费意图挖掘方法，能够捕捉词汇级语义特征、整合句子级语义特征，并可进行领域迁移。

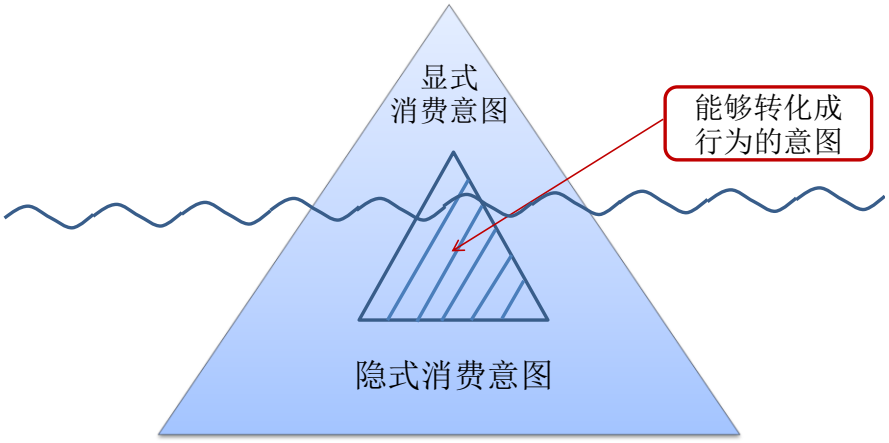


图 1. 消费意图研究层次

消费意图毕竟还只是停留在个人意愿层面，有多少用户会真正的将消费意图转化成消费行为，这是更加值得关心的话题。如图 1 所示，消费意图识别的研究可以分成三个层次，即显式消费意图、隐式消费意图和能够转化成行为的意图。显式消费意图是用户消费意图这座冰山中露出水面的一角，更多的是隐式意图，而无论是显式意图，还是隐式意图，都只有一部分能够转化为购买行为。

3.6 基于社会媒体的预测

社会媒体内容为定性和定量的发现社会结构和分析行为模式提供了大量的数据，这为预测与人类息息相关的未来事件提供了依据。根据预测中采用的方法不同，可以将其分为基于相关关系的预测和基于因果关系的预测。

3.6.1 基于相关关系的预测

基于相关关系的预测通过找到一个现象的良好的关联物来帮助了解现在和预测未来。相关关系很有用，不仅仅是因为它能为我们提供新的视角，而且提供的视角都很清晰。基于社会媒体的政治选举预测是被广泛研究的一个课题，由于社会媒体的出现，一种全新的低成本的民调方法应运而生。简单来讲，关于某位候选人的社会媒体内容提及数是其支持率的一个非常有效的参照物，可以根据其高低来预测候选人支持率的变化。

3.6.2 基于因果关系的预测

基于因果关系的预测是对基于相关关系预测的重要补充。其主要研究可以分为三大部分，即因果关系抽取，由因导果和执果溯因。

（1）因果关系抽取是一个非常基础且重要的任务。因为抽取出的因果关系或因果知识可以用于预测、问答等任务中。在文本中进行因果抽取就要用到自然语言的处理技术和方法，如词性标注、句法分析、短语抽取等。对于因果关系抽取和检测任务来说，前人的工作所使用的线索可以粗略的分为基于 Lexico-syntactic 模板的方法、基于上下文词信息的方法、基于词之间关联信息的方法和基于动词和名词的语义关系信息的方法。

（2）由因导果即因果的预测逻辑。看到一个现象或者一个事件的发生，我们总想知道未来可能出现的现象或者发生的事件。对于预测未来来说，因果无疑是最有效的指南和依据。尤其是在基于相关性分析的预测失效时，分析出原因并利用原因进行预测，预测结果会更加可靠。

（3）执果溯因即因果的解释逻辑。看到一个现象的结果我们总想知道“为什么”。在自然语言文本中，我们对因果解释逻辑的诉求也是随处可见。以商业领域为例，在电商的网站上有大量用户对商品的评论信息，有些人对某款商品 A 持有积极的评价，也有一些人在纷纷吐槽商品 A。那么作为商品的生产商和销售商就一定很想知道，是否喜欢商品 A 的原因是什么？

为了分析一个时序变量是否是另一个时序变量的因果作用，Kay H. Brodersen 等人提出了一个基于贝叶斯网络的时间序列模型。通过预测出一个虚拟结果进而和真实结果进行对比来评价一个变量对另一个变量的因果作用。比如有一个网站，它在某一时刻 t 加入了一个广告，我们想知道的是，我们引入的那个广告究竟可以为我们带来了多大的点击量。

4. 总结与展望

纵观社会媒体处理研究发展的态势和技术现状，以下研究方向或问题将可能成为未来社会媒体研究的热点和难点：

消除噪声数据：在经典的数据挖掘文献中，一个成功的数据挖掘操作必须要有大规模数据预处理过程和数据去噪过程，以避免出现“垃圾数据输入，垃圾数据输出”这样的情况。由于社会媒体自身的一些特点，它包含了很大一部分噪声数据。对于社会媒体数据去噪有两个重要问题需要解决：（1）盲目地去除噪声数据会加剧大数据悖论问题，这是因为去噪的同时也会将有价值的信息过滤掉；（2）对于噪声数据的定义是复杂且相对的，这取决于数据要服务于什么样的任务。

避免数据采样有偏：社会媒体数据的偏置性体现在两个方面。一方面是人口统计学特征的偏置，例如年龄结构，2000 年，36%的美国公民在 18 到 24 岁之间，50%的公民在 25 到 34 岁之间，但是，在推特上 60%以上的用户小于 24 岁。如果我们在社会媒体中进行平均采样的话，那么采样得到的数据与真实数据一定会存在着偏置，而这种偏置对于很多任务（例如，基于社会媒体的总统大选结果预测）都会产生非常大的影响。偏置的另一方面主要体现在社会媒体草根性上，对于某一热点事件，无论是专家还是草根大众都会给出自己的观点判断，而这种观点判断更多的是基于其自身的背景知识和经验。专家与普通大众在特定领域的背景知识和经验积累是有明显差距的，因此，在数据收集过程中忽略这种差距是不正确的，而是应该充分考虑到不同人群的观点权威度或置信度，然后赋予不用的权重。

因果分析对相关分析的补充：随着大数据概念的兴起，相关关系越来越多的得到重视和使用。然而，相关关系并不能揭示现象发生的本质，尤其当预测对象在不断随着时间推移而发生变化时，相关关系可能就会失效。因此，当我们对于某一事物预测不准或者认识不准时，一个合理的做法是分析因果并使用因果进行再认识。

评价困境：数据挖掘中一个经典的模式评价方法是收集准确客观的数据用于验证。例如，一个数据集可以被分成训练集和测试集。只有训练集被用来学习，而测试集则被用来当作标准答案进行测试。然而，在社会媒体处理中往往没有一个标准的答案，例如我们很难给出用

户全部社交圈的标准答案。如何评价从社会媒体中挖掘出来的模式，给我们提出了一个看似难以逾越的挑战。另一方面，如果没有可靠的评价手段，如何能够保证从社会媒体中获取到的信息是正确的？

与社会学跨学科合作难题：计算机学科与社会学等学科进行跨学科合作的推进过程中面临学科属性、学科建制和学术市场等方面的障碍。首先，因学科属性不同，学科在研究活动的组织方式上存在重大差别，从而影响相互之间的合作。其次，还有学科建制上的障碍。按当前体制，不同学科往往分属不同的研究单位。组织归属不同，科研议程的设置、资源的配备、绩效的考核也就不同。最后是市场选择。在社会媒体大数据开发的两种取向中，社会学研究生产周期长，生产成本高，短期内却难以见到效益，因而在研究资源的获取上受到很大限制。而计算机科学，其工作更容易被市场接受，更容易走应用路线。这样一种局面，对学科能否亲密合作，坚持到底是一个严峻的考验。

纵观近年来社会媒体处理研究的趋势和现状，我们有理由相信，随着计算机科学、社会学、传播学等学科的快速进展，这一多学科交叉的问题将在可预见的未来得到相当程度的解决，社会媒体处理的产业化应用前景将更加广阔。

第十六章 语音技术研究进展、现状及趋势

语音技术包含了很广泛的内涵，涉及语音合成、语音识别、说话人识别、语音增强、语音翻译等等。本部分仅介绍语音合成、语音识别和说话人识别三个部分的研究进展、现状和未来发展趋势。

1. 任务定义、目标和研究意义

1.1 语音合成

语音合成（Speech Synthesis），也称为文语转换（Text-to-Speech, TTS,它是将任意的输入文本转换成自然流畅的语音输出。语音合成涉及到人工智能、心理学、声学、语言学、数字信号处理、计算机科学等多个学科技术，是信息处理领域中的一项前沿技术。随着计算机技术的不断提高，语音合成技术从早期的共振峰合成,逐步发展为波形拼接合成和统计参数语音合成，再发展到混合语音合成；合成语音的质量、自然度已经得到明显提高，基本能满足一些特定场合的应用需求。目前，语音合成技术在银行、医院等的信息播报系统、汽车导航系统、自动应答呼叫中心等都有广泛应用，取得了巨大的经济效益。另外，随着智能手机、MP3、PDA 等与我们生活密切相关的媒介的大量涌现，语音合成的应用也在逐渐向娱乐、语音教学、康复治疗等领域深入。可以说语音合成正在影响着人们生活的方方面面。

1.2 语音识别

语音识别 （Automatic Speech Recognition, ASR）是指利用计算机实现从语音到文字自动转换的任务。在实际应用中，语音识别通常与自然语言理解、自然语言生成和语音合成等技术结合在一起，提供一个基于语音的自然流畅的人机交互方法。

早期的语音识别技术多基于信号处理和模式识别方法。随着技术的进步，机器学习方法越来越多地应用到语音识别研究中，特别是深度学习技术，它给语音识别研究带来了深刻变革。另外，随着数据量的增加和机器计算能力的提高，语音识别越来越依赖数据资源和各种数据优化方法，这使得语音识别与大数据、高性能计算等新技术产生广泛结合。综上所述，语音识别是一门综合性应用技术，集成了包括信号处理、模式识别、机器学习、数值分析、自然语言处理、高性能计算等一系列基础学科的优秀成果，是一门跨领域、跨学科的应用型研究。

语音识别研究具有重要的科学价值和社会价值。语音信号是典型的局部稳态时间序列，研究这一信号的建模方法具有普遍意义。事实上，我们日常所见的大量信号都属于这种局部稳态信号，如视频、雷达信号、金融资产价格、经济数据等。这些信号的共同特点是在抽象的时间序列中包括大量不同层次的信息，因而可用相似的模型进行描述。历史上，语音信号的研究成果在若干领域起过重要的启发作用。例如，语音信号处理中的隐马尔科夫模型在金融分析、机械控制等领域都得到了广泛应用。近年来，深度神经网络在语音识别领域的巨大成功直接促进了各种深度学习模型在自然语言处理、图形图像处理、知识推理等众多应用领域的发展，取得了一个又一个令人惊叹的成果。

在实用价值方面，语音交互是未来人机交互的重要方式之一。随着移动电话、穿戴式设备、智能家电等可计算设备的普及，基于键盘、鼠标、触摸屏的传统交互方式变得相对不够便捷。为了解决这种困难，手势、脑波等一系列新的人机交互方式进入人们的视野。在这些新兴交互方式中，语音交互具有自然、便捷、安全和稳定等特性，是最理想的交互方式。在

语音交互技术中，语音识别是至关重要的一环：只有能“听懂”用户的输入，系统才能做出合理的反应。今天，语音识别技术已经广泛应用在移动设备、车载设备、机器人等场景，在搜索、操控、导航、休闲娱乐等众多领域发挥了越来越重要的作用。随着技术越来越成熟稳定，我们相信一个以语音作为主要交互方式的人机界面新时代将很快到来。

1.3 说话人识别

说话人识别（Speaker Recognition），或者称为声纹识别（Voiceprint Recognition, VPR），是根据语音中所包含的说话人个性信息，利用计算机以及现在的信息识别技术，自动鉴别说话人身份的一种生物特征识别技术。

简单来讲，说话人识别研究的目的就是从语音中提取具有说话人表征性的特征，建立有效的模型和系统，实现自动精准的说话人鉴别。说话人识别根据实际应用的范畴可以分为说话人辨认（Speaker Identification）和说话人确认（Speaker Verification）两种。说话人辨认是解决把待识别的人判定为其所属的若干个参考说话人中的哪一个的问题，是一个“多选一”的选择问题。而说话人确认是确定待识别者是否是所声称的参考者，识别结果只有是与否两种，是一个“一对一”的判决问题。根据实际应用场景，说话人识别还包括说话人检测（Speaker Detection）和说话人追踪（Speaker Tracking）等任务。

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。在移动互联网高速发展的现代，通过电话和移动设备进行远程的身份认证成为需要，声纹识别技术以其特有的形式，成为最方便的远程生物特征认证方式之一。声纹识别技术在军事、国防、政府、金融等多个领域都已得到广泛应用。例如，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;在刑事侦察和技术侦察中，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。

2. 研究内容和关键问题

2.1 语音合成

语音合成系统是一个典型的人工智能系统，为了合成出高质量的语音，不仅需要依赖于各种规则（包括语义学规则、词汇规则、语音学规则），还涉及到对文字内容的理解。一个典型的语音合成系统如图 1 所示。

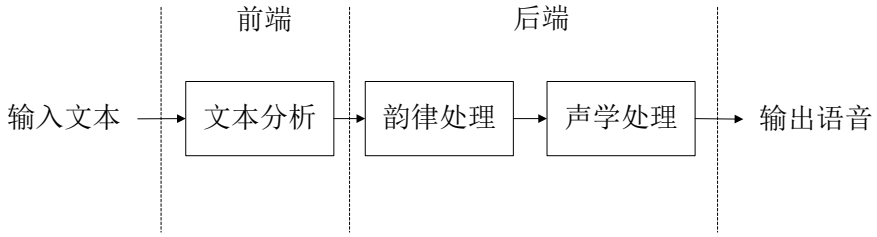


图1. 典型语音合成系统示意图

因此语音合成的相关研究内容主要包括文本分析、韵律处理和声学处理三个部分。对研究内容及相应的关键问题介绍如下：

2.1.1 文本分析

文本分析的主要任务是对输入的任意文本进行分析，输出尽可能多的语言学信息，如拼音、节奏等，为后端的语音合成器提供必要的信息。对于简单的系统而言，可能文本分析只提供拼音信息就足够了；而对于高自然度的合成系统，文本分析需要给出更详尽的语言学和语音学信息。因此，文本分析实际上是一个人工智能系统，属于自然语言理解的范畴。

对于汉语语音合成系统，文本分析的处理流程包括：文本预处理、文本规范化、自动分词、词性标注、字音转换（多音字消歧）、韵律预测等。文本预处理包括删除无效符号，断句，内码转换等。文本规范化的任务就是将文本中的这些特殊字符识别出来，并转化为一种规范化的表达。分词是指将待合成的整句以词为单位划分为单元序列，以便后续考虑词性标注、韵律边界标注等。词性标注也很重要，因为词性可能影响字或词的发音方式。字音转换的任务是将待合成的文字序列转换为对应的拼音序列，即告诉后端合成器应该读什么音。由于汉语中有多音字问题的存在，字音转换的一个关键问题就是解决多音字的消歧问题。典型的文本分析流程如图 2 所示。

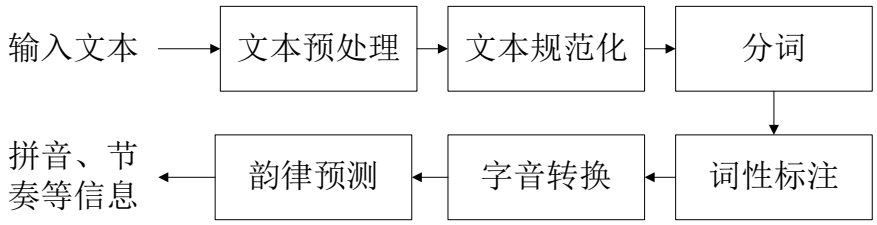


图2. 文本分析流程图

2.1.2 韵律处理

韵律处理是文本分析的目的所在。停顿、时长的预测，以及基频曲线的生成都是基于文本分析的结果。直观来讲，韵律即是实际语流中的抑扬顿挫和轻重缓急，例如重音的位置分布及其等级差异，韵律边界的位置分布及其等级差异，语调的基本骨架及其跟声调、节奏和重音的关系等等。韵律建模模块负责为待合成语音生成合适的基频曲线、音长信息、静音长度信息等。由于这些特征需要通过不止一个音段上的特征变化得以实现，通常也称之为超音段特征。韵律表现是一个很复杂的现象，对韵律的研究涉及到语音学、语言学、声学、心理学、物理学等多个领域。但是，作为语音合成系统中承上启下的模块，韵律模块实际上是语音合成系统的核心部分，极大地影响着最终合成语音的自然度。从听者的角度来看，与韵律相关的参数包括：基频、时长、停顿和能量。韵律模型就是利用文本分析的结果，来预测这四个参数。

2.1.3 声学处理

声学处理是根据文本分析和韵律处理提供的信息来生成自然语音波形。语音合成系统的合成阶段简单的可以概括为两种方法：一种是基于时域波形的拼接合成方法，声学处理模块根据韵律处理模块提供的基频、时长、能量和节奏等信息在大规模语料库中挑选最合适的语音单元，然后通过拼接算法生成自然语音波形；一种是基于语音参数的合成方法，声学处理模块的主要任务是根据韵律和文本信息的指导来得到语音参数，如谱参数、基频等，然后通过语音参数合成器来生成自然语音波形。

2.2 语音识别

语音识别研究主要包括三方面内容：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。图 3 给出一个语音识别系统的典型架构。

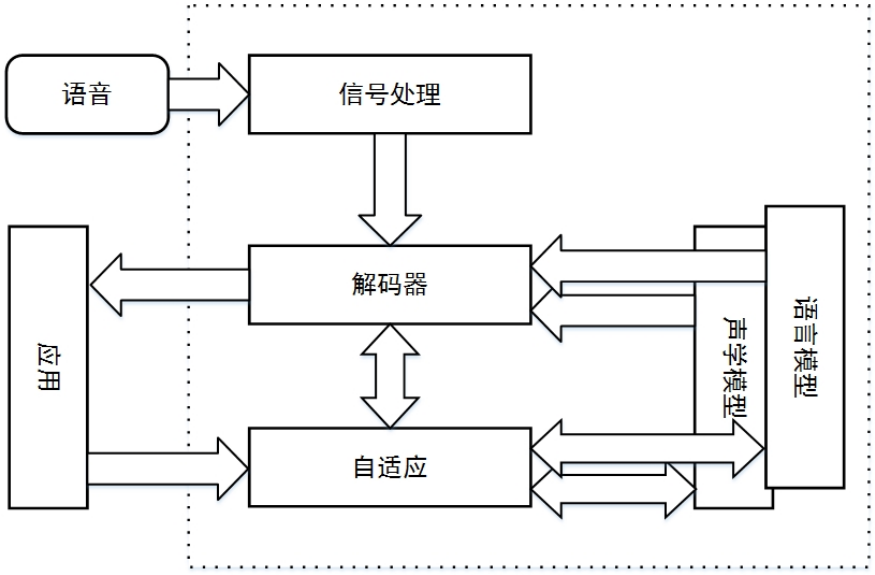


图 3. 语音识别系统结构 (Huang, X.,1996)

2.2.1 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。

2.2.2 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，特别是隐马尔科夫模型/高斯混合模型(HMM/GMM)结构。最近几年，深度神经网络(DNN)和各种异构神经网络成为声学模型的主流结构。

