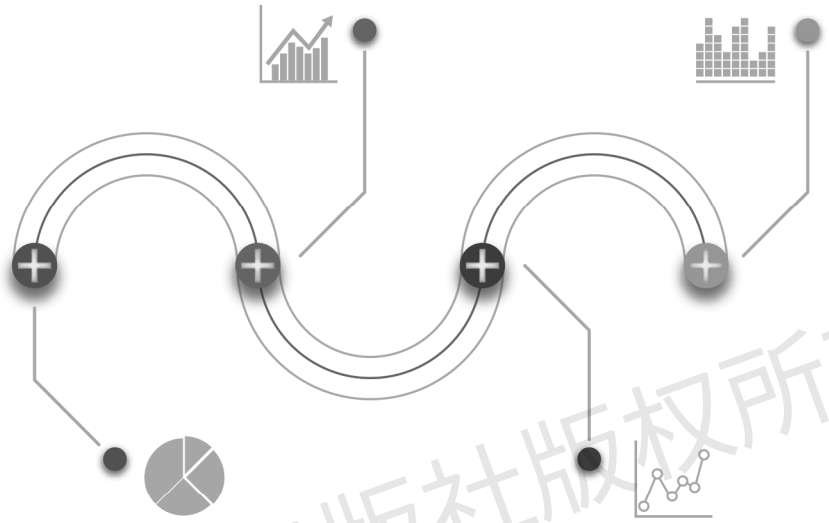
“数据标注”人才培养系列丛书



数据标注工程
——语言知识与应用

组编◎辽宁盘石数据科技有限公司
主编◎于　东　王会珍

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

电子工业出版社版权所有
盗版必究

内 容 简 介

本书介绍了人工智能与语言知识的结合特点。通过理论概念讲解、具体实例分析，介绍语言知识的构建方法、类型案例、应用领域，辅助学习者快速了解行业基础和发展动态。

本书首先介绍语言知识库的基本理论和构建方法，通过例子介绍资源类语言知识、语料库语言知识的概念和结构。为了方便理解，本书结合大量案例介绍语言知识在自然语言处理及在司法、医疗、金融等垂直领域中的应用，目的是帮助数据标注者理解行业发展，建立语言知识理论和应用的基本框架，为从事相关工作提供便利。

本书是一本面向数据标注人才培训的理论教材，适用于有意从事数据标注、语言知识库构建和应用的相关技术人员。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

数据标注工程. 语言知识与应用 / 于东，王会珍主编. —北京：电子工业出版社，2023.8

ISBN 978-7-121-45955-9

Ⅰ. ①数… Ⅱ. ①于… ②王… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国国家版本馆 CIP 数据核字（2023）第 126911 号

责任编辑：杨　波　　文字编辑：张　京
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1 092　1/16　印张：8.75　字数：142.8 千字
版　　次：2023 年 8 月第 1 版
印　　次：2023 年 8 月第 1 次印刷
定　　价：58.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254584，yangbo@phei.com.cn。

电子工业出版社版权所有
盗版必究

序

目前，我们正经历人工智能的第三次浪潮，机器学习大行其道。机器学习的发展和进步主要依赖算法和数据。如今，算法基本相同，数据的作用尤其突出。这里所说的数据是指机器学习所用的带标数据，这种带标数据是通过数据标注的方式获得的。

数据标注是被人工智能催生出来的新兴职业，对人工智能的实现至关重要，也因人工智能技术落地的大量需求而进入从业者的视野。近几年，在数据标注的助力下，人工智能的应用场景不断落地，让大家享受到了人工智能的便利。

人工智能变得越来越智能，数据标注行业面临的挑战也就越来越大，这种挑战主要体现在两个方面：一是数据标注的质量要求越来越高，人工智能正在经历着从 1 到 2 的发展过程，需要更多高质量的带标数据支撑，人工智能发展初期的准确率已无法满足当今人工智能技术发展的需求；二是数据标注任务的难度越来越高，随着人工智能技术的日趋成熟，人工智能任务的难度不断提高，数据标注的难度也在不断提高。

这些都对数据标注人员提出了更高的要求，一方面要求数据标注人员在工作时要更加细致，另一方面也要求数据标注人员具有更高的素质。基于这种趋势，数据标注人员想在数据标注行业取得持续性发展，就要不断提高自身的能力和素质，向专业化方向发展。

事在人为，业以人兴。数据标注乃至人工智能行业的发展关键在于专业人才的培养。

在未来几十年，数据标注会伴随着人工智能需求的不断提高而不断发展、

电子工业出版社版权所有 盗版必究

精进。我相信会有更多的年轻人愿意加入数据标注行业，享受学习的福利与时代的红利，也相信本书能为他们的职业生涯助一臂之力，为求知者打开一扇新领域的大门。我期待数据标注人员将来利用自己卓越的数据标注技能通过计算机及智能设备给人类提供更丰富的智能服务。

中国中文信息学会名誉理事长
哈尔滨工业大学教授
李生

2023 年 3 月

电子工业出版社版权所有
盗版必究

PREFACE

前　言

电子工业出版社版权所有

语言是人类知识的重要载体，语言知识也是人工智能发展的基石。研究语言知识的特点，是人工智能相关从业人员必不可少的基本技能。为了满足相关领域学生和工程技术人员对语言知识发展和应用方法了解的需求，我们根据多年授课经验，精心编撰了这本《数据标注工程——语言知识与应用》，希望能够通过这本书，向更多的读者介绍语言知识在人工智能诸多论题中的基础性作用。本书可以作为相关学科学生教材以及相关培训的理论基础教材，也可以作为相关工程技术人员在构建语言知识库、使用语言知识库进行研究开发时的参考资料。

本书第 1 章介绍人工智能的基本概念和基本方法，以及面向人工智能的知识表示方法。详细为读者介绍主流的语言知识表示类型和特点。在此基础上，介绍语言知识基本理论。

第 2 ~ 4 章介绍语言知识的构建、获取、存储方法。内容涵盖了语言知识库开发的全过程。在此基础上，具体介绍资源类语言知识、语料库语言知识的概念和方法。针对资源类语言知识，以目前典型的语言知识资源如语义网络、知识图谱为例详细介绍其开发、构建方法；针对语料库语言知识，进一步从词法、实体关系、句法、篇章几个角度展开论述。

第 5 ~ 6 章结合大量案例介绍语言知识的应用，包括语言知识在自动问答、机器阅读理解、机器翻译等问题中的应用方法，以及在智能司法、智能医疗、

智能金融等垂直领域中的应用。

本书理论框架清晰，内容循序渐进，经过精心策划，在内容上体现人工智能与语料库语言学相结合的最新进展。书中介绍了许多目前主流的语料库资源，为相关技术人员学习建立语料库，从事语言知识研究提供了极大方便。

编　者

电子工业出版社版权所有
盗版必究

CONTENTS

目 录

第 1 章　人工智能与知识表示 ····· 001

1.1　智能与人工智能 ····· 001

1.1.1　智能 ····· 001

1.1.2　人工智能 ····· 002

1.2　基本方法和流派 ····· 003

1.2.1　基本方法 ····· 004

1.2.2　基本流派 ····· 004

1.3　知识表示 ····· 005

1.3.1　知识符号化 ····· 005

1.3.2　知识表示的概念 ····· 006

1.3.3　人工智能中的知识表示 ····· 007

1.4　一阶谓词逻辑的知识表示 ····· 008

1.4.1　命题逻辑 ····· 008

1.4.2　谓词逻辑 ····· 010

1.4.3　使用谓词表示知识 ····· 012

1.4.4　小结 ····· 014

1.5　产生式知识表示 ····· 015

1.5.1　产生式的概念 ····· 015

1.5.2　规则性知识的产生式 ····· 015

1.5.3　事实性知识的产生式 ····· 017

1.5.4　产生式系统 ····· 017

电子工业出版社版权所有
盗版必究

第 2 章　语言知识库的构建……023

2.1　语言知识的概念……023
2.1.1　语言知识……023
2.1.2　语言知识库……024
2.1.3　语言知识库的类型……025
2.2　语言知识的来源……026
2.2.1　结构化数据……027
2.2.2　半结构化数据……027
2.2.3　非结构化数据……028
2.3　语言知识库的构建……028
2.3.1　构建流程……028
2.3.2　规范和原则……029
2.4　语言知识获取方法……031
2.4.1　人工标注知识……031
2.4.2　自动获取知识……032
2.4.3　人机交互获取知识……033
2.5　语言知识的存储……033
2.5.1　数据库及其类型……033
2.5.2　可扩展标记语言……034
2.5.3　数据交换格式……036
2.5.4　本体知识表示……037