声学模型需要解决如下几个基本问题：

如何描述语音信号的短时平稳性；

- 如何描述语音信号在某一平稳瞬态的静态特性，即特征分布规律；
- 如何应用语法语义等高层信息；
- 如何对模型进行优化，即模型训练。

同时，在实际应用中，还需要解决众多应用问题，例如：

- 如何从一个领域快速自适应到另一个领域；
- 如何对噪音、信道等非语音内容进行补偿；
- 如何利用少量数据建模；
- 如何提高对语音内容的区分性；
- 如何利用半标注或无标注数据，等等。

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解码空间形成约束，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)，近年来基于递归神经网络（RNN）的语言模型发展很快，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：

- 如何对字母、字、词、短语、主题等多层次语言单元进行多层次建模
- 如何对应用领域进行快速自适应；
- 如何提高训练效率，特别是对神经网络模型来说，提高效率尤为重要；
- 如何有效利用大量噪声数据，等等。

2.2.3 解码

解码是利用语音模型和语言模型中积累的知识，对语音信号序列进行推理，从而得到相应语音内容的过程。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，即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。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，一般对搜索路径进行剪枝，保留最有希望的路径，即束搜索（beam search）。

对解码器的研究包括但不限于如下内容：

- 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；
- 如何实现静态解码图的动态更新，如加入新词；
- 如何利用高层语义信息；
- 如何估计解码结果的信任度；
- 如何实现多语言和混合语言解码；
- 如何对多个解码器的解码结果进行融合。

2.3 说话人识别

说话人识别主要需要解决如下三个关键问题：

2.3.1 特征提取与模式划分问题

说话人识别是一类典型的模式识别问题，因此需要解决特征提取和模型建立两个方面的主要问题。

语音信号中含有各种各样丰富的信息，在特征提取阶段，需要研究如何从语音中提取鲁棒的说话人特征来表征说话人。对于说话人识别，提取的特征需要满足：具有很高的区分不同说话人的能力；能够充分体现说话人个体间较大的差异；对说话人自身差异体现得不明显。目前绝大部分说话人识别系统采用的特征仍然是语音识别中广泛使用的声学特征，虽然这些特征可以间接的提供一定的说话人区分性信息，但是如何从原始语音中提取出仅与说话人相关的区分性信息仍然是一个悬而未决的问题。

为了对说话人特征做一致性描述，通常需要选择合适的模型结构对说话人建模，不同的说话人模型结构对应于说话人识别的不同方法。随着计算机和信号处理、人工智能等技术的不断发展，说话人模型已经从单一的模板模型发展出矢量量化模型（Vector quantization, VQ)、高斯混合模型(Gaussian mixture model, GMM)、隐马尔科夫模型(Hidden Markov model, HMM)、人工神经网络（Neural network, NN）以及它们的混合模型。联合因子分析（Joint factor analysis, JFA）和 i-vector 模型将说话人模型映射到低维子空间，可以得到更加精确的说话人描述。但是，由于这些模型大多是纯粹的模式识别方法，并没有考虑说话人识别应用本身的某些特点。因此，这些方法除了在自身的区分准确性问题之外，还面临着如何处理语音这个特定信号的问题，进而如何处理说话人识别这个特定任务的问题。

2.3.2 鲁棒性问题

- **噪音鲁棒性**。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，也会使训练与识别失配。此外，这些环境噪声通常是不可预知的，这使得说话人识别系统的识别性能具有极大的不确定性。这种情况下则需要研究对噪声更加鲁棒的特征，减少噪声影响。
- **跨信道鲁棒性**。信道失配是影响说话人识别性能的一大因素。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。
- **说话人自身状况及时变鲁棒性**。由于人的声纹特征是一种行为特征，具有易变性，因此说话人自身的语音变化也会影响说话人识别系统的性能。一个人的声音会受到其身体状况、年龄变化、说话时感情和语气的变化、语速快慢等各种不同因素的影响而产生不同的变化。这些变化很容易带来测试语音与预留语音的失配。此外，有实验表明，同一个人在使用不同语言进行声纹预留和声纹验证的时候，较之于使用同种语言时的识别性能会大幅降低。
- **短语音鲁棒性**。实际应用中，较短甚至超短语音条件下的说话人识别是必须要面对的难题，这个问题的解决将直接改善用户的体验性。此外，在很多声纹识别的应用领域，实际使用时是无法获取足够长度的测试语音的。在短测试语音情况下，目前主流的几种说话人识别系统性能的变换均十分强烈，其根本原因一方面在于短语音测试条件下，语音中所包含的说话人信息很不均衡；另一方面在于短语音条件下测试语音中所包含的说话人信息量太少，难以提供足够的区分性信息。
- **其他鲁棒性**。另外，语音的不同编码方式（编码）、两个或多个说话人对话或交流（多说话人）、训练和识别使用不同语言（跨语言）等等情况下如何保证说话人识别性能，也是备受关注的研究课题。

2.3.3 防攻击（Anti-Spoofing）问题

随着说话人识别技术的快速发展和广泛应用，针对其假冒闯入的防攻击研究开始兴起。说话人识别防攻击的目的是检测出利用一切手段，冒充真实说话人而闯入识别系统的情形，并将其拒绝。目前假冒真实说话人的主要场景分为声音模仿、语音合成、声音转换、录音重放等几个方面，说话人识别的防攻击问题严重关系到该技术在实际应用中的安全性。

防声音模仿是较早的研究方向，由于模仿的语音更多是体现在韵律特征和说话风格上，而未能从根本上改变声道的形状和特性，因此声音模仿可以骗过人耳，而对声纹识别系统的影响不大。

随着语音合成技术的迅速发展，利用少量语音进行自适应得到特定说话人语音成为可能，闯入者进而可以将合成得到的特定说话人语音用来攻击说话人识别系统。通常而言，正常语音与合成语音在声学特征之间是具有差异性的，通过检测这种差异性，可以实现针对语音合

成的闯入。

声音转换通常分为离线训练和在线转换两个过程，闯入语音和在线转换语音之间的转换函数直接影响声音转换假冒闯入的效果。

录音重放是指闯入者使用简单的录放音设备，对目标说话人的语音进行录制后重新播放，从而闯入和攻击说话人识别系统。由于录音重放无需任何语音学的技术知识，实际应用中更加容易出现。此外，重放的语音是目标说话人自身的语音，因此闯入率较高，是现阶段亟需解决的关键问题。

3．技术方法和研究现状

3.1 语音合成

3.1.1 波形拼接语音合成

波形拼接技术的基本原理就是根据文本分析的结果，从预先录制并标注好的语音库中挑选合适的基元，进行适当调整，最终拼接得到合成语音波形。受限于计算机存储能力与计算能力，早期的拼接合成方法的基元库都很小；同时为了提高存储效率，往往需要将基元参数化表示；此外，由于拼接算法本身性能的限制，造成了合成语音不连续，自然度较低。

随着计算机运算和存储能力的提升，实现基于大语料库的单元拼接合成系统成为可能。在这种方法中，基元库由以前的几 MB 扩大到几百 MB，甚至是几 GB。由于大语料库具有较高的上下文覆盖率，能够支持更为精细的选音算法，使得挑选出来的基元几乎不需要做调整就可用于拼接合成。因此相比于传统的参数合成方法，该方法合成语音在音质和自然度上都有了极大的提高。这种方法既保证了语音的音质，又提高了合成语音的自然度。因此基于大语料库的单元拼接系统得到了十分广泛的应用。

基于大语料库的波形拼接语音合成技术的主要思想就是将在大量自然语流中的丰富语音单元按照一定的规则拼接，进而得到高自然度的语音。这种方法在训练阶段，从原始语料库中提取一定的参数，用机器学习算法进行声学建模，得到训练的声学模型，作为目标基元参数预测的模型；在合成阶段，经过文本分析，得到待合成语句的特征，通过基元预选，从音库中选取部分基元作为候选基元，根据训练得到的声学模型，用文本分析后提取的参数，进行目标基元的声学参数预测，得到目标基元的声学参数，然后计算候选基元与目标基元的目标代价与相邻基元间的拼接代价，最后通过动态搜索算法，选择出最佳基元，拼接合成出语音。

尽管拼接合成方法在应用上获得了很大的成功，但它依旧存在着一些不足：

（1） 稳定性仍然不够。虽然音库很大，但远不能覆盖任意文本的上下文。基元库中没有非常匹配的基元，以及拼接点不连续的情况还是有可能发生；

（2） 音库构建较为复杂。在音库上进行音段和韵律的标注需要耗费很大的人力物力。对于每个音库，基元选取算法中的主要参数都需要重新人工校对；

（3） 难以改变发音特征。针对拼接基元的调整，往往造成合成语音音质的严重下降；

（4） 应用的局限性。主要应用在数字串合成、新闻体合成等有限领域，任意文本合成还存在一定问题。

3.1.2 统计参数语音合成

由于波形拼接语音合成存在着一些固有的缺陷，这些缺陷限制了其在多风格语音合成方面的应用。在这种情况下，国内外学者逐渐将关注点转向了统计参数语音合成。统计参数语音合成包含了两层概念：首先，它采用统计机器学习的方法对语料库的声学参数进行建模；其次，它采用声码器对模型预测出的声学参数进行参数合成。因此从本质上来说，它也是一

种参数合成方法。

统计参数语音合成的方法在最初提出之时，其合成语音的质量与大语料库的拼接系统相比有着较大的差距，因此并没有引起广泛的兴趣和重视。主要的原因就在于分析合成的音质不高。不过，随着一些学者在这一领域持续不断的研究和探索，在影响合成语音质量的几个关键因素——声码器、模型的精确度、以及参数生成方法方面都有了较大的进步，合成语音的质量也有了较大的提高。在基于统计声学建模的语音合成思想中，基于 HMM 的参数合成方法在 20 世纪末和 21 世纪初得到了充分的发展。尤其是 HTS 开源工具的推出，大大推进了各种语言的统计参数语音合成的发展。基于 HTS 的统计参数语音合成系统具有以下优点：（1）合成语音稳定平滑；（2）存储空间需求小。只需要存储统计模型而不是波形文件；（3）整体语言无关，仅在上下文标注以及问题集与语言有关；（4）容易改变发音风格以及情感表达。可应用在自适应合成等领域。

近几年来，随着深度学习研究的热潮，DNN 在语音识别，图像压缩等领域取得了突破性进展。在语音合成领域，凌震华等人提出用 HMM-DBN 来替代 HMM-GMM 来进行建模，从而提高建模精度，增强合成语音的音质。Heiga Zen 等人提出了基于深度神经网络（DNN）的统计参数合成方法，直接通过一个深层神经网络来预测声学参数，克服了 HTS 训练中决策树聚类环节中模型精度降低的缺陷，进一步提高了合成语音的质量。可以说，基于 DNN 的统计参数合成方法，是当前统计参数合成方法发展的一个主流方向。深层神经网络模型可以采用不同的拓扑结构，如深层置信神经网络、长短时记忆递归神经网络等，后者可以充分利用整个序列的信息生成目标参数，在序列生成问题上具有明显优势。

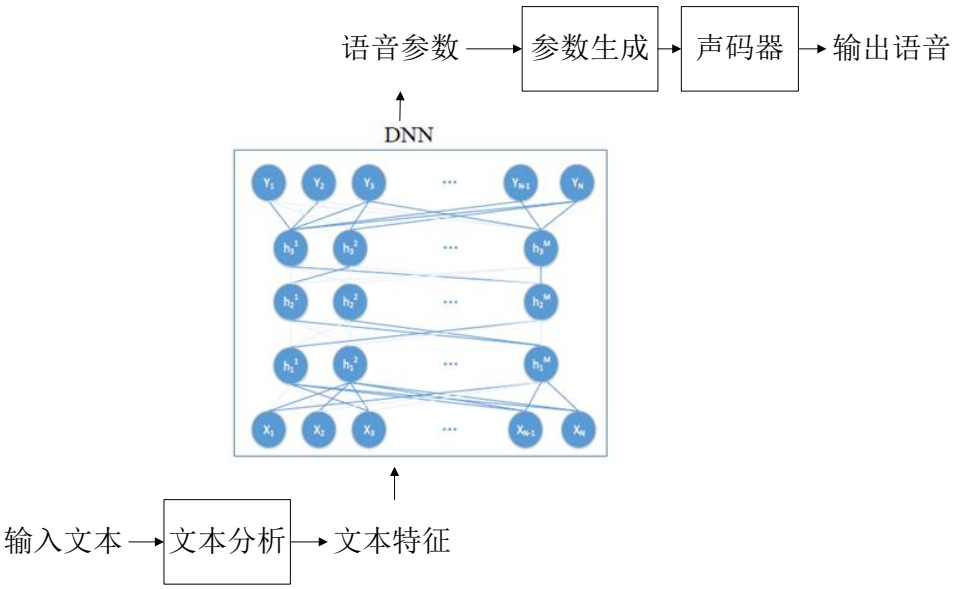


图4. 基于DNN的参数语音合成方法

尽管当前统计参数语音合成是目前语音合成研究的的主流方法，但它仍存在着许多不足之处，主要表现在以下三个方面：

- 合成语音的音质不够高，听感上时常会存在蜂鸣声；
- 合成语音的清晰度不够高，语音分析合成模型不够理想，对语音的建模不够精确；
- 合成语音的自然度不够高，生成的谱参数轨迹和基频轨迹仍存在过平滑问题，合成语音在听感上显得较为平淡。
- 而造成合成音质不高的主要原因在于：
- 声码器结构过于简单；
- 建模精度不够；
- 过平滑问题没有很好解决。

这些弱点限制了统计参数语音合成的进一步提高，也让一些研究人员重新将目光投向了波形拼接合成，在波形拼接语音合成系统中融入了统计参数语音合成的声学模型。

3.1.3. 混合语音合成

波形拼接语音合成使用了自然语音波形，可以合成出高自然度的语音，但是对于不同领域文本合成效果的稳定性不强，很难胜任任意文本合成的需求；而统计参数语音合成可以输出稳定流畅的语音，但因为参数合成器本身的缺陷，以及参数建模和生成时的平均效应，使得合成语音的音质较自然语音还有很大的差距。为了融合二者的优势，研究者们提出了基于HMM的混合语音合成方法。这种合成方法利用统计参数模型来进行候选基元的选取，融合了参数建模和选音拼接合成两种方法，进一步提高了合成效果。基于HMM生成参数轨迹跟踪的拼接合成系统框图如下图所示。在语音合成过程中，首先根据句子HMM决策用参数生成算法生成用于指导选音的参数轨迹，然后根据参数轨迹跟踪的选音准则构建目标代价与拼接代价，进一步地用动态搜索算法搜索出代价最小路径，最后将最小路径上的基元进行拼接合成出语音。

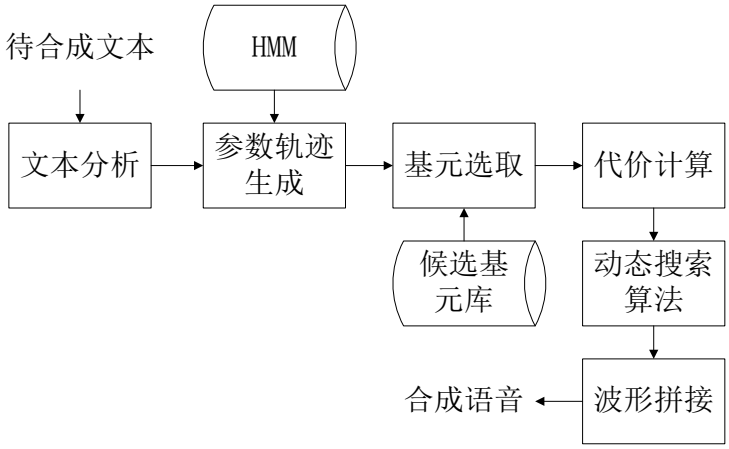


图5. 基于HMM的混合语音合成方法

随着计算机运算和存储性能的进一步发展，基于互联网的云计算平台开始应用和普及，这大大促进了混合语音合成的发展与应用。国内外大学和研究机构纷纷提出了自己的融合思路，并建立了自己的混合语音合成系统。在建库单元的选择上，不同研究人员采用了不同大小的建库单元。一般来说，小的建库单元会导致搜索空间的急剧增长，因此在实际使用的实时系统中很少采用帧大小或状态大小的单元进行建库，大多采用音素或音节作为建模单元。

在选音准则的选取上，大致有以下三种选音准则：

- 采用参数生成算法生成的谱参数、基频参数和时长参数来作为选音合成的目标；
- 采用候选基元相对于指导模型的HMM似然值来作为候选基元的选音代价；
- 采用代表候选基元的HMM模型与指导模型间的KLD距离作为候选基元的目标代价。

深度学习方法具有较强的特征学习能力，它在统计参数语音合成中已经得到了成功的应用，一些学者开始尝试将深度学习方法用于混合语音合成。这种方法在训练阶段，采用深层神经网络对基频、谱等声学参数进行建模，作为目标模型指导后续的选音。在合成阶段，首先经过文本分析得到待合成语音的相关文本特征，利用上述文本特征采用数据驱动的方法对各个基元的时长进行预测，得到不同基元的目标时长；在此基础上通过深层神经网络模型对声学参数进行预测，预测出作为目标基元的声学参数，然后通过库中基元的代价计算，确定最优的基元，进行拼接合成，最终得到合成语音。

虽然混合语音合成融合了波形拼接语音合成和参数语音合成的优势，但是在具体应用过程中仍然存在一定局限性：这种方法需要大语料库作为支撑，因此难以在嵌入式终端设备上应用；虽然基于DNN的语音合成模型可以提高指导选音的可靠性，但是相邻基元拼接处仍然存在不连续问题，从而影响合成语音的自然度。

3.1.4 技术现状

根据上述阐述可以看出，基于波形拼接的语音合成系统在大音库支撑下可以合成出高自然度的语音，但是对于不同领域文本合成效果的稳定性不强；基于统计参数的语音合成系统可以在不同嵌入式设备上得到应用，但由于参数建模时的误差，使得合成语音较为平淡，跟自然语音还有一定差距。混合语音合成方法融合了两者的优势，但在具体应用过程中仍存在着一定的局限性。虽然语音合成技术取得了飞速发展，合成语音的自然度和可懂度逐年提升，但距离任意文本、多表现力语音合成的目标仍有差距，很多理论和技术问题仍有待于更深入的研究和探索；随着人工智能技术的不断发展，语音合成的性能也将得到显著改善。

3.2 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、机器学习、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。

3.2.1 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，对不同词保留若干各自的样本，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标注为该语音信号的发音。这一方法有两个主要问题：（1）不能有效描述语音信号在时序上的不确定性，即短时平稳属性；（2）不能有效描述语音信号在发音特征上的不确定性，即不同条件下同一发音的不确定性。为解决上述因难，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（词或音素）包括若干状态，同一状态内部的发音特性保持相对稳定，不同状态间的转移具有随机性；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，其中 HMM 用来描述短时平稳的动态性，GMM 用来描述 HMM 每一状态内部的发音特征。

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，如结合上下文信息的动态贝叶斯方法、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产生做好了准备。

3.2.2 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，Mohamed 等在 NIPS workshop 上发表的“Deep Belief Networks for phone recognition”报告了基于 DNN 的声学模型在 TIMIT 数据集上得到了 23%的错误率，远好于其它复杂模型。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。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。