第 3 章　资源类语言知识……042

3.1　资源类语言知识的概念……042
3.2　资源类语言知识的发展……043
3.2.1　语义网络……043
3.2.2　语义 Web……044
3.2.3　知识图谱……049
3.3　常用的资源类语言知识……049
3.3.1　WordNet……049

电子工业出版社版权所有
盗版必究

3.3.2 FrameNet ····· 050
3.3.3 ConceptNet ····· 052
3.3.4 HowNet ····· 054
3.3.5 同义词词林 ····· 055

第 4 章　语料库语言知识 ····· 059

4.1 词汇中的语言知识 ····· 059
4.1.1 词性知识 ····· 059
4.1.2 分词知识 ····· 061
4.2 句子中的语言知识 ····· 062
4.2.1 命名实体知识 ····· 063
4.2.2 实体关系知识 ····· 063
4.2.3 事件知识 ····· 064
4.3 句子结构中的知识 ····· 065
4.3.1 句法结构树 ····· 065
4.3.2 浅层句法结构 ····· 066
4.3.3 依存句法树 ····· 067
4.3.4 抽象语义表示 ····· 069
4.4 常用汉语语料库 ····· 070
4.4.1 大规模汉语语料库 ····· 070
4.4.2 汉语标注语料库 ····· 072

第 5 章　语言知识的应用：面向自然语言处理 ····· 077

5.1 自然语言处理的基本问题 ····· 077
5.1.1 语言模型问题 ····· 077
5.1.2 分类问题 ····· 080
5.1.3 序列标注问题 ····· 081
5.1.4 语言结构分析问题 ····· 083
5.1.5 语言生成问题 ····· 085
5.2 自动问答 ····· 085

电子工业出版社版权所有
盗版必究

5.2.1 概念和历史……085
5.2.2 开放领域自动问答……087
5.2.3 基于知识的自动问答……088
5.3 机器阅读理解……090
5.3.1 概念和发展史……090
5.3.2 完型填空型任务和数据集……092
5.3.3 选择型任务和数据集……093
5.3.4 片段抽取型任务和数据集……094
5.3.5 自由问答型任务和数据集……095
5.4 机器翻译……096
5.4.1 概念和发展史……096
5.4.2 机器翻译的基石：双语平行语料库……098
5.4.3 统计机器翻译方法简介……099
5.4.4 神经机器翻译方法简介……100

第 6 章 语言知识的应用：面向垂直领域……104

6.1 智能司法信息处理……104
6.1.1 概述……104
6.1.2 法律判决预测任务……106
6.1.3 相似案件匹配任务……107
6.1.4 司法领域自动问答……108
6.2 智能医疗信息处理……110
6.2.1 概述……110
6.2.2 医疗信息知识库构建……111
6.2.3 智慧医疗的典型应用……115
6.2.4 智慧医疗的未来发展……116
6.3 智能金融信息处理……117
6.3.1 概述……117
6.3.2 金融领域知识库构建与分析技术……118
6.3.3 智能金融的典型应用……123

电子工业出版社版权所有
盗版必究

第1章 人工智能与知识表示

电子工业出版社版权所有
盗版必究

【本章学习目标】

（1）了解人工智能的概念，理解人工智能研究的基本方法和基本流派。

（2）理解知识表示的概念和内涵，理解知识表示在人工智能中的应用方法。

（3）理解命题逻辑和一阶谓词逻辑的概念和形式。掌握利用谓词表达知识的步骤。

（4）理解产生式的概念和形式。掌握产生式知识的编写步骤。

1.1 智能与人工智能

1.1.1 智能

智能的定义非常模糊，涵盖的范畴非常广泛，所以我们很难给智能下一个

科学、精确的定义。斯滕伯格在 1994 年就人类意识这个主题给出了以下有用的定义。他认为，智能是个体从经验中学习理性思考、记忆重要信息以及应付日常生活需求的认知能力。这个定义把智能与意识和智慧等同在一起。这个定义突出了学习的重要性，同时强调了智能必须能够记忆信息，以及智能的目标是应付日常生活中的需求，但并没有强调智能的拥有者必须是人类。也就是说，任何满足这样条件的人或者事物都是具有智能的。

在本书中，我们将智能认为是智慧和能力的综合体现。智慧包括感知、记忆、思维、学习、创造等，而能力包括行为、语言、情感等。此外，智能还包括一些特有的特征，如有思维、有创造性、有情感等，人类对于智能有着多样化的研究和阐述。每一种智能都有自己独特的特点。比如，人类智能的独有特点是有创造性、有情感；人类智能体现的一个重要方面是有语言。这使得智能的定义在科学范畴仍然具有一定的模糊性，也正是因为这种模糊性，使得智能可以被演绎成各种模样。

如果我们把智能的范围从生物扩展到事物，如何去评判一个机器是否具有智能？如果我们认为生命并不是拥有智能的必要条件，换句话说，如果一个机器表现出与人类智能相关的特征，是否可以认定这个机器也具有智能？这就属于人工智能研究的范畴。为了了解这个概念，我们首先要来介绍人工智能的概念。

电子工业出版社版权所有 盗版必究

1.1.2 人工智能

提到人工智能，大多数人都会想起许多科幻小说和电影中塑造的机器人的形象。例如，20 世纪 80 年代的系列科幻电影《星球大战》塑造了人形和桶形机器人形象，成为许多青少年的科幻启蒙之作。2000 年上映的电影《人工智能》则描述了一位机器人小孩大卫，为了成为真正的人而奋斗的故事，也反映了创作者对人工智能发展的乐观态度。近年来，许多科幻影视作品都对未来社会中人类与人工智能机器人共存共处的情况进行了具体的描绘。如迪士尼电影《机器人总动员》在讲述两个机器人的爱情故事的同时，也展示了人类与人工智能机器人相处中，几乎成为附庸的现象。这样的思想在《黑客帝国》《终结者》系

列电影中得到了更明确的阐述——当人工智能机器人发展到一定程度，具有自主意识后，必将超越人类，并且进一步毁灭人类。由此可以看出创作者对人工智能持悲观态度。

所谓人工智能，就是用人工的方法，在机器或者计算机上实现的智能，也称为机器智能。目前我们对人工智能有一个基本的共识，即人工智能区别于自然智能，是一种用人工手段产生的智能现象。这种人工制造的智能能够像自然智能一样，实现某种智能能力。比如，人类的许多活动，如下棋竞技、编写程序、驾驶汽车等，都是需要智能才能完成的。如果我们制造了一个机器，它能够像人类一样完成这些任务，我们就认为这个机器具有了某种性质的人工智能。

国内许多学者也对人工智能的概念给出了自己的论断，比如谭铁牛院士在《求是》上曾经提出过自己对人工智能的定义："人工智能是研究开发能够模拟、延伸、扩展人类智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能的研究目的是促使智能机器能够会听、会看、会说、会思考、会学习、会行动。"

在这个定义中，谭铁牛院士用比较通俗的语言概括了当前人工智能所要实现的目标，即会听、会看、会说、会思考、会学习、会行动。这些目标分别对应了一个应用问题，如语音翻译器就对应了会听、会说，人工智能竞技则对应了会思考，自动驾驶对应了会行动。从科学的角度出发，人工智能并非天马行空，更非无所不能。目前，人工智能研究的重点在于在有限的条件下，对人类智能的某种能力在某些具体问题上的模拟。

1.2 基本方法和流派

在 1956 年的达特茅斯会议上，与会者给出了人工智能学科的定义。在此基础上，与会者们又将当时所有符合人工智能思想的方法进行了汇总，把这些方法归入学科的研究范畴，也就形成了最初的人工智能研究的流派，为后续的学科发展奠定了基础。人工智能的方法也随之逐渐演变成几种主要的流派。根据达特茅斯会议中对人工智能的纲领性描述——学习的每个方面或智能的任何特征都能被精确地描述到用机器来模拟的程度——可以将人工智能研究分为两种方法、三大流派。其中，两种方法分别是理性主义方法、经验主义方法；三大

电子工业出版社版权所有 盗版必究

流派分别是符号主义、连接主义、行为主义。

1.2.1 基本方法

人工智能的第一个维度是如何用机器来表达知识和知识的获取过程，即学习过程。人工智能基本方法包括理性主义方法和经验主义方法。

1．理性主义方法

理性主义认为，人类通过学习得到的知识是绝对理性的，是可以精确描述的，因此人工智能需要由人工对人类学习到的知识进行汇总、加工、抽象、归纳并建立某种理性思考的框架。智能机器依赖这种框架展开思考。理性主义是人工智能早期的主流思潮，至今依然扮演着重要的角色，如现在人们所熟知的知识库、知识图谱等，就是以理性主义为基础建立的。