2013 年以前，DNN 是语音识别中应用最广泛的深度模型。随着研究的深入，研究者对

DNN 声学模型特性的理解也越来越全面。首先，人们发现 DNN 具有很强的特征提取能力，可以从频谱甚至时域信号中直接学习语音特征。这种纯数据驱动得到的特征在很多识别任务上远好于基于听觉感知特性设计的特征（如 MFCC 和 PLP）。第二，人们发现 DNN 具有强大的环境学习能力，可以对多种噪音、口音条件下的信号进行统一学习，极大提高了系统鲁棒性。第三，人们发现 DNN 非常适合多任务学习和转移学习，利用一种语言的数据训练出的 DNN，可以直接用到另外一种语言上做为特征提取模型。

DNN 的成功激励研究者探索更有效的深度模型，其中具有重要意义的是卷积神经网络（CNN）和循环神经网络（RNN）。这两种网络在深度学习提出之前已经被研究多年，在深度学习框架下取得了更好的效果。（更详细的介绍可参见《语言表示与深度学习研究进展、现状及趋势》一章。）

RNN 更深远的影响是对 HMM 模型统治地位的冲击。同 HMM 一样，RNN 模型是一种时序模型，通过累积历史信息进入不同的状态，进而改变模型输出特性。和 HMM 的离散状态结构不同的是，RNN 是一种连续状态模型，因而适合描述语音信号从起始到结束的动态发展过程。因此，利用 RNN 代替 HMM，用连续状态序列代替离散状态序列，进而把语音识别的所有模块统一成神经网络模型，是件非常吸引人的事。研究者曾做过一些这方面的探索，但直到 2014 年端对端训练方法出现以后，这一思路才最终确定下来。以 CTC 准则为目标的端对端训练方法不再依赖一个初始 GMM 模型对信号和标注进行逐帧对齐，而是考虑所有可能的路径来计算损失函数，因而有望得到更精确的模型。特别重要的是，基于 RNN 结构，音素内部的状态变化不再用 HMM 来描述，而是依赖 RNN/LSTM 内部的状态累积。这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。

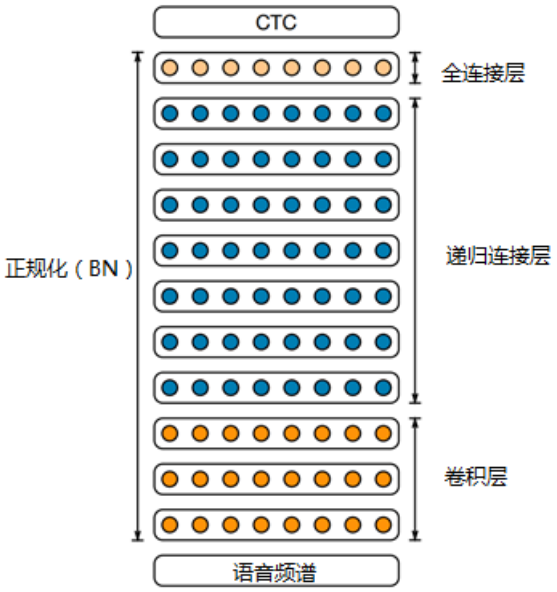


图 6. 当前语音识别系统中的声学模型结构（DeepSpeech2）

图 6 是当前语音识别系统所采用的一种典型的神经网络模型结构。该结构从频谱开始，通过 9 层 CNN 学习发音特征，通过 7 层 RNN 学习信号的静态和动态特性，最后通过 1 层全连接网络输出音素（或其它语音单元）的后验概率。RNN 层可采用 GRU 或 LSTM 结构，并可采用双向结构。训练时以 CTC 为目标，同时加入 BN 等控制梯度的方法，保证训练的收敛。

3.2.3 语言模型和解码器

前述内容主要是声学模型上的进展。相对而言，语言模型和解码器并没有发生太大变化。在语言模型方面，绝大部分系统依然基于传统的 n-gram 模型加上各种平滑算法。近年来，基于深度神经网络的语言模型(NNLM)取得很大进展，但 NNLM 不论训练和推理都显著慢于 n-gram，特别是应用到一次解码中，还需要较多的工程化工作。这使得 NNLM 还不能取代传

统 n-gram 成为主流的语音模型结构。另一方面，随着训练速度、新词处理、应用框架等问题的解决，NNLM 应该很快会取代 n-gram 成为主流，甚至成为端到端网络的一部分。

在解码器方面，绝大多数系统依然采用基于 FST 的静态解码方法。这一方法预先将 LM、词表、决策树、HMM 模型等各层次知识统一表达成 FST，并将这些 FST 编译成一个“端到端”的解码图，其中输入为音素状态，输出为词。在 CTC（端到端）系统中，解码图仅需包括 LM 和词表，因而极大简化了构图流程。基于这一解码图，应用动态规划算法（如 Viterbi）即可实现解码。这一静态解码方法不需考虑众多各异化的资源，极大简化了解码器的结构和工作流程，而且可以对解码图进行确定化、最小化等离线优化，提高解码效率。这一方法的缺点是编译后的解码图难以进行动态更新，如增加新词等。研究者提出了嵌套子图算法和相似对方法等技术，可以部分解决这一问题。

3.3 说话人识别

说话人识别的研究起始于20世纪30年代。早期的研究工作主要集中于人耳的听辩实验，研究和探讨通过声音识别说话人的可能性。进入 20 世纪下半叶，随着生物信息和计算机信息技术的发展，通过计算机进行自动的声纹识别成为可能。之后，说话人识别方法主要是基于模式匹配和概率统计分析，研究内容主要集中于各种识别参数的提取、选择和实验上。至今为止，说话人识别研究的重点依然集中在语音的特征提取、变换以及模式匹配新方法的建立上。

3.3.1 特征提取

怎样从语音信号中提取说话人的鉴别性信息是说话人识别的重要问题。对人的语音学感知研究表明，人与人之间的语音差异会受到先天和后天因素同时影响：不同人的发声器官具有生理差异；在不同环境中成长的人发音习惯也有所不同。这些都称为说话人之间的差异。除此之外，还有一些通过语音信号间接表现出来的信息，例如说话人的说话习惯、情感状态、用词特点等，这些可以称之为说话人本身的变化因素。虽然这些特征可以很好的区分说话人，但是人们很难找出它们和语音信号之间的对应关系，将这些特征定量化。因此，说话人识别的研究不得不退而求其次，仅利用物理上可以定量测定的参数来表征说话人，通过抑制说话人本身的变化，并突出说话人之间的差异来区分不同的说话人。但是在实际研究中，由于语音信号的复杂性，其中语音识别时所需要的语义特征和说话人的个性特征以难以定量的复杂形式交织在一起。如何从原始语音信号中提取说话人的特征信息仍然是一个悬而未决的问题。因此，说话人识别中所采用的还是语音识别中采用的特征，这些特征也能够间接的提供有效的说话人区分性信息。

语音信号通常可以看成短时平稳的序列，特征提取过程首先对语音信号分帧处理，并利用窗函数减少截断处理的 Gibbs 效应。利用高频预加重提升高频信息，然后对每帧语音信号进行频谱处理，得到不同的特征参数。目前研究者常使用的特征参数有线性预测倒谱系数（Linear predictive cepstrum coefficient, LPCC）、Mel 频率倒谱系数（Mel-Frequency cepstrum coefficient, MFCC）和感知线性预测（Perceptual linear predictive, PLP）参数等等。在说话人识别中，MFCC 比 LPCC 和 PLP 的识别性能更佳，是目前应用最广的特征参数。

3.3.2 模式划分和模式识别

在说话人识别系统中，不同的模式建立和匹配方法对应不同的说话人识别方法。常用的识别方法有模板匹配法、GMM-UBM 框架、支持向量机的识别方法、因子分析和 i-vector 模型、人工神经网络等。

- **模板匹配法**

在训练阶段，模板匹配法从说话人的训练语句中提取出相应的特征矢量，作为可以充分表征说话人特性的参考模板；在测试阶段，同样从测试说话人的语音信号中提取出具有表征性的特征矢量作为测试模板，并将其与参考模板比对。依据两者之间的相似匹配程度给出接受或者拒绝的判决。基于模板匹配的方法的一个缺点是其对于说话人模型的存储量需求较大，需要存储大量的特征矢量模板；此外，在参考的说话人集合人数规模比较大时，识别的性能不佳。

- **GMM-UBM框架**

20 世纪 90 年代，Reynolds 提出使用高斯混合模型对说话人建模，并迅速成为文本无关说话人识别领域占统治性地位的建模方法，使说话人识别的研究进入了一个新的阶段。2000 年，Reynolds 等人又提出高斯混合模型-通用背景模型(Gaussian mixture model-Universal background model, GMM-UBM）的说话人识别框架。从此，说话人识别从实验室研究阶段逐渐进入到应用阶段。

GMM-UBM 框架依赖于一个充分训练的通用背景模型（UBM），这个模型通常使用大量的说话人语音训练得到，用于代表一个说话人无关的高斯混合模型。然后，每一个注册登录的说话人则用一个与此说话人相关的高斯混合模型来表示，说话人模型一般通过该说话人的训练语音在通用背景模型上通过最大后验概率估计（Maximum a posterior, MAP）得到。在很多情况下，说话人注册登录的训练语音长度有限，不能对整个音素空间进行覆盖，因此，在 GMM-UBM 的框架下，训练语音能够覆盖的音素空间采用说话人训练语音建模，在训练语音不能覆盖的音素空间采用 UBM 模型近似表示，从而可以减小测试语音与训练语音不同带来的影响，提高识别率。

在 UBM 的训练过程中需要大量的训练数据保证，另外训练数据的声学信道需要和测试环境的信道保持一致性，然而在实际应用中，通常很难收集到足够的信道匹配数据来训练一个完全一致的 UBM。除此之外，实际应用中信道是可以发生改变的，这些都将导致预先训练的说话人识别系统性能下降。

- **支持向量机的识别方法**

SVM 的基本思想是将输入空间的向量映射到高维扩展空间，然后再高维扩展空间中采用分类方法构造最优超平面分界面，来解决模式划分问题。SVM 最初被用到说话人识别领域时普遍采用的是基于帧的方法，把每帧特征向量作为 SVM 的输入进行识别，然后统计测试语音中各帧的打分，得到最终的判决结果。

美国 MIT 大学 Lincoln 实验室的 Campbell 等人将 GMM-Supervector 作为 SVM 说话人识别系统的输入特征，并采用线性 K-L 核的 SVM 进行识别。一方面，该方法在综合利用 GMM-UBM 的优点之上，有效利用了均值之间的相关性进行后续处理；另一方面，K-L 核综合考虑了 GMM 模型训练时的协方差和权重影响，体现了说话人的特征向量在空间中的分布情况。系统性能得到较大提高。

- **因子分析和i-vector模型**

近年来，Kenny 提出的联合因子分析（JFA）方法在跨信道的说话人识别领域取得了很大的成功。该方法假设说话人变化和信道变化是互相独立的随机变量，限定其各自的变化都是在一个低维度的子空间上，并且满足标准的正态分布。利用因子分析方法可以从语音信号中推导出信道因子，从而分离出语音信号的说话人表示。

后来一些研究表明，一段语音中所包含的说话人因子和会话因子很难仅通过 JFA 的方法完全分离开来，分离后的会话因子中仍然包含一部分说话人信息。Dehak 等人在此基础上提出了 i-vector 模型，采用单一的“全变量因子”同时表示说话人信息和会话信息，使得在计算后验概率时可以保留更多说话人信息。

在 i-vector 框架下，说话人模型是通过本征音适应(Eigenvoice adaptation)得到的。首先，相对于高阶的 GMM 空间，该方法定义了一个低维度的全变量子空间(Total variability space)，并假设说话人因子只在该空间上变化，该空间需要用大规模的开发集训练得到。估计特定说话人的语音在全变量子空间上的后验分布，以其均值向量作为该特定说话人的 i-vector。I-vector 模型框架下，说话人模型和测试语音都使用 i-vector 表示，系统采用 i-vector 之间的余弦夹角作为说话人相似性的度量，以此做出识别判决。

I-vector 模型的一个优势是极大的保留了更多的说话人信息，而同时带来的缺点是会话等信息的混合引入，导致了说话人区分性能力的下降。因此，i-vector 模型方法中很重要的一个部分是需要一些区分性的方法来减少和抑制会话等因素的变化，增强说话人的特征信息。类内协方差归一化（WCCN）方法通过线性核函数的优化得到一个线性变换，干扰属性投影（NAP）选择在不同信道的信号中最小差异的投影方向，这些方法都可以不同程度的增强说话人信息的表示。另一个可以显著提高 i-vector 系统性能的说话人区分性的方法是概率线性区分性分析（PLDA）。该方法是线性区分性分析（LDA）的概率版本，在类变量中隐含了一个高斯的先验，因此可以采用较少的语音数据对特定说话人建模。近些年，i-vector 方法和 PLDA 建模方法的结合取得了复杂信道下说话人识别的最优性能。

- **人工神经网络的方法**

人工神经网络（Artificial Neural Network, ANN）试图模仿人脑信息处理机制，把大量结构简单的计算单元互相连接起来，实现高度并行和分布的处理。在说话人识别应用中，人工神经网络可以通过训练，更好地划分语音中所包含的说话人特征在特征空间中的分布。

近年来，随着深度神经网络在机器学习、语音识别、图像处理等领域的快速发展和成功应用，其相关方法也逐渐应用到说话人识别中，并取得了较好的成果。通过结合语音识别中基于音素状态的 DNN 模型和 i-vector 模型来对说话人空间建模，使得说话人识别取得了较大的性能提升；基于说话人标签的 DNN 模型可以提取 Bottleneck 特征代替原始的声学特征参数，从而得到更有利于说话人识别的特征表示。

4. 展望和发展趋势

4.1 语音合成

纵观语音合成研究发展态势和技术现状，以下研究方向将可能成为未来语音合成研究必须攻克的堡垒。

4.1.1 基于端到端的语音合成方法研究

语音合成系统由文本分析、韵律处理和声学处理三个模块形成一个有机整体，各个模块之间是相互依存的，当前主流的语音合成框架将不同模块分别进行处理，一方面前一阶段的误差会影响到后一阶段的建模，另一方面这种框架并不符合人对言语感知和生成的过程。当前，基于端到端的建模方法在机器翻译、语音识别、目标跟踪等不同领域得到成功应用，这种建模思想也适合于语音合成处理，基于端到端的语音合成目前正在得到国内外学者的普遍关注。

4.1.2 面向自然口语的语音合成

当前的语音合成系统针对朗读体内容能够合成出高质量的语音，但是针对更具表现力的自然口语，合成效果不尽人意。一方面由于系统对韵律信息的捕获不准确，另一方面由于生成的声学参数存在误差。提高自然口语语音合成的表现力，可以有效的提升语音交互系统的体验感，极大的拓宽语音合成的应用场景。因此，如何充分的挖掘自然口语中的语义信息，如何针对自然口语语料提高韵律模型和声学模型的精度将是语音合成领域急需解决的一个难题。

4.1.3 多说话人、多语言语音合成问题

当前语音合成大多面向单一说话人、特定语言的语音合成，这极大的限制了语音合成技术在工业界的应用；虽然一些自适应方法可以实现特定说话人的语音，但是语音音质有所下降，难以达到实用化的要求；同时，现有的多语言语音合成方法大多需要有相应语言的音库作为支撑，语料获取的难度制约了这项技术的推广。如何利用数据驱动方法和自适应技术实现任意说话人、不同风格的高质量语音合成，如何充分挖掘不同语言的发音空间，在语料受限条件下实现多语言语音合成，是未来亟待解决的问题。上述问题的解决将极大拓宽语音合成的应用场景。

4.1.4 融入发音机理和听觉感知的语音合成

现有在工业界成熟应用的语音合成技术，是以统计机器学习方法作为有力支撑的，虽然在语音生成的过程中考虑了发音机理和听觉感知的特性，但是这方面的应用尚属起步阶段，如何将发音机理和听觉感知融入到语音生成非线性建模过程中，将语音生成的物理模型和统计模型有机的结合，这对于深入探索人类的发音机理，对于细化语音合成的建模过程使之更好的实用化，都将具有十分重要的意义，这将是一个充满期望的研究方向。

纵观近年来语音合成研发的趋势和现状，我们有理由相信，随着机器学习、语音信号处理、自然语言处理等相关技术的快速发展，语音合成这一人工智能领域中具有挑战的问题将在不久的将来得到相当程度的解决，同时语音合成系统的产业化应用前景将更加广阔。

4.2 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，机器识别的性能还远远不能让人满意。本节对这一技术的未来发展做一展望，希望引起更多兴趣。

4.2.1 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较差。这是因为远端声音包含更多背景噪音，且有回声干扰。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。

4.2.2 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转移学习，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。

4.2.3 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。

4.2.4 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。端对端训练有可能是一种有效的方法。

4.2.5 神经网络结构学习

当前深度学习越来越依赖计算图模型(computing graph)架构。基于这一架构，研究者可对神经网络的结构进行任意设计，计算图模型架构可自动计算梯度，从而极大节约了设计优化算法上的成本。一些大公司依赖其计算资源优势，利用这一方法寻找最合适的网络结构。未来这一方法在工程上可能会设计出极其复杂的网络，显著提高最终系统的性能。然而，这种穷举式的结构搜索方法可能会被结构学习方法所替代，即将网络结构也作为参数的一部分进行学习，从而通过数据驱动得到优化网络。

4.2.6 神经网络持续学习

当前网络优化多基于 SGD 算法。这一算法的一个显著缺陷是当网络学习完成以后，很难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，逐渐忘记以前的环境，适应到新环境。研究者提出了一些方法（如 AdaGrad）来解决这一缺陷， 但这些方法是否能实现一个持续学习的语音识别系统，需要进一步研究。

4.3 说话人识别

从说话人识别研究的现状和发展态势可以看出，该技术已经逐步从实验室研究阶段发展成一项较为成熟的现代应用技术。现阶段主流说话人识别系统的系统性能、系统鲁棒性等都已经能够满足一些基本的应用场景需要。然而，说话人识别技术在研究向应用产品转化的过程中还有许多亟待解决的问题，以下问题或研究方向将可能成为未来说话人识别研究需要攻克的难题。

4.3.1 寻求更具说话人区分性的特征

实践应用表明，在复杂环境下，现阶段主流的说话人识别系统性能会明显降低。由于现阶段说话人识别的系统框架中，使用的说话人特征和模型都不是说话人的本质特征和模型，这些特征和模型很容易受到客观条件因素的影响。一方面，噪声、信道等环境差异会降低特

征的区分性；另一方面，说话人年龄、身体状况甚至说话时情感的变化，都会导致识别与预留时的不匹配，从而使识别性能下降。因此，寻求更具说话人区分性的特征甚至是说话人的本质特征，对于复杂环境下说话人识别系统的鲁棒性提升具有重要意义。

4.3.2 提升说话人识别系统防攻击能力

根据前文的阐述可以看到，在实际应用中，不法分子会使用各种各样的方法对说话人识别系统进行闯入攻击，如何成功的检测并拦截这些攻击，对于保证说话人识别系统在应用中的安全性尤为重要。语音作为信息的载体，其中除了含有说话人信息之外，还有丰富的内容信息、情感信息、信道信息等，这些信息往往以不为人知的复杂方式交织在一起，如何充分挖掘和利用这部分隐含信息，根据这些信息找出攻击者语音中所含有的特征，以此检测出攻击者并将其拒之门外。此类问题很可能会成为今后的重点研究发展方向，该问题的解决将会促进说话人识别系统在应用领域的进一步发展。