2．经验主义方法

与理性主义不同，经验主义则认为，外部世界的知识是无法精确描述的，只能通过体验、经历、感受才能获得。智能机器需要通过不断感知外部世界，以尝试、探索的方式来获取知识。经验主义其实就是机器学习的最基本思想。

理性主义和经验主义的区别在于，智能机器获得经验知识的方式不同。理性主义认为知识需要人工总结，机器来运用知识；而经验主义则希望智能机器能够自己获取解决问题的经验知识。

另外，利用智能机器解决实际问题，还需要依赖问题的形式化，即如何将问题表达为智能机器能接受的形式。目前，主要包括以下三个流派。

1.2.2 基本流派

人工智能的第二个维度，是智能的特征如何用机器来模拟，或者说机器应该从什么角度出发去模拟智能。这一问题使得人工智能研究衍生出三种思潮，

电子工业出版社版权所有 盗版必究

分别是符号主义、连接主义和行为主义。

1．符号主义

符号主义认为人工智能本质是知识符号化，只要将世界知识转换为某种符号系统，智能机器就可以根据该符号系统，解决真实世界的问题。可见，符号主义与理性主义是统一的，是理性主义解决问题的方式。

2．连接主义

连接主义认为，大脑是智能产生的根源，因此实现人工智能应该研究大脑的结构、信息处理机制、运行方式，然后在机器上模拟大脑，实现人工智能。这就是现阶段主流的人工神经网络方法。

3．行为主义

行为主义又称模拟学派，该学派认为智能行为的基础是“感知—行动”的反应机制，认为智能只是在与环境交互作用中表现出来，不应采用集中式的模式，而是需要具有不同的行为模块与环境交互，以此来产生复杂的行为。其研究重点是模拟人在控制过程中的智能行为和作用，如对自寻优、自适应、自镇定、自组织和自学习等控制论系统的研究。

根据以上描述，我们可以知道在人工智能研究的主要方法中，明斯基所提出的人工神经网络方法是经验主义和连接主义的；麦卡锡提出的搜索方法是符号主义和经验主义的；而西蒙和纽威尔所提出的逻辑理论家方法则是理性主义和符号主义的。

电子工业出版社版权所有 盗版必究

1.3 知识表示

1.3.1 知识符号化

广义的知识表示问题贯穿了整个哲学和科学的发展历史。早在古典哲学中

就已经出现了事实上的知识表现方法，如古典哲学中最经典的三段论推理就可以视作一种知识表示的形式，图 1-1 所示为三段论中的知识表示。

图 1-1　三段论中的知识表示

分析图 1-1 可知，三段论推理结果为：“因为苏格拉底是人，所以苏格拉底是会死的。”在该推理过程中，你可以将其中的实体和动作替换掉，使之变成：“如果所有的 B 都满足 A 且 C 属于 B 则 C 必定满足 A。”这样的过程实际上就是一种知识抽象化的过程，通过这样的方法就可以把一个推理形式用一个符号体系来表示，而这个符号体系就体现了这个推理中所蕴含的知识，在事实上已经完成了知识的符号化表示。

在 17 世纪，德国的数学家莱布尼茨（Gottfried Wilhelm Leibniz）提出一个设想，即如果将人类的知识用一组概念来表示，那么这组概念就是人类思想的字母表，有了这些字母表，人类的知识就可以通过字母之间的逻辑运算来得到。莱布尼茨设想的这种符号体系被称为“普遍文字”。这样的一些思想后来被德国哲学家弗雷格（Friedrich Ludwig Gottlob Frege）发扬光大。

1879 年，弗雷格发表了被誉为“亚里士多德之后在逻辑学领域最重要的出版物”的著作《概念文字》。在该著作中，弗雷格建立了一套符号体系，并尝试用这套符号体系表示包括数学在内的全部逻辑系统。这套体系对后来的科学发展产生了深远的影响，是知识符号化的鼻祖。

1.3.2 知识表示的概念

在现代，我们所说的“知识表示”，是指将人类的知识形式化或者模型化，即研究如何把一个任务领域中人类的“知识”表示成为计算机能够接受和理解的形式。知识表示有许多形式，可以是一种符号描述；也可以是由人们制定的某种规则或者约定；也可以是某一些数据，或者某种数据结构，只要能够表达我们所需要处理的知识，并且能够让计算机访问到、能够处理，就是一种

电子工业出版社版权所有 盗版必究

知识表示形式。每一种知识表示都可以对应一个知识库。知识库中的内容，可以描述为“问题域中对象和关系”与“程序中的计算对象和关系”之间的一种映射。对于一个问题，如果让人类去思考，会得到该问题相关的对象及其关系，而知识库则是建立一种映射，将人类思考的对象和关系表示成为计算机程序中的对象及其关系。

一般来说，知识库中的知识应该与现实世界中人类的观察结果相一致。而用来建立人类思考与计算机程序之间的媒介，就是“知识表示模式”。

1.3.3 人工智能中的知识表示

前面介绍了人工智能各个流派及其发展历程，实际上在人工智能的每一个流派的发展过程中，知识表示都是需要解决的核心问题。随着流派的发展和方法的变迁，知识表示也经历了许多的变化。下面就具体介绍在人工智能领域中的知识表示。

电子工业出版社版权所有 盗版必究

1．符号主义的知识表示

最早期的人工智能的研究方法主要是理性主义。其中，主要采用符号来表示知识，由此形成了理性主义方法加符号主义知识表示的潮流。其本质就是用符号来描述知识，然后再对表示得到的符号公式进行计算，从而求解问题，由此可见知识表示在符号主义中的核心地位。在这一阶段，知识的表示包括谓词逻辑、产生式以及框架等。

进入 20 世纪 90 年代后，人工智能的经验主义方法复兴，知识表示的形式由符号主义逐渐向多样化发展。知识不再由人工去抽象归纳，而是通过大量样本的状态特征来间接体现，即状态空间表示方法。在这种方法中，样本数据的状态特征表示对应了某种知识，智能系统通过体验或者经历这种样本，把每一种样本的特征学习到，就获取了这个样本或者这一类问题中所包含的知识。这种知识表示的方法，仍然是符号主义的，通过每一种状态特征学习得到的知识仍然需要以符号形式表示和存储。

2．连接主义的知识表示

进入 2010 年，随着连接主义的深度学习兴起，知识的表示更加抽象。由于连接主义的核心是模拟人类大脑神经元的连接结构，因此在这种情况下，从样本数据中获得的知识实际上表现出来就是神经网络中的语义向量及神经网络中神经元的权重。连接主义认为整个神经网络及它每一个神经元所对应的权重，就表示了某种知识。连接主义中的知识表示形式通常就是语义向量及网络权重。

本书由于篇幅所限，无法完全涉及所有的知识表示形式，下文将主要在讲解符号主义的知识表示方法基础上，介绍基于逻辑规则的知识表示方法。基于逻辑规则的知识表示方法是人工智能第一次发展浪潮中研究的热点，包括了一阶谓词逻辑、产生式两种方式，下面分别进行介绍。

电子工业出版社版权所有 盗版必究

1.4 一阶谓词逻辑的知识表示

在人工智能研究的早期阶段，计算机能够处理的信息类型非常单一，绝大部分都是符号串。在这种情况下，用符号来表示知识就成了人工智能研究的关键课题。学者们认为知识往往由语言来描述，因此如果能够将语言描述的知识转化为某个符号体系，就可以实现知识的形式化，计算机只需要读入这个符号就可以获得相应的知识，而在这方面最先应用于人工智能研究的就是命题和谓词逻辑。

1.4.1 命题逻辑

了解命题逻辑的首要任务是明确命题概念，命题是指对客观世界的带有真假意义的确定性的陈述句，并且满足下面三个条件：

（1）命题一定是陈述句而不能是疑问句。

（2）命题要有确定性。如“他晚上可能会来”这样带有可能性的句子就不是一个命题。

（3）命题要有真假意义。

命题可以是一个单一的论述，如“云是白色的”就是单一命题；也可以是复合的论述，如“香蕉的表皮是黄色的而果肉是白色的”就是复合命题。复合命题涉及的实体和关系较单一命题更为复杂。

利用命题逻辑即可表示一些知识，如可以用英文的大写字母来表示某一个实体的事实，再配合一些逻辑符号，就可以表示知识，下面是一个命题逻辑的例子。

例 1-1 A、B、C 参加面试，面试官对三人进行了判断，得到如下几条想法：

（1）三人中至少录取一人；

（2）若录取 A 而不录取 B，则一定录取 C；

（3）B、C 要么都录取，要么都不录取；

（4）若录取 C，则一定录取 A。

那么，最终谁会被录取呢？

本例中面试官的 4 条“想法”就是 4 个命题，要将这 4 个命题的逻辑用符号表示，可以先用 P、Q 和 R 分别来表示录取 A、录取 B 和录取 C 这三个事件，然后就用符号表示这些命题，如下所示：