4.3.3 多生物特征融合技术

说话人识别技术发展到今天，其对所面临问题的解决方法并未完全成熟，在实际情景中，一些安全性较高的应用场景内，说话人识别作为单一的身份认证方式还很难成为现实。多生物特征融合技术是目前解决此问题的有效方法之一。例如，目前在远程身份认证中，出现了一些以说话人识别和人脸识别同时认证的应用方案，该方案在一定程度上增强了系统的安全性。结合特定的应用场景，研究者可以研究如何融合多种生物特征，使得系统性能更加安全，更加有保障。

第十七章 文字识别研究进展、现状及趋势

1.任务定义、目标和研究意义

文字识别（Character Recognition），广义地称为文档分析（Document Analysis），是对文档图像中的文字进行分割、识别，将文档从图像转换为电子文本的技术。具体内容包括文档图像预处理、版面分析、字符切分、字符识别、文本行识别等。文本行是文档图像的基本和相对容易分割的单元，因而文本行识别是最核心、也最难的问题，因为字符切分和字符识别不能分开，而且同时要考虑上下文信息（语言模型和几何上下文）。

文字识别技术是中文信息处理中非常重要的一环，因为大量的文档以纸张（如书籍、报纸、档案、票据）形式存在，而年代较早的纸张文档都没有对应的电子文件，只有通过文字识别转换成电子文本后才能对其内容进行语义分析。文字识别应用在上世纪 90 年代中期达到一个高潮。当时模式识别方法和技术逐渐成熟，个人计算机和扫描仪迅速普及，为文字识别技术推广应用提供了良好条件。2000 年前后跌入一个低潮，当时存在纸张文档会越来越少的担忧。事实上，这是一个错误的判断，虽然有些文档（如手写书信和表格）新增量在减少，但文档总量在继续增加，而且历史上的文档绝大部分没有电子化。另一方面，有些类型的文档（如公文、快递单、网络合成文档、拍照文档图像）在加快增长。最近五年，随着数码相机和智能手机的普及使随时随地拍照识别成为可能，加上技术的进一步发展，文字识别迎来了一个新的应用高潮，不断产生新的应用模式和技术需求。

文字识别的方式按照文档的媒体形式分为两大类：脱机（offline）文字识别和联机（online）文字识别。脱机文字识别是指对已经存在于纸张或物体表面（如建筑物标牌、交通标志）的文字进行提取和识别，处理对象是通过扫描或拍照得到的文档图像。联机文字识别是对书写过程中采集到的笔划轨迹（如触屏书写、手写板书写、数码笔书写）进行文字提取和识别。脱机文字识别根据文字的书体又分为印刷体文字识别和手写体文字识别，而有些文档（如各种手填票据、手写批注文件）中手写体和印刷体是并存的，因此需要系统能准确将手写体和印刷体文字分开，或同时对手写体和印刷体文字进行识别。联机手写的笔划轨迹也可看作是一种图像。

2. 研究内容和关键科学问题

按照从文档图像到电子文本的处理流程所涉及的技术环节，文字识别技术的主要研究内容和涉及的关键科学问题如下：

2.1 文档图像预处理

传统的扫描纸张文档图像比较清晰，通过简单的二值化即可分割文本和背景。然而，拍照的自然场景图像和纸张文档图像存在光照不均、视角变化、聚焦模糊、扭曲变形等问题，需要通过图像增强、复原等手段来改善图像的对比度和校正变形。这些是文档图像预处理的范畴。

2.2 版面分析

自然场景图像中文本检测与定位是一个复杂的版面分析问题，近年来有大量的研究工作。纸张文档图像（包括印刷体和手写文档）和联机手写文档（笔划轨迹构成，也可能是图文混合）中如何准确分割文本段落和文本行，区别和分割文本与图形、公式、表格、符号等，从而方便后续的文本行识别和公式、符号识别，是版面分析的主要研究内容。

2.3 文本行识别

文本行识别是文档分析的最核心技术，对文本行图像进行字符切分和识别，得到对应的电子文本（字符串）。相关研究内容包括：字符识别器设计（包括特征提取、分类器设计和学习），字符过切分，几何上下文建模，语言上下文建模，上下文融合模型，文本行序列表示和整体识别模型等。

2.4 后处理和应用

经过版面分析和文本行识别（如果需要，还可以有表格分析、公式识别、符号识别等）后，文字识别的后处理主要有两个目的：一是根据文档中不同元素（文本、图形、符号）的几何关系和语义关系对识别结果进行消歧和纠错，二是结合识别结果和几何关系对文档进行重构得到结构化电子文档（如 PDF）。文字识别的应用方面，除了语义分析和信息提取，文档检索是一个比较普遍可行的应用，因为检索不需要文字识别精度很高。文档检索的研究内容包括文本分类和聚类、基于文本语义的检索、关键词检索（可基于文字识别或不需要文字识别）等。

文字识别技术主要有两个关键科学问题：

字符切分和识别的复杂性：字符切分和识别的复杂性是针对文本行识别而言的。文本行图像中由于字符间粘连、有些字符多部首、多语种混合，加上手写体的字符变形、大小和间隔不均匀等因素，字符在被识别之前很难准确且分开。因此，字符切分和识别要同时进行，这需要对字符切分和识别的模型、文本行识别的整体建模（包括切分与识别的融合和上下文建模）进行深入研究和精心设计。

文档版面分析的复杂性：版面分析的复杂性有几个方面：图像背景复杂，文本段落排版的多样性，文本行方向多样性（水平、垂直、倾斜，甚至有弯曲），图文混合（包括文本、图形、表格、公式、特殊符号等）。复杂版面文档中准确区分文本和图形，准确分割段落和文本行都是比较困难的。

3. 技术方法和研究现状

近十年来，文字识别领域的研究重点跟过去相比有一些重要变化。首先，在成像手段上，数码相机和智能手机的普及化导致拍照文档图像越来越多。跟扫描图像相比，拍照图像具有光照不均、几何变形、容易模糊等特点，因而在图像增强和几何校正方面带来一些新的研究问题。除了纸张文档的识别，自然场景图像中的文本检测与识别成为一个新的研究热点。在识别单元上，单字识别已不是研究的重点，而是集成字符切分、识别和上下文的文本行识别。在应用方面，历史文档（古籍）的图像分析和识别、检索受到广泛关注。由于文档成像的便利和应用增多，近几年文字识别领域的学术会议，如国际文档分析与识别会议（ICDAR）、国际手写识别前沿会议（ICFHR）、国际文档分析系统研讨会（DAS）吸引了越来越多的人参加。

下面从不同的研究内容和产业应用角度介绍主要方法和现状。

3.1 文档图像预处理和版面分析

在文档图像预处理方面，历史文档因纸张陈旧、污损等原因，即使是扫描的图像也呈现严重的背景噪声。因此如何将文本与背景分开成为一个重要的研究问题，传统的二值化和局部二值化方法性能不佳。近几年提出了一些结合灰度和边缘信息、结合多特征对像素进行分类的自适应二值化方法。对手写历史文档二值化的竞赛 Competition on Handwritten Document Image Binarization （H-DIBCO 2010, 2012, 2014）吸引了大量研究者参加。基于局部对比度、梯度和边缘信息的自适应二值化方法取得了较好的性能。基于图模型（如马尔科夫随机场、条件随机场）结合像素分类和空间上下文的方法由于更好地融合了局部和空间信息，具有良好的发展潜力。

另一个预处理问题，拍照文档的几何校正和光照矫正多采用立体视觉模型和几何分析模型。如中科院自动化所的孟高峰等人用广义圆柱形表面模型对拍照书籍图像进行几何形变矫正，用凸包模型对扫描书籍图像（因放置不平而带来光照不均）进行光照矫正。

版面分析将文档图像分割为文本段落、图形、表格等区域，其方法可分为两大类：基于前景的方法和基于背景的方法。基于前景的方法将像素或连通部件进行逐级聚合，得到分本行和段落，并且对连通部件或区域进行分类判断是文本或图形。代表性方法有文档谱（Docstrum）方法、基于块邻接图（Block Adjacency Graph, BAG）的方法，基于最小张成树（Minimal Spanning Tree, MST）的聚类的方法、基于 Hough 变换的方法、基于纹理分割的方法等。基于背景的方法对文档图像进行自上而下的划分，如通过投影找到栏、段落、文本行之间的空白，代表性方法有递归水平-垂直切割（Recursive X-Y Cut）、Voronoi Diagram 方法、背景矩形（White Space）分析等。

3.2 自然场景文本检测与识别

自然场景图像文本检测与识别吸引了大量的研究者。文本检测方法可分为基于纹理（区域分类）和基于连通部件两大类。基于纹理的方法对图像进行多尺度滑动窗分析，判断每个窗口的纹理是否为文本，在此基础上对图像进行分割。基于连通部件的方法先通过图像区域分割或边缘分析提取连通部件，然后通过几何分析或分类器判断每个连通部件为文本或非文本，最后将文本连通部件聚合为文本行。近几年的主流方法是用最大稳定极值区域（MSER）方法提取候选连通部件，对连通部件进行过滤、聚合得到文本行。这类方法在多次竞赛和公开数据集上取得了领先的性能。

对于场景图像中的文本行识别，有的方法是在文本检测定位并得到二值图像的基础上用集成字符切分、分类器和上下文的传统方法（下面文本行识别部分详述）进行识别，有的方法则把文本检测和识别同时进行，即用字符识别器进行文本检测，称为 End-to-End 方法。在文本定位基础上，不用二值化直接对彩色图像进行字符切分和识别（结合上下文）的方法也取得了优良的性能，尤其是采用深度神经网络的方法。基于长短时记忆（LSTM）循环神经网络的方法在英文词和文本行识别中取得了领先的性能，但用于中文文本行识别还不成功。

值得注意的是，互联网上的文档图像有多种类型，除了扫描或拍照的纸张文档、自然场景文本图像，还有大量的合成文档图像，是将电子文本设置字体后叠加在简单背景或复杂背景的图片上形成。合成文档图像中的文本检测与识别跟自然场景文本检测与识别的方法类似，难度相对小一些。

图 1. 自然场景文本检测与分割示例

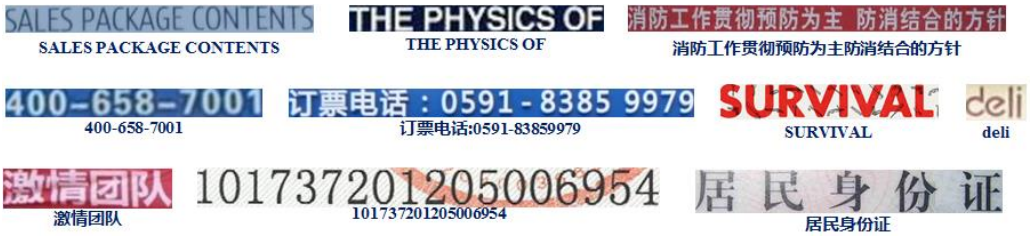
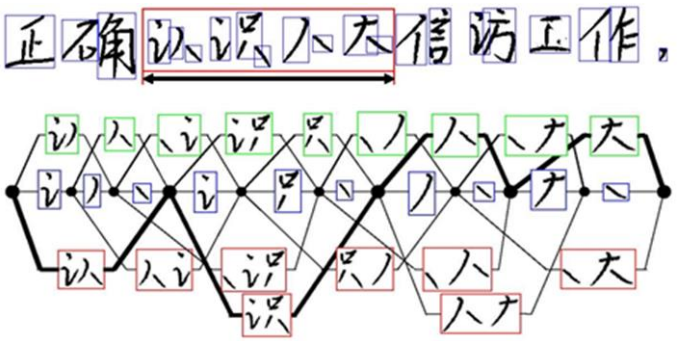


图 2. 自然场景文本行识别示例

3.3 手写文本行识别

手写文本行图像中字符难以在识别之前被准确切分，因此字符切分和识别需要同时进行，或者对文本行图像整体进行识别。在中文手写识别方面，2011 年中科院自动化所发布了一个大规模手写文档和字符样本数据库并组织了几次学术竞赛，带动了该领域单字识别和文本行识别方法与性能的进步。目前比较成功的方法是基于过切分和候选切分-识别路径评价搜索的方法。该方法在字符过切分（尽可能将不同字符分开，每个片段为字符或字符的一部分）基础上组合片段生成候选字符，用字符分类器对所有候选字符分类给出候选类别和置信度，融合几何上下文和语言上下文对候选切分-识别路径进行评价，搜索最优路径得到字符切分和识别结果。在此框架下，基于贝叶斯决策的上下文融合方法和基于半马尔科夫条件随机场的方法都取得了较高的字符切分和识别正确率。除了融合方法，字符分类器的精度、几何上下文和语言上下文的表示对文本行识别性能有决定性影响。近几年在字符识别器的书写人自适应和语言模型适应方面也取得了新的进展。一个值得注意的动向是，深度卷积神经网络（CNN）把手写汉字识别的精度提升了一大步。它可作为分类器集成在文本行识别系统中，提高文本行识别性能。



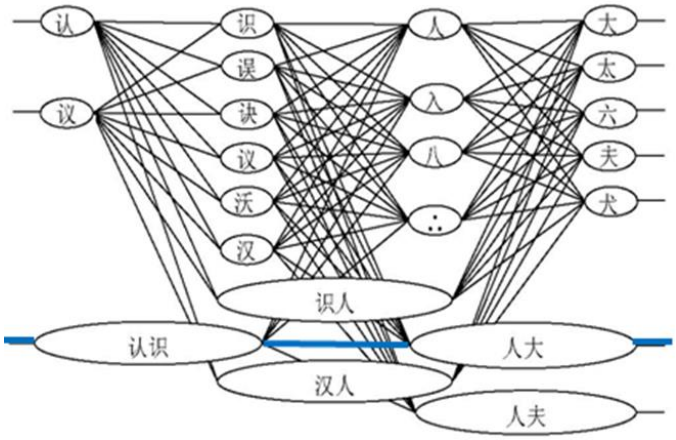


图 3. 候选切分-识别网格示意图

英文和阿拉伯文手写识别方面，近几年的趋势是采用长短时记忆（LSTM）单元的循环神经网络（RNN）取得了领先的性能，而基于隐马尔可夫模型（HMM）和神经网络混合的方法仍然被广泛采用。然而，这些方法用于中文手写文本行识别还没有取得明显的成功。

3.4 文档检索

由于手写文档和历史文档的文字识别错误率仍然较高（字符错误率一般在 5%以上，词识别错误率一般在 10%以上），完全电子化需要大量人工校对。从应用的角度，文档检索从文档数据库中查找与用户关心的主题或关键词匹配的文档图像或区域，具有广泛的应用价值。文档检索主要有两类方法：基于文字识别的检索和文字识别无关的检索。前一类方法依赖文字识别的精度，在识别精度较低的情况下检索性能会明显下降。文字识别无关的检索中，大部分工作是对用户查询的关键词进行检索。关键词检索方法分为基于图像查询的方法（Query-by-Example）和基于文本查询的方法（Query-by-Keyboard）。 基于文本查询的方法一般需要有字符或词的形状模型，用不同字体和书写风格的样本训练后具有较好的形变适应能力，而且用键盘输入文本查询的方式在使用上更方便。近年提出把词图像和文本嵌入共同特征空间的方法，可同时用于词识别和关键词检索。

3.5 产业现状

文字识别技术把纸张文档和拍照文本图像变成电子文本，具有广泛的应用价值。文字识别的传统应用包括印刷文档数字化、表格单据识别、邮政地址识别、名片识别、联机手写文字录入等。所有这些应用的需求量跟过去相比都在增加，技术也在进一步成熟。有些应用过去由于技术性能限制长期停滞不前，近年才取得明显进展，如手写票据识别。有些应用模式在变化，如名片识别从扫描成像转向手机拍照识别，联机手写识别从单字识别过渡到单词识别和句子识别。

最近新出现的应用可分为三类：一类是联机手写图文混合文档分析，这是由于手写轨迹采集设备进步使得大面积采集手写轨迹成为可能，如大屏幕手机、数码笔（如 Anoto Pen）等。但是这样的文档分析的性能还需改进。第二类是历史文档的识别与检索，这在西方已开展较多，但国内还在初步研究阶段。第三类是智能手机和移动互联带来的拍照文档识别和网络文档图像识别，这对个人、电子商务、网络内容管理部分都有广泛用途。近来新出现的应用模式包括百度公司的涂书笔记、百度翻译（可拍照识别），网络图片文本提取工具 Project Naptha，基于拍照识别的试题搜索等。

由于应用的增多，近年很多公司投入文字识别技术研发的力量也明显增加，比较知名的公司包括 Abbyy, Parascript, A2iA，微软、Google、苹果、三星、富士通等。

由于文字识别技术较少作为独立产品销售，而多数是搭载在硬件产品（如智能手机、ATM机）、互联网业务（如电子商务和内容搜索）或综合信息处理系统中，该技术本身的商业价值和销售额很难估计。不过可以肯定的是，其应用范围很广而且在很多时候发挥关键作用（比如互联网图片中的文字必须经识别才能用于内容分析和检索），并且由于文字信息到处存在，人们经常跟文字打交道，可开发的潜在应用还很多。

4. 总结及展望

近几年，文字识别技术得到了快速发展。这得益于两方面的因素。一方面，互联网和移动通信、便携式成像设备的推广应用带来了很多新的文字识别应用需求。另一方面，模式识别、机器学习、计算机视觉等领域理论方法的发展给文字识别技术研发产生了很多启发和促进。

文字识别技术面临非常广泛的应用需求，但是实际应用的推进仍然比较慢。主要原因在于已有技术和性能仍然难以满足应用的需要。近几年，计算设备和通信网、大数据推动了文字识别技术的发展，但是在原理方法上需要不断创新才能真正使技术变得好用。最近几年的进展主要是利用了机器学习（如深度学习）方法和采用大量样本数据训练得到的，而这种方法在学习的灵活性和自动程度、对模式的理解能力还是有局限的。

从理论方法的角度，文字识别领域将来的主要研究方向包括：

文字识别的认知机理和受认知启发的文字识别方法：人的文本检测、分割和识别能力很强，很容易发现和识别图像中的文本，目前的自动检测和识别性能跟人相比有较大差距。因此，需要从人的感知和认知机理得到启发设计新的处理方法。

文档识别模型的小样本学习与自适应：手写字符和文本识别中，现有的深度神经网络和循环神经网络经过大量数据训练可以得到超过人的识别精度，但是人是从少数样本开始学习的，而且在阅读中不断积累知识和适应环境及各种字体和书写风格。希望将来的计算机识别系统具有这种能力。

文档分析与识别的结构化学习与优化模型：文档中的各种元素（文本、图像、表格、符号等）相互关联，需要用一个结构化模型来表示他们之间的相互关系，充分利用几何上下文关系和语言上下文关系来提高各个元素识别的准确性，这个结构化模型的有效学习和优化推理都是研究的难点。

从应用的角度，下面几个问题值得重视：（1）图文混合、印刷和手写混合的文档识别与结构理解，这是有大量应用需求的问题。(2)网络文档图像识别与检索，这对网络信息检索、敏感内容监测等具有重要意义。(3）多语言文档分析与识别，尤其是互联网上多语言文档和各种民族语言文字的识别。(4）历史文档（古籍）的识别与检索，这在中国有迫切的需求，而且面临与一般文档识别不同的一些难点（如字符类别数特别多，学习样本很少）。