（1）“三人中至少录取一人”用符号表示为：// P∨Q∨R；

（2）“若录取 A 而不录取 B，则一定录取 C”用符号表示为：// P∧¬Q→R；

（3）“B、C 要么都录取，要么都不录取”用符号表示为：// Q↔R；

（4）“若录取 C，则一定录取 A”用符号表示为：// R→P。

通过上文的方法，原始的四个命题形式就被转化为了一个符号体系、一个符号串，其中“∨”为逻辑析取，表示“或”；“∧”为逻辑合取，表示“和”；“¬”为逻辑否定，表示逆命题；“→”为实质蕴涵，表示如果该符号前的部分成立，则该符号后面的部分也成立；“↔”为实质等价，表示该符号的两边的部分等价。符号串中的每一个符号都对应着某一个命题，这就是利用命题逻辑来表示知识的基本方法。

命题逻辑用符号来表示命题，然后通过逻辑符号连接命题，就可以得到复

电子工业出版社版权所有 盗版必究

合命题，从而实现对复杂知识的表示。但命题逻辑也有它的局限性，比如命题逻辑对命题内部的结构实际上没有做描述，而把整个命题表示成一个字符符号，同时对不同命题的共同特征也缺少描述手段。如这样两个命题："小张是一个父亲""小李是一个父亲"，这是两个不同的命题，所以在作为符号的时候，需要用 P、Q 分别表示，整个命题则是"P∧Q"。实际上这两个命题"小张是一个父亲""小李是一个父亲"非常类似，但在符号中无法体现出其相似性。

命题逻辑的这种局限性，使得命题的任何一个非常小的改变都会导致命题的符号的变化。为了解决这个问题，在命题逻辑的基础上，人工智能研究者们又发展出了谓词逻辑。下面就对谓词逻辑进行具体介绍。

1.4.2 谓词逻辑

通过仔细观察命题，研究者们发现最简单的命题至少由一个主语和一个谓词组成。如"小张是一个父亲""云是白色的""香蕉的味道是甜的"。这几个命题中的"小张""云""香蕉"作为各个命题的主语，表示了命题中独立存在的某个事物、某个实体或者相应的概念，而"父亲""白色""味道是甜的"这样的词则是命题的谓词，谓词刻画了主语的某种性质、某种状态或者某种关系属性。

在一个命题中，主语往往是具体的、特定的，但谓词往往是通用的，如上文所提及的"小张是一个父亲""小李是一个父亲"中，"小张"和"小李"是主语，是特定的，但"父亲"这个谓词实际上是通用的，可以出现在许多不同的命题中，并代表相同的含义。"云是白色的""棉花是白色的"中，"云"和"棉花"都是实体，都是特定的概念，但"白色"则是一个通用的属性。同样，"香蕉的味道是甜的""苹果的味道是甜的"也是同理。

因此，如果将命题中的谓词抽象出来，作为核心成分，而将主语视作是谓词所支配的部分，就可以将命题形式化为以谓词为中心，主语为谓词附属成分的形式。以这种形式来表示知识，就是谓词逻辑。

使用谓词逻辑来表述简单的命题，有一种基本形式，即 $P(x_1,x_2,\cdots,x_n)$这样一个简单的公式。在该公式中，P 代表一个谓词，x_1 一直到 x_n 则是这个谓词所操作或者所控制的个体或者实体。谓词操作的个体数目，称为谓词的元数，只操

电子工业出版社版权所有 盗版必究

作一个实体的谓词，我们把它称为一元谓词，即 $P(x_1)$，里边只有一个实体。有些谓词有两个实体，即 $P(x_1,x_2)$，需要有涉及两个实体的某种关系。

要将一个命题转化成某个谓词，首先就涉及谓词的定义、谓词的含义、谓词的元数设计。这些通常由使用者根据自己的经验和需要来规定，一般而言，谓词用具有相应意义的英文单词或者词串来表示。