总之，文字识别技术既有大量的应用需求，又有一系列的理论和技术难题，需要投入大量的研发力量，进一步推动学术研究和应用的发展。

第十八章 多模态信息处理研究进展、现状及趋势

1. 任务定义、目标和研究意义

多模态（multimodality）的概念起源于计算机人机交互领域信息表示方式的研究，其中术语“模态”一词被定义为在特定物理媒介上信息的表示及交换方式。在研究中人们发现，用语言、视频、音频等媒体指称来描述信息表示方式过于宽泛、粒度太大，不足以区分实际采用的表示方式，为此引入了比媒体（或媒介）更细粒度的“模态”概念。而多媒体媒介可以分解为多个单模态，如视频作为一种多媒体媒介，可以分解为动态图像、动态语音、动态文本等多个单模态。为了模态概念定义的科学性和实用性，单模态的分类必须满足完整性、正交性、关联性和直观性的要求。在同一事物上多类单模态信息共生或共现的现象是十分普遍的。人与人交谈时有声语音与文字文本是共生的；互联网网页中图片与其对应的解说文字是共现的，凡此等等。共生或共现的多种单模态信息的统称即所谓的多模态信息。融合多种单模态的信息处理即所谓的多模态信息处理，其中涉及对多模态信息的获取、组织、分析、检索、理解、创建等。

多模态信息处理技术主要应用于对象识别、信息检索、人机对话等与智能系统及人工智能相关的领域。大量研究成果显示，基于多模态理念的信息处理算法和方法，往往会得到比传统方法更好的性能和效果。例如，语义计算相关领域基于指称语义的研究发现，采用语言表达式的视觉指称（即一组图片）来定义指称相似性度量，在某些语义推导任务中，效果好于基于纯文本的分布式语义表示；情感计算领域相关研究发现，不同模态的数据在情感表达中具有互补性，在愉悦度表达方面文本模态优于音频模态，而在激活度表达方面音频模态则优于文本模态。在基于内容的多媒体信息检索领域，针对基于内容的视音频检索中的语义鸿沟问题，利用与视音频数据共生或共现的文本信息，进行多模态的语义分析和相似性度量，是克服语义鸿沟问题的一种十分有效的方法。以媒体为单位的跨媒体信息处理任务，普遍存在语义鸿沟问题，所处理信息对象的语义，无论是基于外延语义（指称语义）还是内涵语义（关联语义）概念，在单一媒体信息范围内得不到完整或最终表达，而多模态信息处理方法为该问题的解决提供了新的思路和方法。

2. 研究内容和关键科学问题

多模态信息处理是在文本、图像、音频等现有单媒体信息处理的基础上发展起来的，现有单媒体数据的处理方法是多模态数据处理的基础。例如在特征提取层面，针对文本、图像、音频等单模态数据，往往直接利用成熟的文本、图像、音频特征提取方法来实现。多模态信息处理特有的研究内容主要关注于多模态信息的建模、获取、融合、语义度量、分析、检索等方面。

2.1 多模态信息建模

如何科学、严谨的定义单模态信息，是多模态信息建模要解决的问题。由于用媒体方式界定人机交互方式粒度太大，从而引入了模态的概念。所谓多模态信息建模，就是要构建一个单模态的分类体系，在该分类体系中，各单模态类别之间满足完整性、正交性、关联性和直观性的要求。Niels Ole Bernsen 2008 年基于前人的工作，在“多模态理论(Multimodality Theory)”一文中给出了一个满足这些要求的单模态的分类体系，如表 1 所示。

表 1 一个输入/输出模态的分类

顶层	通用层	原子层	亚原子层
语言模态	1 静态拟真图形元素		
	2 静-动态拟真声音元素		
	3 静-动态拟真触觉元素		
	4 动态拟真图形	4a. 静动手势话语	
		4b. 静动手势关键字	
		4c. 静动手势符号	
	5 静态非拟真图形	5a. 书面文本	5a1. 打印文本 5a2. 手写文本
		5b. 书面关键字	5b1. 打印关键字 5b2. 手写关键字
		5c. 书面符号	5c1. 打印符号 5c2. 手写符号
	6 静-动态非拟真声音	6a. 口语话语	
		6b. 口语关键字	
		6c. 口语符号	
	7 静-动态非拟真触觉	7a. 触觉文本	
		7b. 触觉关键字	
		7c. 触觉符号	
	8 动态非拟真图形	8a. 动态书面文本	
		8b. 动态书面关键字	
		8c. 动态书面符号	
		8d. 静动口语话语	
		8e. 静动口语关键字	
		8f. 静动口语符号	
拟真模态	9. 静态图形	9a. 图像	
		9b. 地图	
		9c. 组合图表	
		9d. 图形	
		9e. 概念图表	
	10. 静-动态声音	10a. 图像	
		10b. 地图	
		10c. 组合图表	
		10d. 图形	
		10e. 概念图表	
	11. 静-动态触觉	11a. 图像	
		11b. 地图	
		11c. 组合图表	
		11d. 图形	
		11e. 概念图表	
	12. 动态图形	12a. 图像	12a1. 脸部表情
		12b. 地图	12a2. 手势
		12c. 组合图表	12a3. 肢体动作
		12d. 图形	
		12e. 概念图表	
主观模	13. 静态图形		
	14. 静-动态声音		
	15. 静-动态触觉		

态	16.动态图形		
显式结构模态	17.静态图形		
	18.静-动态声音		
	19.静-动态触觉		
	20.动态图形		

随着人机交互设备的发展和丰富，新的传感器可以采集到更多新的、可与人交互的信息，如定位信息、重力加速度信息、脑电信息、热量消耗信息、步行运动信息等，表 1 给出模态分类体系已不能完全覆盖新模态信息的种类，因此需要持续研究新的模态分类体系。

2.2 多模态信息获取

尽管人与人、人与机器之间交互信息的多模态现象是普遍存在的，但对于多模态信息处理而言，所处理的对象数据往往需要特殊处理才能获得。多模态信息的获取主要包括数据的采集、解析与数据集构建。

2.1.1 多模态数据的采集

尽管可以对单模态数据类别进行比较形式化的定义，但实际研究中只要尽可能地遵守完整性、正交性、关联性和直观性的原则，新模态数据类别的引入是比较灵活，同时也是比较活跃的。比如除了图像、声音等信息外，针对社交媒体，可通过智能终端，采集到位置、重力加速度、睡眠、运动等人体信息；针对车联网，可通过车载传感器，采集到车速、位置、温度、发动机转速、雷达等汽车状态信息；针对监控网，可以采集红外、震动、烟雾浓度、生物指纹等与安防相关的信息。

多数情况下，多模态信息处理任务要求所有处理样本数据的各单模态数据是完整的。好在各单模态数据源经常是共生或共现的，满足完整性要求是可以做得到的。但也有例外的情况，例如歌曲多模态信息中，尽管音频与歌词是共生的，但歌词很难从音频中分离，因此，歌词文本数据还要通过其它单独途径采集。

2.2.2 多模态数据的解析

多模态数据的解析就是将原始混合状态的多模态数据，分解为单模态的数据。例如视频数据，需要分解为动态图像、音频语言、文本语言等三种单模态数据，其中文本语言部分，可能来自于视频字幕、图像内容中的文字和语音识别的结果等。

多模态数据的解析往往需要与数据采集相结合，例如歌曲 MTV 视频的解析，歌词文本很难从视频本身得到，可以通过采集系统来弥补。再例如，艺术、影视评论类文本数据的解析，其中涉及的图像、视频、音频数据的获取，更需要借助采集系统来完成。

2.2.3 多模态训练数据集的构建

为了进行对多模态信息的机器学习处理，如分类、回归、聚类等，需要构建训练用样本数据集，特别是针对有监督学习，还需要进行数据标注。多模态训练数据集的构建有自己独特的方法。

以多模态人脸情感识别为例，需要选择一组参试人员，选择一组表达不同情感的诗词，准备一个相对封闭的环境，一个显示诗词的屏幕，一个面对受试人员脸部的摄像头，一个录音麦克风，一个采集视频、音频和交互数据的软件，交互数据通过受试人员拖动屏幕上采集

软件的滚动条来产生。标注的情感数据可采用二维连续的 VA 情感模型来量化，由于标注的情感模型是二维的，因此每个诗词样本都需要标注两次。标注开始后，受试人朗诵屏幕上的诗词，并根据朗诵诗词的情感体验拖动滚动条。最终可以获得包含有声语言、文本语言和人脸视频的多模态情感标注数据及相应的训练数据集。

2.3 多模态语义分析

术语“语义分析”在不同领域有不同的含义，这里特指机器学习中的语义分析。在机器学习中，语义分析是指构建一个文档集概念结构的任务，该概念结构逼近文档集所表达的概念。也即，运用机器学习的方法提取或挖掘文档的深层次概念。虽然语义分析一般不等同于文档的语义理解，但往往是语义理解的基础步骤。在语义分析相关研究中，所分析的文档集已从文本类数据，扩展到图像、视频、音频等其它媒体形式的数据集。以图像数据为例，所谓图像语义分析是指完整地将图像内容转换成可直观理解的类文本语言表达,即将图像内容“像素-区域-目标-场景”的层次关系，采用合适的词汇、合理的构词方式进行词汇编码和标注的过程。

语义分析过程中首先要面对的是如何克服语义本身在表达上的多义性和不确定性问题，如同词不同义，同义不同词的问题。对于图像、音频这样的非文本类数据，更要解决在数据表达和语义解释之间建立合理的联系的问题，即语义鸿沟问题。大量研究表明，多模态语义分析方法对解决上述两类问题具有明显的优势。例如,在对足球比赛视频语义分析的基础上，辅以音频欢呼声事件的鉴别，能够更好地分析出进球事件的语义。

所谓多模态语义分析是指在同一个媒体对象的多个模态数据上,同时并行或协同进行语义分析，并最终通过融合得到分析结果的语义分析方法。

2.4 多模态情感识别

人机交互、多媒体信息处理等多个领域的研究和应用，对情感计算技术的发展起到了重要的推动作用。

目前人机交互的主要方式仍是书面语言,书面语言交流与人类面对面交流的最大差别是,所谓副语言（Para-language）的缺失。副语言包括语气声、哭笑声、面部表情、肢体语言等。实现副语言的人机交流是实现和谐自然人机对话的基础。鉴于副语言更多地侧重情感语义表达的属性，引入情感识别技术来实现对副语言的理解是顺理成章的。为了处理语音和副语言这样的多模态数据，将情感识别技术扩展到处理多模态数据，既是所谓的多模态情感识别技术。

在多媒体检索研究领域，传统的基于文本知识的索引方法已显现出它的局限性，而基于情感的索引吸引了多媒体研究的学者们。在多媒体应用领域，用户也期望内容推荐和分发系统，能够更好地适应他们的体验和情感。多媒体情感分析与识别的研究目标是，在多媒体内容的推荐和检索中使用情感因素。例如，当把“我想听一首欢快的歌”、“我想看一部恐怖片”等检索条件输入给计算机系统时,计算机系统能够给出满足要求的响应。其中关键的前提是，多媒体内容的情感属性，不是人工标注的，而是计算机自己通过计算获得的。歌曲、电影数据的多模态属性，同样要求情感识别技术是多模态的。

2.5 多模态信息检索

随着经典的文本检索文本、图像检索图像的单模态信息检索技术的成熟与大规模应用，各单模态之间相互检索，诸如用图像检索文本、文本检索音频这样的跨媒体检索系统，也成为信息检索领域的研究热点。与单模态信息检索方式相比,跨媒体信息检索不仅能够更好地表达用户的检索意图,改善用户的检索体验，提高检索召回率和准确率，而且对媒体数据语义的理解也具有重要作用。跨媒体信息检索首先要解决的是所谓语义鸿沟问题,由于各单模

态内容的异构性导致语义的不可度量，使得传统多媒体检索方法不能直接适用于跨媒体检索。有多种方法被用来解决这一问题。一种方法是对多媒体数据不同模态的语义关系进行统一建模，以实现跨媒体检索。这种方法的缺点是受限于语义概念的建模规模；另一种方法是利用共生或共现的多模态信息作为语义桥梁，来实现跨媒体检索。广义上讲，上述两种检索方法，都可以被称为多模态信息检索，狭义上讲，后者为典型的多模态信息检索，前者可称为跨模态信息检索。

一个典型的多模态信息检索系统是欧盟基金项目 I-SEARCH （Axenopoulos， 2010，见图 1），该项目的目标是提供一个统一的多模态内容索引、搜索和检索框架，该框架能够处理指定的多媒体和多模态内容类型，如文本、图像、图形、视频、3D 对象和音频，现实对上述任何类型信息内容的检索和查询。I-SEARCH 将多种媒体类型封装到一个称为“内容对象（CO）”的媒体容器中，并共享相同的语义，同时，不同的媒体类型可拥有各自的的元数据，如文本、分类、位置或时间等信息。多模态信息的索引、检索和查询，都基于内容对象来完成。

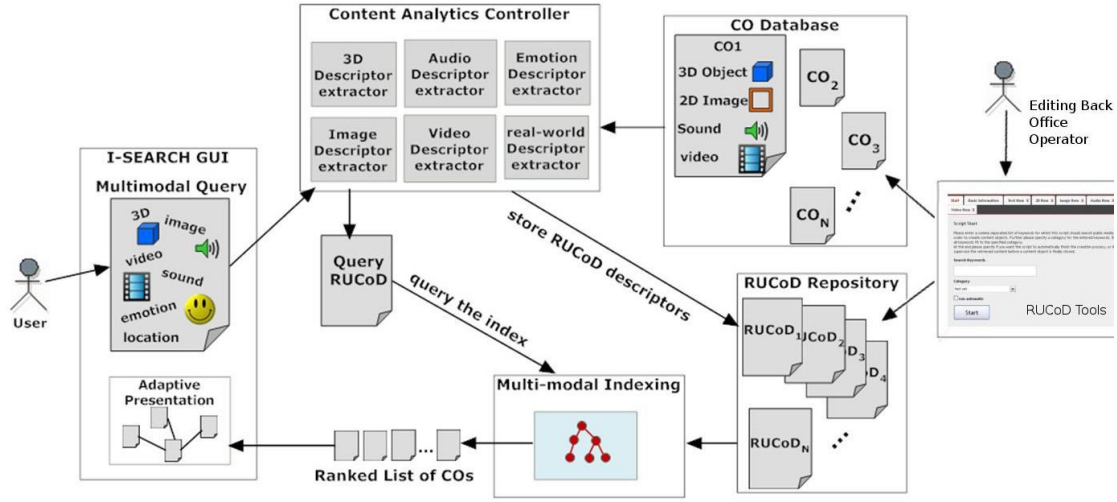


图1 I-SEARCH多模态检索系统框架

2.6 多模态人机对话

多模态人机对话系统与基于文本语言的传统人机对话系统类似，由信息获取、信息处理和信息输出三部分组成，不同之处在于，多模态人机对话系统的信息获取模块通过麦克风、摄像机等输入设备，采集语音、面部表情、肢体动作等多模态信息作为输入；信息处理模块对输入信息进行多模态融合的语义分析，并基于多模态知识库产生协同对话内容，该内容除语言内容外，还包括反映情感的面部表情内容；信息输出部分将两部分内容同步输出到输出设备上，目前主要是输出到有模拟对话人脸部图像的屏幕上，长远目标是输出到仿真机器人上，实现整合了语音、手势和面部表情的、类似人类的自然互动与对话。已有许多多模态人机对话系统框架被提出，如较早的 MIND 系统（Chai， J. 2002，见图 2），国内的中科院自动化所的系统（陶建华等， 2011，见图 3）。

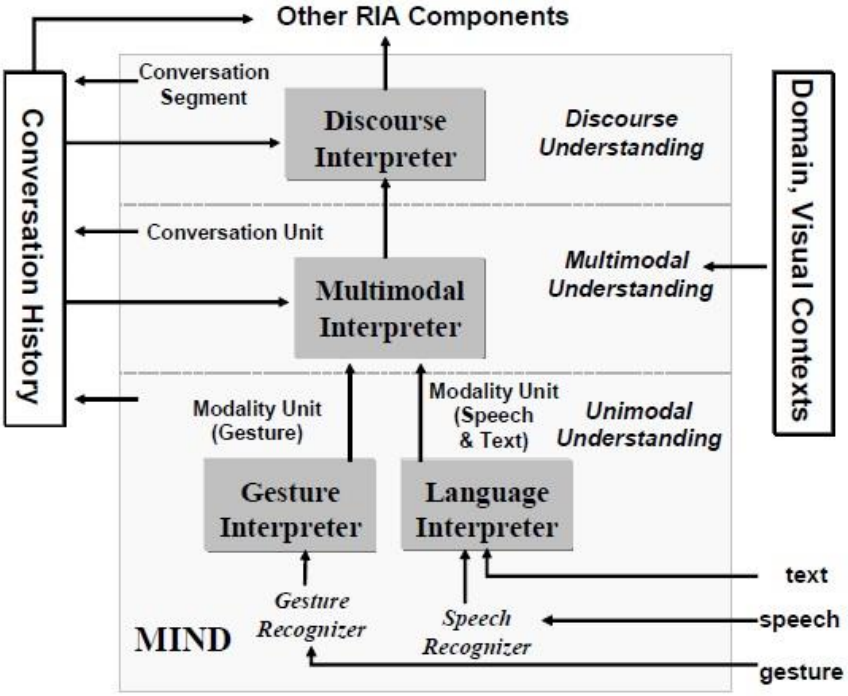


图2 MIND多模态人机对话系统框架

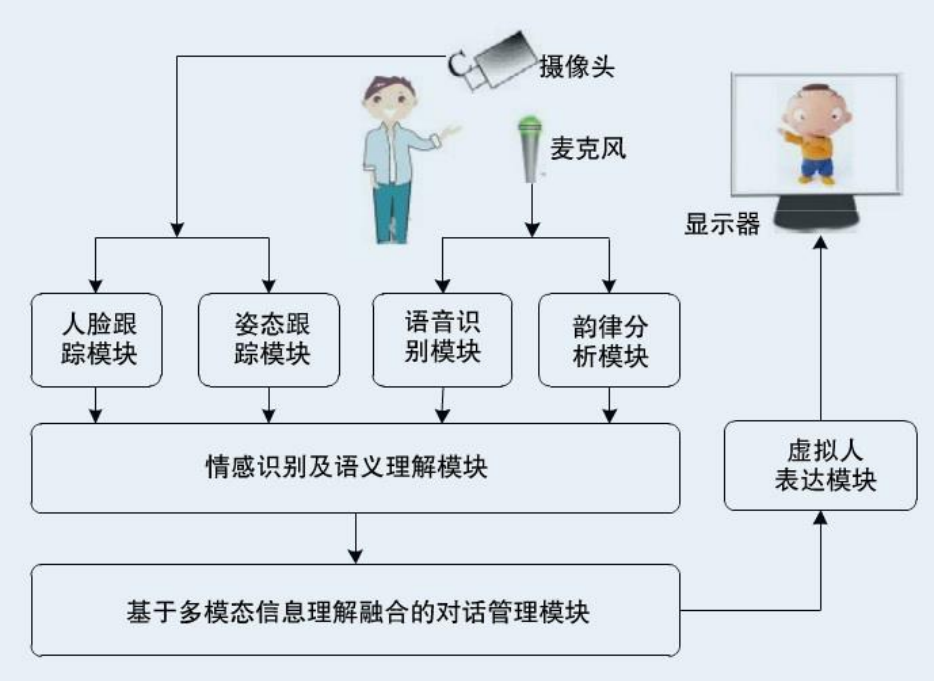


图3 多模态人机对话系统框架

多模态人机对话系统的核心研究内容是两个方面，即多模态会话内容的理解和多模态会话内容的生成。在会话内容理解方面，除了会话人情感识别外，对会话内容所涉及图像的理解，也成为研究的热点。如对基于图像字幕生成（看图说话， image caption generation，见图 3）的研究，以及更进一步的基于图像的问答系统的研究。这些研究的目标是实现机器对会话场景及会话视觉内容的理解。

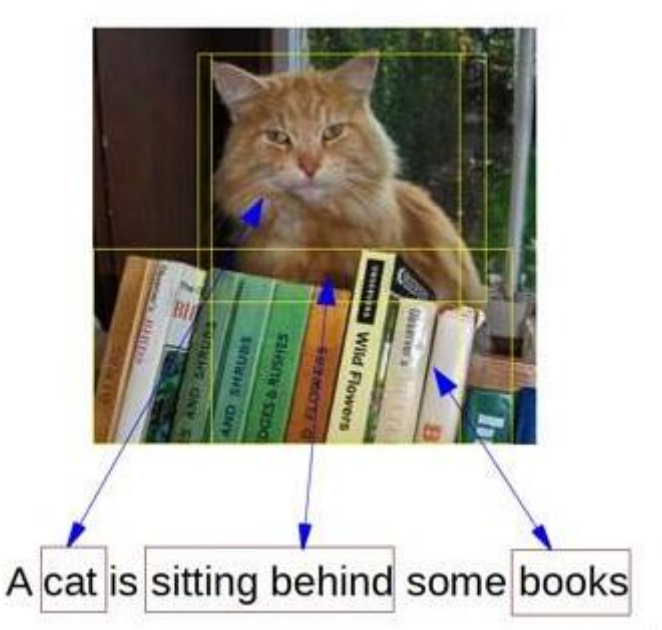


图4 基于图像的字幕生成

3. 技术方法和研究现状

为实现多模态信息处理的目标，大量的文本和多媒体信息处理的技术和方法被多模态信息处理系统集成和采纳。下面仅就多模态信息处理中比较重要的关键技术和方法作一介绍。

3.1 多模态融合方法

多模态信息由于底层数据的异构性，比如图像是 24 位的 RGB 颜色值矩阵、音频是 16 位的声压值串、中文文本是 16 位或 24 位的汉字编码串。如何让这些异构的数据完成同一个识别或检索任务，是多模态信息处理首先要解决的问题。解决这个问题的方法被称为多模态融合（Multimodal fusion）。所谓多模态融合是指：整合各种输入模态的信息，并将它们合并在一个完成同一目标的系统中的处理方法。以多模态人脸情感识别为例，输入的多模态信息为人脸图像和语音，一个最直观的融合方案是，分别对人脸图像和语音各构造一个情感识别系统，然后对两个系统的输出进行加权平均，得到最终的识别结果。

Pradeep K. Atrey 等人 2010 年在“多媒体分析的多模态融合综述”一文，对多模态融合方法作了较系统的总结和分析。关于实现多模态融合的方法，一般是在两个层次上进行融和，即特征层融合（或称早期融合）和决策层融合（或称后期融合）。第三种融合策略是所谓混合融合方法，该方法是将特征层融合与决策层融合结合起来一起使用。图 5 是 Pradeep K. Atrey 等人对各种融合方法给出的图示。

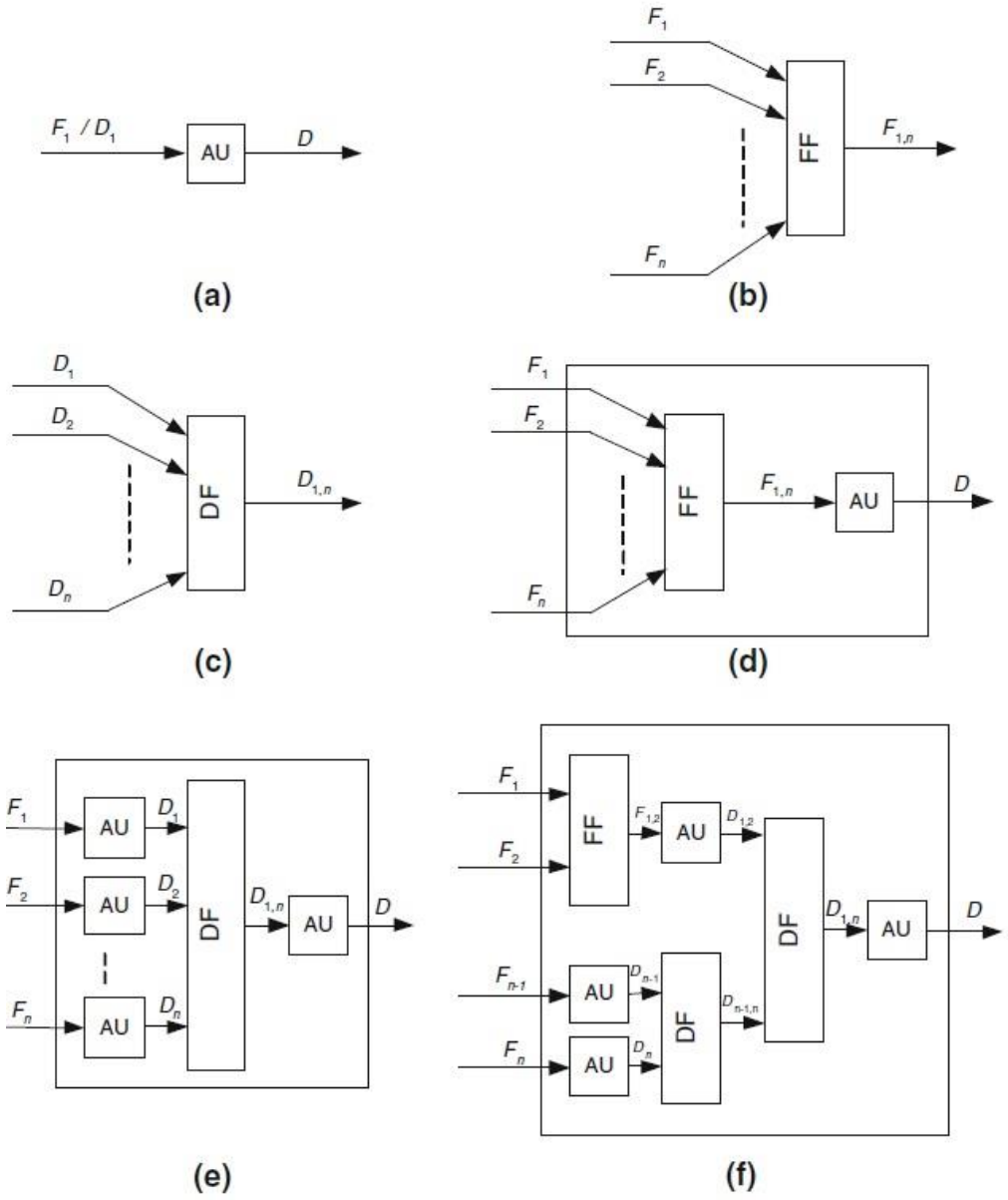


图 5 多模态融合方法示意图。 其中，（a）是分析单元，（b）是特征层融合单元，（c）是决策层融合单元，（d）特征层多模态分析，（e）决策层多模态分析，（f）混合多模态分析。

特征层融合方法先对各模态信息分别进行特征提取，再对特征数据进行综合分析和处理，形成规模更大的多模态联合特征矩阵或向量。对于采用机器学习作为决策层分析的系统，特征层融合的主要作用是使学习器有结构统一的输入样本数据。由于大多数机器学习方法对输入样本有格式要求，如等长的向量，此时特征融合是必须执行的处理步骤。最常见的融合方法是将不同模态的特征向量进行拼接。

决策层融合策略是，首先让每个模态单独完成各自的分析或属性判别，然后通过融合来它们的分析或属性判别结果，来形成最后的分析或判别结果。主要的融合技术有表决法、加权线性表达式、集成学习、协同学习、多层学习等。

为了结合特征层融合和决策层融合各自的优势，一些研究人员提出一种将特征层融合与决策层融合的组合策略，即所谓的混合融合。混合融合的方案有多种选择，图 5（f）是其中的一种，这种方案可以实现部分模态采用特征层融合，部分模态采用决策层融合的融合策略。

3.2 多模态深度学习

采用深度学习方法研究多模态信息处理问题是近年来的热门方向。学者们充分利用了深度学习的特点，针对多模态信息处理任务，提出了一系列新的方法和算法。

深度学习是一个非常好的多模态融和工具。多模态深度学习模型的一种实现方案是，为每一个参与融和的单模态训练一个深度波尔兹曼机（DBM），然后在这些 DBM 之上增加一个额外的隐藏层给出融和后的联合表示（图 6）。上述融和过程，如果是无监督的，则可视为特征学习过程，输出的即为特征层融和的结果特征；如果是有监督的，输出的即为决策层融和的最终分类结果。(更多关于深度学习的内容，请参见报告的《语言表示与深度学习》一章。)

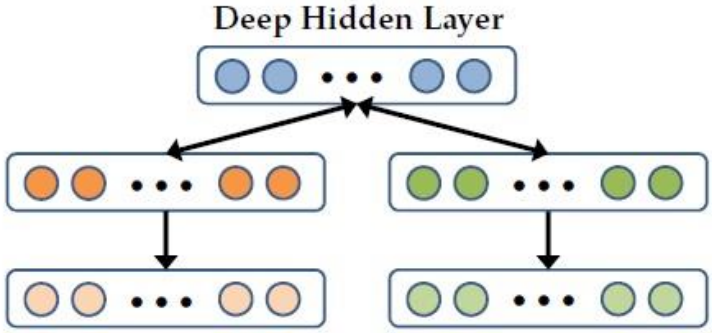


图 6 多模态深度学习模型

基于图像的字幕生成问题也可以用深度学习方法来解决，即采用所谓交叉模态特征学习。由于字幕与图像之间存在内在的多模态关联关系，因此，运用上述多模态深度学习模型，可以学习到融和的特征（也称为共享特征表示，Shared Representation），那么理论上该模型应该支持训练一个模态，而测试另外一个模态，且仍能获得好的分类效果。

3.3 多模态语义表示

所谓语义表示是指在计算机系统中对语义的形式化描述或表达。因此，多模态语义表示是指，人机交互过程中不同模态之间交互语义信息的形式化描述。对语音、文本和视觉信息进行处理、理解和生成的多模态系统，必然会涉及到多模态信息输入输出过程中的语义表示问题。由于多模态信息的异构性，在多模态系统中，一种模态的输入信息需要先映射到一种语义表示，当在另外一种模态进行输出时，再将这种语义表示映射到指定模态进行输出。

多模态语义表示的发展是基于应用驱动的，许多多模态应用或实验系统，都提出了自己的语义表示方案，其中采用比较多的是基于框架语义学（Frame）和 XML 语言的表示方案。在特定的多模态应用系统中，语义表示问题可理解为，基于框架语义学对应用系统语义表示空间的 XML 编码。

上述多模态语义表示方法事实上是对语义的显式表示，在基于机器学习，特别是基于深度学习的多模态系统中，语义表示常常以模型的形式存在，这种语义表示可理解为隐式的多模态语义表示。

4. 技术展望与发展趋势

多模态信息处理研究的发展，受到来自移动智能终端、可穿戴设备、物联网、自然语言处理、人机对话、仿真机器人、信息检索、模式识别、情感识别、深度学习、大数据、认知科学等工作的促进和推动。随着移动智能终端、可穿戴设备、物联网的普及，人机交互的信息从传统的文字、声音、影像，发展到位置、重力加速度、睡眠、运动等人体信息，共生、

共现的单模态信息种类大大增加；由于人的感知和认知机理的多模态本质，自然语言处理、人机对话、仿真机器人、信息检索、模式识别、情感识别等研究领域，越来越多地采用多模态信息处理的方法和思路，取得了许多具有实际应用价值的成果，从而大大提升了多模态信息处理的能力；深度学习、大数据的兴起，即为多模态信息处理提供了新的技术手段，也为多模态信息处理提供了更丰富的数据来源。

第十九章 医疗健康信息处理研究进展、现状及趋势

1.任务定义、目标和研究意义

医疗健康信息处理（Clinical and health information processing）主要利用信息技术对与人类医疗健康相关的数据进行处理，挖掘蕴含在这些数据中的有用信息和规律，以服务于医学研究、临床诊疗、公共卫生决策、个人健康咨询等各个领域。

医疗健康信息处理研究的目标就是针对不同类型的医疗健康相关数据，建立有效的信息抽取和利用的方法、模型和系统，充分挖掘数据潜力，弥补优质医疗资源严重短缺、大大降低误诊率，为实现智能化医疗、提高人类医疗健康服务水平提供必要支撑。

健康是促进人类发展的必然要求，医疗健康相关领域的发展关乎社会和谐和民生幸福，长期受到全人类的广泛关注和世界各国政府的大力支持。在“人、机、物”三元世界高度融合的“一切皆可数据化”时代，医疗健康大数据首当其冲成为许多国家战略发展计划的重要内容之一。在美国，近两年来先后公布了基于大数据的“精准医疗”和“新登月”计划。我国的“十三五”规划以民生建设为主线，明确提出加快全民健康型社会建设和提高老龄健康预期寿命的目标。国务院 2015 年发布的《关于印发促进大数据发展行动纲要的通知》中指出“在公共服务大数据工程中要构建医疗健康服务大数据”，同时已开始着手成立中国精准医疗战略专家组，设立了多个国家重要科技研究计划专项。要发展医疗健康大数据，首先需要发展医疗健康信息处理技术。从理论上讲，医疗健康信息处理涉及到医疗、生物、计算机、统计等多个学科，是一个典型的多学科交叉研究方向，相关技术的发展有利于相关学科的发展，特别是医疗健康相关领域的发展，为智能化、精准化医疗提供理论支撑。因此，开展医疗健康信息处理方面的研究具有非常重要的理论意义。从应用上讲，人类社会离不开医疗健康服务，迫切需要医疗健康信息处理技术来进一步提高医疗健康服务效率，提升医疗健康服务水平，保障人们生活质量。因此，开展医疗健康信息处理方面的研究具有很大的实际应用价值。

从技术的角度看，医疗健康信息处理是建立在医疗健康大数据之上的一门人工智能科学，是人工智能在医疗健康相关领域深度发展的基石。因此，医疗健康信息处理是“智能医疗”和“精准医疗”战略实施中不可或缺的重要技术。目前，医疗健康信息处理在某些方面（如医学图像处理）已经取得了一定的发展，但在诸多其他方面（如医学文本处理、多模态数据处理等）才刚刚开始起步。毫无疑问，医疗健康信息处理仍然面临诸多挑战，需要政府、科研院校、医疗机构和科技企业一起齐心协力共同解决。

本部分对医疗健康信息处理研究的主要内容、遇到的关键科学问题、面临的主要挑战，以及当前采用的主要技术、现状和未来发展趋势，做简要介绍。

2.研究内容和关键科学问题

医疗健康信息主要来源于在医疗服务过程中产生的临床医疗数据、在人们日常生活过程中通过移动设备等记录的个人健康数据、以及由其他媒介（如医学文献、社会媒体）记录的医疗健康相关数据。这些数据通常以以下四形式存在：1）结构化数据（如检验检查记录）；2）文本（如入院记录、出院记录、病程记录、医学文献等）；3）图形（如心电图、脑电图等）；4）图像（如超声图像、核磁共振图像等）；5）新媒体数据，如微博、微信等。因此，理论上来说，医疗健康信息处理研究的涉及到对上述各种来源的所有四种类型数据的处理技术。其中单独的图形、图像处理技术的研究历史久远，技术已经非常成熟，结构化数据处理则相对容易，因此在医疗大数据时代，研究工作的研究工作的重心逐渐转移到对于大规模非结构化医疗文本信息的处理，以及将文本信息与结构化信息、图形图像信息的联合处理上来。非结构化的文本信息由于在描述患者病因、症状，医生诊断结论、治疗方法，以及患者长期

的病程发展等过程中都起到关键作用，但国内外在该方向上的研究都起步较晚，且仍然面临着巨大挑战，因此这部分主要讨论医疗文本信息处理，以及医疗文本与健康数据的处理技术，及其与图形图像数据、其他数据之间相互融合等技术的主要研究内容和关键科学问题。概括起来，主要包括以下研究内容和科学问题：

2.1 医疗知识图谱构建

医疗健康领域对于知识的依赖远比大部分文本信息处理应用要强，这就使得建立更加丰富准确的医疗健康知识体系（知识图谱）成为首要的研究内容。和传统的专家知识库不同的是，医疗健康知识图谱需要从我们上文提到的大规模多模态医疗健康信息中自动抽取目标知识，并建立起各种知识之间的关联，涉及到的关键技术包括医疗实体识别以及基于多模态特征的实体链接。 医疗健康实体识别的主要挑战首先在于它涉及的化学、生物、医学等多个差异较大的学科领域；其次，在医疗文本中的用语习惯也和我们的日常用语习惯差异巨大，这使得我们针对通用的文本信息处理任务所构建的大量标注数据库难有用武之地，而直接从医疗领域获得的各种类型的加工数据的规模远比通用领域要少，这也让进年来基于深度学习的端到端学习方法难以有效发挥其作用。对于实体链接任务来说，其挑战一方面是对医疗实体链接准确性具有极高要求，因为这直接决定了知识库的可用性；另一方面，医学实体之间的关系通常要考虑到患者健康数据、症状表现、多模态检测结果、各种药物的化学成分和效用分析、诊断和治疗手段等多方面信息的交互，这就大大增加了医疗健康实体链接的难度。

2.2 辅助诊疗技术研究

智能辅助诊疗是实现患者导诊、医生辅助决策、远程诊疗等智慧医疗主要任务的核心技术，主要内容是通过分析每个患者当前的主诉、病史、家族病史等特征，在实际诊疗前进行智能推理与判断，给出导诊建议，并在各种医疗检测结果出来后，结合检测结果为医生提供疾病的诊断建议，甚至直接给出诊断结果。为了能够进行自动的推理和判断，首先要解决的关键问题是从大量的诊疗实例中进行学习总结的技术与能力，这也是当前医疗文本信息处理所做的最重要的工作之一，比如通过从病历中挖掘疾病和用药的关系，来发现药物的新的疗效、副作用、多种药物之间的相互作用等，其基本技术包括医疗信息抽取、非结构化序列标注与关联关系挖掘等；其次，在医疗知识图谱与医疗信息挖掘的基础上，如何建立包含文本、图形、图像、检验数据等多模态特征的知识推理能力则是当前智能诊疗所面临的另一个关键问题。

2.3 基于大数据的流行病学研究

流行性疾病的分析预测是保障公众健康和政府公共卫生管理的重要任务，因此，开展更加智能化的、基于医疗健康大数据和新媒体分析的流行病学分析方法的研究是医疗健康大数据处理的重要研究内容。流行性疾病分布和发展规律分析主要是为了对流行性疾病进行有效的预测和预警。流行性疾病发展规律分析是当前世界范围内广泛关注的热点和难点问题。研究流行性疾病发展规律分析技术的主要目的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，进而提高流行性疾病预防控制工作的主动性和预见性。其中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。

3.技术方法和研究现状

医疗健康信息处理涉及的技术方法范围广泛，针对上述研究内容和关键问题，主要涉及的技术方法包括医疗实体识别、医疗实体链接、医疗文本挖掘、知识推理技术、基于大数据的流行病分析技术等，本节将就上述技术的方法与现状进行介绍。

3.1 医疗实体识别

随着医疗信息技术的发展，出现了大量可用的电子健康档案。这些数据不仅能够支持电子化的临床系统，同时供临床和转化医学方面的研究使用。利用电子病历的一大挑战在于：大量蕴含于临床记录中的有用信息无法被依赖于结构化数据的电子化的临床系统直接使用。为充分利用电子病历数据，能够从原始文本中抽取结构化信息的自然语言处理技术在临床医学领域受到了广泛关注，许多临床自然语言处理技术被实际应用系统所采用。

作为临床自然语言处理的一项基本任务，临床医疗实体识别一直备受学术界和工业界关注。在过去二三十年内，研究者们开发了大量针对不同类型临床记录的临床医疗实体识别系统。早期的临床医疗识别系统大多利用临床医疗专家人工构建的字典或规则来识别临床医疗实体。在过去几年里，随着可用的标注临床医疗语料的增多，研究者们开始使用机器学习算法来识别临床医疗实体，并组织了一些关于临床自然语言处理任务的国际公开竞赛来促进相关技术的发展。比如，i2b2 分别在 2009 年和 2010 年组织了关于临床医疗实体识别的国际公开评测。ShARe/CLEF eHealth Evaluation 实验室在 2013 年组织了一次关于临床医疗实体识别和标准化的自然语言处理国际公开评测，2014 年和 2015 年分别在扩充了 2013 年的 ShARe/CLEF 评测语料库的基础上，将临床医疗实体识别和标准化任务引入到国际语义评测中。在这些评测中，被要求识别的临床医疗实体类别从仅包含药物逐渐扩展到包含药物、问题、检查和治疗，实体表现形式也从仅考虑连续临床医疗实体扩展到同时考虑连续和非连续临床医疗实体。2009 年的 i2b2 NLP 评测任务仅考虑了电子病历中连续药物的识别，2010 年的 i2b2 评测任务则考虑了连续药物、问题、检查和治疗的识别，2013 年的 ShARe/CLEF 评测任务则首次同时考虑了电子病历中连续和非连续临床医疗实体识别问题。这些评测主要面向英文电子病历，面向其他语言电子病历的临床医疗实体识别研究相对较少，主要原因在于缺乏公开可用的标注数据。

连续临床医疗实体识别是一个典型的序列标注问题，一些经典的序列标注算法如条件随机场、结构化支持向量机等在上述的几次评测中得到了广泛应用，并取得了不错的识别效果。2016 年，新提出的基于深度学习的序列标注算法也被用于连续临床医疗实体识别。然而，非连续临床医疗实体的识别不再是一个纯粹的序列标注问题，而是一个新的研究问题。在 2013 年的 20 多个 ShARe/CLEF 评测参赛队伍中，仅有两个参赛队伍考虑了非连续临床医疗识别的问题，他们分别采用了基于规则和基于机器学习的识别方法。其中，基于机器学习的识别方法取得了明显更好的效果。该方法的关键在于如何在统一框架下表示连续和非连续临床医疗实体，并根据统一表示形式设计相应的识别算法。近两年来，人们也对上述两类方法提出了一些改进的统一表示方法及相关算法，并取得了一定的性能提升。

3.2 医疗实体链接

一般而言，实体链接主要解决以下三类问题：1）歧义性：相同实体提及对应多个标准实体概念；2）多样性：一个标准实体概念有多种不同的提及形式；3）缺失性：实体提及对应的实体概念没有在给定的标准知识库中出现。在通用领域，歧义性现象较为常见，但在医疗领域，多样性现象最为常见，实体链接的研究重点则集中在解决实体多样性的问题。现有医疗实体链接方法通常利用实体上下文信息来进行实体消歧，利用实体提及与实体概念之间的相似度来解决实体多样性问题。

医疗领域实体链接方法的发展是由近年来的几次国际公开评测来推动的，具有代表性的评测有 2013 年以来的 ShARe/CLEF、2014 年以来的 SemEval 等，主要任务是找到临床医疗文本中医疗实体在一体化医学语言系统中“医学术语系统命名法－临床术语”分支上的编码。所提出的实体链接方法大多基于医疗术语提及与标准医疗术语的相似度计算。如在 2013 年的 ShARe/CLEF 评测中，几乎所有的参赛系统都通过计算疾病和症状提及与医学术语系统命名法中的标准术语之间的相似度来得到相应编码的。主要的相似度计算方法是向量空间模型和编辑距离，就编辑距离而言，亦出现了一些改进的编辑距离计算法方法，具有代表性的工作是通过从给定疗术语提及到标准医疗术语之间的映射样本中学习常见的词或短语编辑模式来对编辑动作进行限定，在此基础上刻画它们的相似度。另外一类是基于规则的方法，如基于采用多路筛选的方法，这些方法在生物以及医疗领域的公开数据集上取得了较好的性能。

虽然基于字符串相似性和模式匹配的方法针对单医疗实体有较好的效果，但是在关联性实体链接上效果并不好，并且没有充分的利用上下文信息，因此，也有学者尝试使用机器学习的方法来解决关联性医疗实体链接，大部分人的研究致力于建立不同的过滤器选出候选实体，如采用条件随机域、逻辑回归等方法。也有研究者将从候选实体中选出正确的实体看做排序问题，使用学习排序的方法来解决。

上述研究工作主要面向英文电子病历，面向中文电子病历的实体标链接研究还不多见，仅有少部分学者在医疗术语标准化方面做了一些尝试。如采用向量空间模型对医保数据中的疾病名按照 ICD-10 分类体系中的大类进行分类，或者采用信息检索的方法在检索的过程中从词、字符和拼音三个方面对标准医学术语进行索引，当大部分研究并没有完成疾病名到 ICD-10 标准疾病术语之间的映射问题。

3.3 医疗文本挖掘技术

医疗文本的挖掘，主要目标是从中抽取并建立起多实体之间的关联，从而为基于医疗文本的学习和推理系统打下坚实基础，同时，这也是建立和完善医疗知识图谱的重要技术。显然，前面提及的医疗实体识别和实体链接，都是医疗文本挖掘的基础支撑技术，而在此之上，这部分重点介绍的是各种医疗实体之间的关系抽取。关系抽取，是指在文本语料中抽取实体之间蕴含的语义关系。关系抽取可以通过多种方式划分，如抽取的关系是否被预定义、二元关系还是多元关系、是否是有监督学习等。这里我们按照学习方法分类，分为有监督学习、半监督学习、远监督学习三类。

1)有监督学习关系抽取:有监督学习关系抽取通常抽取句子级别之间类标的关联关系，如"Apple CEO Steve Jobs said.." 中抽取出 (SteveJobs, CEO, Apple) 三元组。通常典型的关系抽取数据集采用 ACE2004、MUC-7 以及生物医学的一些数据集。有监督方法通常采用分类正负样本来解决这一问题。典型的特征是上下文信息和词性标注，实体之间的依存路径，命名实体标记以及实体距离，这种方法通常能够得到较高的质量和显著的负样本效果。然而，生成标注集合的代价很高，并且难以加入新的关系，也无法将其方式简单泛化到其他领域。

2）半监督关系抽取：其基本思想是采用泛型算法，一开始使用一些初始种子模板，然后从文本中抽取符合这些种子模版的新模板，选取其中的前 k 个加入种子模板中重复这一过程，典型的方法如 TextRunner，该方法近乎无监督学习，其所抽取的关系也并不限定在固定词表上，没有标注数据，同时抽取的文档也几乎都是无标注无结构的文本，其算法采用启发式方法在一个初始文本中自标注学习，通过语法分析建立一定的规则，自动生成正负样例来训练分类器来进行自训练。

3）远监督关系抽取：其思路是通过现有的知识库和未标注的文本来生成新的样例，找到未标注文本中的相关实体对的位置并假定这个关系可能被该段文本所表示。远监督关系抽取的一个典型架构是收集大量在统一语句中共现的实体对，如果这两个实体对之间有关系，那么最简单的假设就是所有这样的句子都能够表示这两个实体之间的关系。基于这样的假设，通过抽取所有共现的句子，聚合典型特征（如词法特征、句法特征、命名实体标记以及共现特征等）后采用多类标逻辑回归分类知识库中定义好的关系。然而，这种假设过于粗糙，绝大多数出现在同一句话中的实体对实际上是没有关联的。因此，有研究者改进了这个假设，

并通过因子图的方法寻找可信句子。进一步的改进包括假设包含共现的实体对的句子中至少有一个能够表示该实体对的关系，两个实体之间可以表示多种关系，从而将该问题定义成多实例的多类标分类问题。远监督关系抽取能够有效利用网络资源，并且无需监督标注，因而能够有效的泛化到各个不同的领域，其主要缺点是需要高质量的实体匹配技术来对应正确的实体，关系表达假设也具有局限性，并且难以生成有效的负样本。目前，医疗文本的关系抽取已经有一些成型的方法或系统，如 DIADEM，DeepDive， McCallum 等，从而为进一步的医疗文本分析挖掘打下了良好基础。

3.4 医疗健康知识推理技术

医疗健康知识推理的主要目标之一是对临床医疗医疗预诊、诊断等提供决策建议和帮助，提升医疗诊断的准确率和效率，降低误诊。目前国际上针对临床决策等提供的支持主要基于现代临床决策支持系统（Clinical Decision Support， CDS），其目标是为医生提供与其患者医护管理相关的信息，这样当医生面对具体医疗案例时，可通过查询 CDS 来获取关于诊断、测试以及治疗的方案。为了推动该领域的研究，著名的国际信息检索评测会议 TREC 设置了 TREC-CDS 评测，该评测通过询问医疗相关的三个基本问题，并基于每个系统检索出来的答案来评价 CDS 系统，这三个基本问题是：（1）“诊断结果是什么？”（2）“应该进行什么试验？”（3）“应该采用哪种治疗方法？” 事实上，TREC-CDS 在任务设置上做了比真实应用场景更理想化的简化处理，使用 30 组短医疗案例报告的形式作为医疗记录表示，每个报告都是一个具有挑战性的医疗案例，并提供了以下数据集，包括有（1）一个描述与患者医疗案例相关部分电子医疗记录的叙述；（2）该医疗案例的摘要；（3）三个基本医疗问题。任务目标是从 PubMed Central 中筛选出可能包含正确答案的文献列表，并根据置信度进行排序，然后将每个医疗案例描述或者摘要中对三个基本医疗问题的描述作为答案。显然，相比于直接给出确定答案，检索出相关的医疗文档要来得简单，同时经过上述处理后的问题比真实的医疗环境进一步简化，但即便如此，目前的系统也还远未获得令人满意的效果。从 2015 年以来，基于智能问答系统 Watson 的认知计算平台， IBM 实现了目前为止最为接近产业化应用的医疗健康知识推理系统 Watson Health，通过与 MSK 癌症中心的合作，该系统能够在输入一个未经诊断的患者病例后，通过分析数千临床病例和相似的患者记录，并结合医学教材等知识，来对患者的疾病推断出一个置信度较高的可能疾病列表，虽然仅仅面向 MSK 提供的癌症等疾病，但这无疑是当前人们在医疗健康知识推理技术领域所取得的一个巨大进展，这同时也是对在该领域几乎刚刚起步、以中文医疗文本为对象的国内医疗健康知识推理技术的重要启迪。

3.5 流行病分析技术

传统流行病学中对流行性疾病预测分析的方法种类很多，分类方法各异。按预测时限长短，可以分为短期（一年以内）、中期（三年以内）、长期（三年以上）。按照数据来源类型，可以分为时间预测，时空预测。时间预测是根据过去一段时间检测变量值的大小，利用统计模型预测伟大监测变量值的大小；时空预测，则是综合利用历史数据的时间和空间信息，利用时空模型进行预测。而近年来，随着大数据和社交媒体的兴起，人们在新的媒体空间中的活动或者讨论等，也成为了流行病预测的重要分析因素，并在流行病预测中扮演着越来越重要的角色。为此，这部分主要介绍几种基于大数据的流行病预测方法。

1）网络和社交媒体大数据分析方法：某种流行性疾病的搜索结果在短期内激增，这可能准确预示着此种疾病将会暴发。例如，在流感暴发季节，关键词流感、勤洗手、带口罩、流感疫苗等会高频率出现。人们也会通过微信、微博、twitter 等聊天工具反映用户本人、朋友是否感染流感，或者与流感相关的信息等。比如，美国科学家将 2004-2009 年查询所得的不同国家和地区的流感估算结果与官方的流感监测数据进行对比，发现 Google 流感搜索引擎查询所得到的估测结果与历史流感疫情非常接近，并且可以赶在政府和流行病学专家之前两个星期提前预测到流感暴发的出现。在国内，百度也依据相似的原理推出了疾病预测系

统。研究者通过对 Twitter 数据流加以过滤，留取与流感相关的信息，并为这些信息加上地理位置标签，同样成功推断出了美国哪些地区出现了流感暴发的初期症状，进而提前预测到某个地区流感即将到来。当然，虽然应用这些社交媒体或者信息检索系统监测手段能比传统监测方法能够提前预测到流行性疾病的暴发，虽然它不能完全取代传统监测系统，但是已经成为了疾病监测预警手段的一种重要的扩展。

2）医疗系统大数据分析方法：症状监测是指通过长期连续系统地收集与所监测的疾病相关的一组临床特征（症状）和相关社会现象的发生频率来获取传统公共卫生监测不能提供的疾病防控信息，及时发现疾病在时间和空间分布上的异常聚集，以期对生物恐怖袭击、新发流行性疾病、原因不明流行性疾病及其他聚集性不良公共健康事件的暴发进行早期探查、预警和快速反应的监测方法。目前在症状监测方面应用最成熟的应当属于急性迟缓性麻痹病例监测，中国自 2000 年被世界卫生组织证实无脊灰野病毒传播以后，在全国范围内建立了包括 14 种疾病在内的急性迟缓性麻痹病例主动监测系统用以早期发现输入性脊髓灰质炎病例，2011 年 7 月，监测人员通过该系统在新疆和田地区成功的发现急性迟缓性麻痹病例增多，经实验室确证系源自巴基斯坦的输入性脊髓灰质炎野病毒传播，并通过 5 轮强化免疫使疫情迅速得到有效的控制，最终成功的阻断了野病毒的传播。为了有效预防和控制西尼罗病毒，以美国疾病预防控制中心牵头，建立了死鸟监测体系来早期发现西尼罗河病毒的活动，监测内容包括定期报告和分析病鸟或死鸟的情况以及有针对性的对鸟类是否感染西尼罗病毒进行实验室检测。2002 年，美国共有 582 个市县出现了西尼罗病毒感染病例，其中有 543 个通过本项监测获得了西尼罗病毒的早期活动信号，死鸟出现的时间比人类病例的出现平均提前了一个多月。英国科学家通过对感冒和流感呼救电话进行监测，在利用泊松模型建立预警阈值的基础上，与临床上流感样病例和实验室确诊病例进行对比，回顾性研究发现能够提前 2 周提前预警到流感暴发，前瞻性研究发现能够提前 6 天预警流感暴发。

3）环境气候及其他大数据分析方法：无论是太阳黑子活动、宇宙线环境大的增强，还是拉尼娜现象造成的气候异常等，都被证明与不同种类的大范围流性病爆发发之间具有一定的对应关系。此外，卫星图像也被广泛用于帮助人们早期预测什么时候会暴发流行病。一种新出现的流行性疾病接下来会在哪里及在何时出现是一个难以解决的问题，以往预测一种新出现的流行性疾病的传播模型一般侧重于地理上的距离，或者采用一些结合了流动性与流行病学数据的现代、复杂的疾病蔓延模型，但是这些预测方法应用于现代社会效果并不是很好。研究人员基于不同地区之间的迁移概率，提出了不同于地区之间物理距离的“有效距离”的概念来分析流行性疾病的蔓延，并据此研发了一种可预测某种新发的流行性疾病会在哪里以及在何时发生的“反应扩散模型”，在新模型中通过计算航空运输网络的运输强度来表征地区间的有效距离，他们应用新数学模型分析了 2003 年的 SARS 疫情和 2009 年甲型 H1N1 流感的疾病暴发路径并得到很好的验证。

如上说述，流行性疾病预测预警工作的研究方法和理论已经向学科交叉和综合应用转变，并且取得了长足的进步，逐渐走向成熟。这些新方法和新理论的应用有利于完善和充实目前的流行性疾病监测体系，提高今后流行性疾病预防控制工作的预见性和主动性，从而更好的维护国家正常的公共秩序和人民群众的健康。

3.6 技术现状总结

从上面的阐述可以看出，针对不同形式的医疗健康与生物数据，处理方法在不断完善，但仍然存在许多急需解决的问题。首先，用于支撑一些医疗健康信息处理关键任务（如数据规范化、医学语义解读）的大规模标准化知识库（知识图谱）还不完善，在我国比较匮乏，严重阻碍了相关技术的发展。其次，针对某些医疗健康信息处理任务的研究还没开始，针对某一形式的医疗健康信息的处理技术(如中文临床医疗自然语言处理技术)还有待继续加强。此外，针对不同形式的医疗健康的处理技术往往相互独立，缺乏融合。实际上不同形式的数据之间是相互关联的，如电子病历中的文本描述与其相关的检验检查产生的结构化表格数据、图形和图像数据之间显然存在很强的相关性，充分利用不同形式数据间的关联关系有望提高各种类型数据的处理能力。最后，在实际应用过程中，现有很多医疗健康信息处理技术仍需要大量的人工干预，人工干预的信息往往没有及时反馈给信息处理系统，如何利用人工反馈

信息来提高信息处理系统性能，是一个值得尝试的方向。

在医疗健康领域，利用信息技术提高医疗健康服务智能化程度是医学发展的大趋势之一，从 IBM Waston 在医疗健康领域某些方面的表现来看，现有研究已经取得了一些具有里程碑意义的研究成果。尽管如此，有待深入研究和探索的问题及相关理论和技术还还多。

4. 技术展望和发展趋势

从医疗健康信息处理研究的趋势和技术现状来看，以下研究问题将成为未来医疗健康信息处理研究必须攻克的堡垒：

大规模标准化医疗健康知识库（或知识图谱）的构建：医疗健康信息处理中，经常涉及到标准化问题，标准化是数据后续使用的基础，完成标准化任务的一个关键因素就是大规模标准化的知识库（或知识图谱）。该知识库要求具有统一、可扩展的知识框架，含有丰富的医疗健康词条和关系。如何根据领域知识设计合理的知识框架是一个全新的研究问题。在知识框架下，准确高效地自动抽取知识词条和关系来对知识库进行填充是另一个需要研究的问题。考虑到医疗健康领域应用较多，可以根据国家战略需求开始进行研究。

中文临床医疗自然语言处理：自然语言处理技术具有很强的语言相关性，已有尚不成熟的英文临床医疗自然语言处理技术并不能直接用来处理中文医疗文本。而中文临床医疗自然语言处理的相关研究才刚刚起步，无论在语料库资源还是理论技术方面都很匮乏。在国家大力支持医疗健康大数据发展的时期，中文临床医疗自然语言处理技术作为基础支撑技术之一，必将是大家关注的一个重要研究方向。

多模态医疗健康信息融合：不同来源不同形式的数据之间往往具有强相关性，通过对这些数据在统一框架下进行相互表示和建模能够把一种形式数据中的语义信息迁移到另一种形式数据的语义空间中，提高不同形式数据处理的性能。多模态数据深度融合是大数据发展的目标之一，医疗健康领域的多模态数据融合也将是医疗健康大数据的一个热点研究方向。

交互式医疗健康信息处理：通过人与机器的交互过程逐步提高机器的智能程度是人工智能发展的一个重要研究方向。在医疗健康领域，信息处理系统智能化程度相对较低，而人机交互频繁，交互式学习具有很大的潜力。

目前，越来越多来自自然语言处理领域、医疗健康行业、大数据分析领域的研究者开始关注这一方向，随着研究工作的不断深入和相关技术的快速发展，我们有理由相信，在具有广泛产业化应用前景的医疗健康领域，医疗健康信息处理将得到相当程度的发展，将很大程度上提高医疗健康领域的智能化程度。

第二十章 少数民族语言文字信息处理 研究进展、现状及趋势

1.任务定义、目标和研究意义

我国有55个少数民族，人口1.08亿，占全国8.4%，居住的面积约占全国总面积的63.7%。中国境内的语言非常丰富，五十五个少数民族中，除回族、满族以外，其他53个民族都有自己的语言。据统计，在中国，少数民族正在使用的语言有72种左右，已经消亡的古代语言更是不计其数。在55个少数民族中，除回、满两个民族用汉文外，蒙古、藏、维吾尔、哈萨克、柯尔克孜、朝鲜、彝、傣、拉祜、景颇、锡伯、俄罗斯12个民族各有自己的文字，这些文字大都有较长的历史。其中中国八省区蒙古族使用一种竖写的拼音文字，居住在新疆的蒙古族还使用适合卫拉特方言特点的拼音文字托忒蒙古文，蒙古国、俄罗斯布里亚特、卡尔梅克加盟共和国除用基里尔文字以外还在使用传统蒙古文。云南傣族在不同地区使用四种傣文，即傣仂文、傣哪文、傣绷文和金平傣文，傈僳、佤、壮、白、瑶等民族都使用两种以上的文字。如果把我国各民族在历史上使用过的文字计算进来，更是名类繁多，达到几十种之多。这些现行的和历史上的文字承载着中华民族几千年的文明，是国家和民族的财富和资源，是我国中文信息处理所必须重视的重要领域和内容。

自从上世纪80年代初开始，国内外开始进行少数民族语言文字信息处理研究。我国率先进行了民族文字编码标准制定、语料库建设、机器翻译等研究，并取得了可喜的成绩。进入新世纪后，我国民族语言信息处理更是得到了长足发展，几乎覆盖自然语言处理的所有领域。民族语言信息处理也成为了中国“一带一路”战略的重要组成部分。它关系着我国在这一领域的话语权和主导权，也关系到国际市场竞争和国家信息安全。跨境语言和文字的信息处理更是如此。

西方发达国家也是从上世纪开始研究我国少数民族语言信息的处理，包括基础理论研究和应用技术开发，旨在占领这一领域的理论和技术制高点。尤其是在藏、蒙、维哈克、朝鲜文方面的研究，国内外一直处于竞争状态。其中也不乏反华势力的参与和技术壁垒的设置。因此，民族语言信息处理不仅仅是一个学术研究问题，更是涉及到政治、经济、学术多个领域的，关乎民生和国家利益的问题。

我们的目标是保持我国在这一领域的传统优势和战略制高点，使我国少数民族语言信息处理健康、高速发展，为全国各民族共同繁荣和发展提供技术保障。

2.基础研究

少数民族语言信息处理几乎涵盖自然语言信息处理的所有领域，主要包括：基础理论研究、编码标准的制定、大型语言资源库的建立、应用系统开发等。下面就这些内容进行简要论述。

2.1 编码标准的制定

字符编码(Character encoding)是一套法则，使用该法则能够将自然语言字符的一个集合（如字母表或音节表）与其它东西的一个集合（如号码或电脉冲）进行配对。一般指用数字、字母、文字按规定的方法来代表特定的信息。

目前我国蒙古文、藏文、维吾尔文、哈萨克文、柯尔克孜文、朝鲜文、彝文、壮文、女字、水字等现行文字和八思巴文、女真文、契丹小字等古文字也已被收录到ISO/IEC 10646国际标准。

我国民族文字从其渊源、书写特征、符号类型等诸多方面各有明显的特点。例如蒙古文是竖向连写，字与字之间没有空格，直到形成一个整词之后才会有空格，而且蒙古文的是从左到右换行。而维哈克文虽然横写，但其书写方向却是从右到左，这些特点在计算机通用系统中需要另行处理。藏、彝、朝等文字也是各有千秋，需要更复杂的技术才能实现。历史上的古文字就更是千奇百怪，错综复杂。

在标准规范方面的重要成果有：

1）内蒙古大学蒙古学学院正制定“信息技术 信息处理用蒙古文相关标准”，涉及到13 项国家标准草案；

2）潍坊北大青鸟华光照排有限公司承担少数民族文字信息技术国家标准，已正式发布61 项：

3）延边大学研制的“信息技术 朝鲜文通用键盘字母数字区的排列布局”、“信息技术 基于数字键盘的朝鲜文字母分布”两项国家标准审查通过。

ئۇيغۇر يېزىقى ئېلىپبەسىنىڭ يۇنىكود سىلىشتۇرما جەدۋىلى

رەت نۇمۇرى	ئاساسىي رايۇن	كېڭەيتىلگەن رايۇن	ئون ئالتىلىك	ئونلۇق
1	ا 0x0627–1575	ﺍ	FBEA	64490
		ﺎ	FE8D	65165
		ﺋﺎ	FBEB	64491
		ﺎ	FE8E	65166
2	ە 0X06D5–1749	ﻛ	FBEC	64492
		ﻜ	FBED	64493
		ﻩ	FEE9	65257
		ﻪ	FEEA	65258
3	ب 0x0628–1576	ﺑ	FE8F	65167
		ﺒ	FE91	65169
		ﺑ	FE92	65170
		ﺐ	FE90	65168
4	پ 0x067E–1662	ﭘ	FB56	64342
		ﭙ	FB58	64344
		ﭘ	FB59	64345
		ﭗ	FB57	64343
5	ت 0x062A–1578	ﺗ	FE95	65173
		ﺘ	FE97	65175
		ﺗ	FE98	65176
		ﺖ	FE96	65174
6	ج 0x062C–1580	ﺟ	FE9D	65181
		ﺠ	FE9F	65183
		ﺟ	FEA0	65184
		ﺞ	FE9E	65182
7	چ 0x0686–1670	ﭼ	FB7A	64378
		ﭽ	FB7C	64380
		ﭼ	FB7D	64381
		ﭻ	FB7B	64379
8	خ 0x062E–1582	ﺧ	FEA5	65189
		ﺨ	FEA7	65191
		ﺧ	FEA8	65192
		ﺦ	FEA6	65190
9	د 0x062F–1583	ﺩ	FEA9	65193
		ﺪ	FEAA	65194

图 1. 维吾尔文编码国际标准

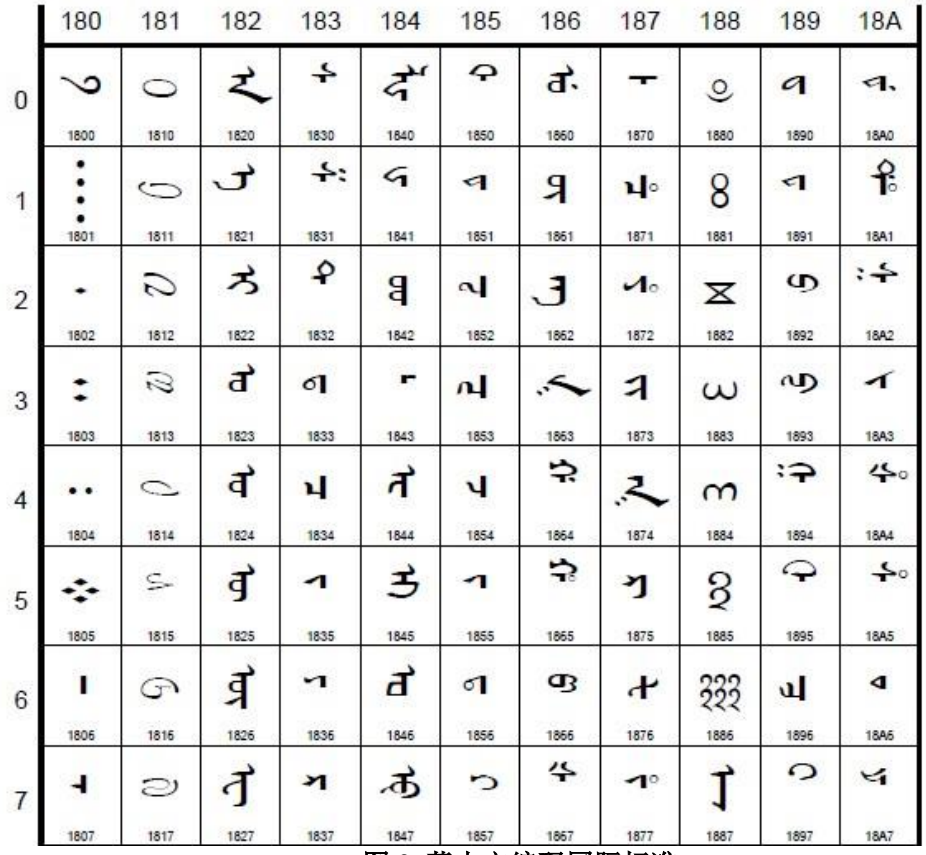


图 2.蒙古文编码国际标准

2.2 基础资源建设

语言资源是指语言信息处理用到的各种语料库和语言数据库，以及各种语言词典等。随着我国现代化、信息化的加速发展，社会对语言实际需求的加大，以及语言功能的变化，人们对语言的认识正在发生深刻的变化，逐渐从视语言为问题转向将语言看作资源，进而发出保护建设和开发利用语言资源的呼吁。这一认识的变化有着重要的意义，对我国语言生活和语言文字工作将产生积极影响（陈章太）。语言资源也是国家的重要资源之一，当然少数民族语言也不例外。

目前民族语言资源匮乏是制约民族语言信息处理的主要瓶颈，近年在各类会议和项目规划中我们一直呼吁和建议各个地区和研究机构重视资源库建设，这些建议得到了大家的肯定，并已成为共识。近几年来民文资源库建设工作方面已经有了大幅度的发展。下面介绍这方面的主要标志性成果：

（1）大规模藏语单语语料库（约 13 亿字）、藏汉篇章对齐语料、藏汉句词对齐语料、藏语分词标注语料以及 1 万词条的情感词词典，为藏汉双语翻译奠定了语料基础（新疆大学）；

（2）1000 万词蒙古语单语语料库；汉蒙、蒙汉双语语料库已达到 80 万句对（内蒙古大学蒙古学学院）；

（3）维哈柯文软件构件库、文本处理工具库以及语言资源库；

（4）藏语分词和词性标注语料库，共收录 7.2 万句藏文文本；藏语句法树库（4000 句）；大规模文本语料库（包含藏文文本 40.52 万篇，合计 859 万句；

（5）维吾尔文文本 23.08 万篇，共计 580 万句（中科院软件所）；

（6）大规模的维汉、蒙汉、藏汉平行语料库和翻译词典（中科院计算所）；

（7）基于哈语文本语料库，实现了词干提取、基本词和兼类词的词级标注、基本短语

的识别、哈语浅层句法分析算法并构建哈语短语库；
（8）汉-东盟语种在线语料库；
（9）蒙古语文献语料库及其相关管理平台。

2.3 词法与句法分析

词法分析（lexical analysis）是计算机科学中将字符序列转换为单词（Token）序列的过程。进行词法分析的程序或者函数叫作词法分析器（Lexical analyzer，简称 Lexer）。我国少数民族语言分别属于汉藏语系、阿尔泰语系、南亚语系、南岛语系、印欧语系等不同语系。由于语言类型的不同，少数民族语言的构词构形各有特点，有的为粘着性语言，其词法意义主要是通过各种附加成分来表示，有些为孤立型语言，其词法意义主要是通过语序和助词等来表示。因此，各个语言的词法分析在理论和技术层面都有所不同。在少数民族语言词法分析方面的主要成果有：
（1）藏文自动分词系统和自动标注系统；
（2）蒙古文自动分词和自动标注系统；
（3）维哈克文自动分词及词法标注系统；
（4）基于语料库的哈文词信息分析与统计技术。

2.4 语义分析

所谓“语义分析”是从语句中抽取有效的知识，经过这些知识分析、理解当前句子或篇章所表达的意义。语义分析和语义知识库是是少数民族语言信息处理的一个薄弱环节，但近几年经过努力在语义知识库的建立和语义分析方面有了长足的发展。这方面的主要成果有：
（1）蒙古语语义信息词典；
（2）蒙古语词汇语义网 ；
（3）熟语知识库及其学习平台。

3.应用软件系统开发

应用软件是专门为某一应用目的而编制的软件，如：文字处理软件、机器翻译软件、文字识别软件、语音识别与合成软件等等。 近年来民文处理应用软件层出不穷，为民族地区广大用户提供信息化服务，取得了很大的社会和经济效益。

3.1 机器翻译

机器翻译（machine translation, MT）是通过计算机程序将一种自然语言转换到另外一种自然语言的自动过程。机器翻译按其技术类型可分为基于规则的机器翻译、基于统计的机器翻译、基于系统融合的机器翻译等，近年来又出现了基于深度学习和神经网络的机器翻译。我国民族语言机器翻译经过基于规则、基于统计、基于系统融合的机器翻译过程，目前已经涉足基于深度学习的机器翻译。民文机器翻译是少数民族语言信息处理的一个活跃的领域。
1）达日罕汉蒙机器翻译系统，这是一款面向政府公文的机器翻译系统，旨在将用汉文发布的政府公文自动翻译成蒙古文，系统采用系统融合技术，其准确率达到 85%以上。
2）多民族语言互译农业信息处理平台；
3）汉藏机器翻译系统及一系列汉藏辅助翻译工具；
4）面向领域的维汉机器翻译系统；

5）壮文电子词典及辅助翻译软件；
6）汉-东盟语种辅助翻译系统（苹果及安卓）；
维汉、蒙汉和藏汉统计机器翻译系统，该系统在著名国际机器翻译评测 NIST 和 IWSLT 中多次取得好成绩。

3.2 电子词典

电子词典就是把传统纸质词典中的内容转换为数字格式存储的电子数据。电子词典都在数据库基础上再匹配相应的查询、维护软件，用户通过键盘输入、屏幕抓词或词条点击等方式找到需要查询的条目，以解决实时翻译的需求。比如输入一个中文单词后便可以找到该单词的民文解释、音标、词类等相关信息，有的产品还可发声。民族语言领域较为代表性产品有：
1）汉蒙英日大辞典；
2）固什汉蒙数字词典 2.5 版（Windows、iOS、安卓）；
3）汉壮双向在线词典；
4）汉越、汉泰、汉马来、汉印尼、汉缅、汉老在线双向词典。

3.3 文字处理与办公套件

文字处理和办公套件在汉字信息处理领域早已得到解决，但由于民族语言文字固有的特殊性，还得需要不断地开发和升级，包括开源系统的研发，输入法的改进，编码转换的实现等很多工作。在这方面的成果有：
1）MenkOffice2013 办公软件；
2）中标普华藏文办公软件；
3）跨平台的维哈柯文办公套件软件；
4）中标麒麟桌面操作系统等（中标软）；
5）Windows 下的古壮文处理系统；
6）华光多民族文字字库、华光多民族文字输入法；
7）中标普华维哈柯文办公软件；
8）蒙科立蒙古文智能输入法 2016（Windows、iOS、安卓）；
9）基于安卓系统的朝文输入法---安音 3.0；
10）蒙古文 28 款 OpenType 字库、传统蒙古文 AAT 字库；
11）蒙古文校对软件；
12）西里尔蒙古文与传统蒙古文相互转换系统；
13）哈萨克文老文字和斯拉夫文转换技术。

3.4 模式识别

模式识别（英语：Pattern Recognition），就是通过计算机用数学技术方法来研究模式的自动处理和判读。在自然语言处理中这一技术运用于语音识别和文字识别中。由于我国少数民族语言隶属于不同语系、文字系统也不尽相同，所以语音识别和文字识别也各有千秋。目前的进展包括：
1）维吾尔语问答系统；
2）蒙古语语音合成系统；
3）藏文信息处理技术平台；
4）多民族文字手写字符识别和古籍文献识别技术。

4.技术展望与发展趋势

4.1 多语种自然语言理解与智能处理研究

1）自然语言理解基础理论研究

研究汉藏、阿尔泰、印欧、闪含语系等国内少数民族语言和周边国家语言的词法、句法、语义、篇章、情感、蕴含、信息抽取等语言分析方法；研究复杂形态语言和长距离语言模型、跨语言文法推导方法等。

2）大规模多层次知识库构建与知识挖掘研究

研究多语言知识的挖掘方法和模型，构建大规模、高质量、多层次的综合型多语言知识库系统。

3）多语种智能信息处理技术研究

多语种文字校对与纠正、社交媒体处理、信息抽取、自动文摘、信息检索、跨语言检索、社会焦点检测、人机智能问答、语种转换、口语对话、多语言机器翻译等技术研究。

4.2 一带一路多语种语音信息处理理论与技术研究

1）一带一路多语种共享语音资源库建设；

2）一带一路多语种语音识别与合成；

3）一带一路多语种语音检索技术；

4）一带一路多语种口语交互技术。

4.3 媒体信息智能处理论及技术研究

1）一带一路多语种图像文字资源库建设；

2）一带一路多语种（印刷、手写）文档图像检索、鉴别与识别技术；

3）一带一路多民族人脸和行为识别技术；

4）视频数据分析和处理

研究多语言视频检索、海量媒体的语义挖掘、自动分析与处理技术、智能视频监控理论与方法。

5）一带一路多语种古文献数字化整理技术。

4.4 一带一路多语种网络信息安全理论与技术研究

基于多语种的跨语言、复杂社会网络的分析技术成为研究热点。

1）多语种环境下网络信息内容安全

网络文本、语音、视觉等大数据的采集，用于舆情分析、人脸识别、说话人识别、目标对象跟踪等感知技术。

2）一带一路多语种环境下在线社交网络群体行为安全；

3）网络信息安全保障、度量及建模。

4.5 一带一路多语种信息处理应用研究

1）多语种社会公共安全信息技术与系统

跨媒体多语言社会舆情监测、情报收集技术与系统、面向反恐的视频与音频分析及检索、基于北斗卫星的特定人和物体的实时跟踪技术等。

2）多语种空间地理信息应用技术与系统

研究与实现集成于语音识别、机器翻译、信息检索的多语言电子地图应用系统（舆情与情绪地图、事件地图、消费地图、观念地图等）。

3）多元文化信息资源数字化智能处理与传播平台构建技术

多元文化信息混合采集、信息资源整合、混合编辑处理及服务等。

4）一带一路多元文化遗产虚拟现实复原

研究丰富多样的文化遗产的数字化虚拟再现、三维模型智能处理、检索及虚拟场景快速构建及真实场景的模拟等技术。

5）一带一路多语种移动互联网及终端软件开发

手机终端和语音翻译、语音检索等其它智能电子产品。

上述国家战略及其实施，可以有效地解决我国民文信息处理中的大量基础性、共同性的关键和核心问题，避免重复开发，保证各个民族语言处理软件的兼容性和相互支持，以有限的投入换取最大的成果，形成产业规模、促进我国多语种信息处理技术和成果对一带一路建设中的辐射、引领性作用。

*本次民文信息处理研究进展材料搜集中参与机构有内蒙古大学蒙古学学院、新疆大学、南宁市平方软件新技术有限责任公司、南宁新芽多媒体有限责任公司、广西达译商务服务有限责任公司、潍坊北大青鸟华光照排有限公司、南宁市平方软件新技术有限责任公司、内蒙古蒙科立科技有限公司，内蒙古大学计算机学院蒙古文信息处理实验室、内蒙古师范大学计算机与信息工程学院、中国科学院新疆理化技术研究所、西北民族大学民族信息处理研究所、中国科学院合肥智能机械研究所、中标软件公司、全国朝鲜文信息技术工作组（延边大学）、中国科学院软件研究所、中国科学院计算技术研究所等 16 个单位，涵盖蒙、藏、维、哈、朝以及东亚语言、壮文和多文信息技术研究领域。其近年来的研究进展综述如同以下所列。