同一个命题可以通过提取出不同的谓词，从而得到不同的谓词逻辑表达。下面有 3 组共 6 个不同的命题，它们都可以转化为不同的谓词逻辑表达。

例 1-2 以下为 6 个命题。

小张是一个父亲

小李是一个父亲

云是白色的

棉花是白色的

香蕉的味道是甜的

苹果的味道是甜的

这 6 个命题中两两一组，其结构非常相似。针对第一组“小张是一个父亲”“小李是一个父亲”两个命题，可以把“是”当作谓词设计，即 is_a，“小张是一个父亲”就可以表示为“is_a（小张，父亲）”；同理，“小李是一个父亲”可以表示为“is_a（小李，父亲）”。针对第二组，“云是白色的”“棉花是白色的”，则可以将“颜色”作为谓词，将它们分别表达为“color（云，白色）”“color（棉花，白色）”。第三组“香蕉的味道是甜的”“苹果的味道是甜的”也一样，提取“味道”为谓词，转化为“taste（香蕉，甜）”“taste（苹果，甜）”的形式。这样就将命题逻辑转化为了谓词逻辑。

上面的转化成果都是二元谓词形式，实际上对于同样的命题，通过提取不同的谓词也可以转化为一元谓词形式。同样是第一组“小张是一个父亲”“小李是一个父亲”，选择提取“父亲”为谓词，就可以分别得到“father（小张）”“father（小李）”两个一元谓词；第二组“云是白色的”“棉花是白色的”提取“白色”为谓词，则可得到“white（云）”“white（棉花）”；第三组“香蕉的味道是甜的”“苹果的味道是甜的”提取“甜”为谓词，则可得到“sweet（香蕉）”“sweet（苹果）”。

电子工业出版社版权所有 盗版必究

1.4.3 使用谓词表示知识

命题逻辑可以转化为谓词逻辑形式，但是如果要表示一些更复杂的，涉及事物与事物间更多关系的命题，仅仅依靠一个谓词是不够的，那么就需要用多个谓词组成的语句来表达其中的知识。

要想用谓词逻辑表达更加复杂的知识，就需要将多个谓词通过逻辑组合和嵌套的方式来实现。而连接各谓词表达所用的符号仍然是前面所提及的“∨”“∧”“¬”“→”“↔”等逻辑符号，用括号( )、[ ]则可组成具有优先运算关系的谓词，如下所示。

例 1-3

小李不在足球场

¬ in(小李，足球场)

李明会打篮球和踢足球

canplay(李明，篮球) ∧ canplay(李明，足球)

我想吃鸡蛋或者蛋糕

wanteat(我，鸡蛋) ∨ wanteat(我，蛋糕)

小张的父亲是教师

is_a(father(小张)，教师)

这里的第一个例子是“小李不在足球场”，显然这是一个否定的命题，可以很轻松地抽取出谓词“在”，“小李在足球场”，是“小李不在足球场”的逆命题，谓词表达为“in(小李，足球场)”，则“小李不在足球场”的谓词表示只需在其前面加上“¬”，即“¬ in(小李，足球场)”。第二个例子是“李明会打篮球和踢足球”，其中“打篮球”“踢足球”同时被谓词“会打”限制，且两者是“与”的关系，用“∧”连接，因此其谓词表示可以写成“canplay(李明，篮球)∧canplay(李明，足球)”。第三个例子是“我想吃鸡蛋或者吃蛋糕”，其中“鸡蛋”与“蛋糕”同时被谓词“想吃”限制且二者是任选其一，即“或”的关系，用“∨”连接，所以其谓词表示可以写为“wanteat(我，鸡蛋)∨wanteat(我，蛋糕)”。第四个例子“小张的父亲是教师”则是一个谓词的嵌套，第一个谓词表示的是小张的父亲，第二个谓词表示的是这个小张的父亲是教师，其关系可以通过括号来说明。提取谓词为“是”，表示为 is_a(father(小张)，教师)。

电子工业出版社版权所有 盗版必究

将所有的谓词连接起来，就可构成一个复杂知识的谓词语句。随着需要描述的知识越来越复杂，谓词语句的形式也会越来越复杂，而在所有的这种谓词语句中，有一种特殊的简单的形式，即原子语句。原子语句是指整个谓词逻辑中只有一个谓词语句，在原子语句中允许出现嵌套，不能包含逻辑运算。下面是一些典型的原子语句。

例 1-4

weather(today , rain)

likes(tom, kate)

friends(father_of(david), father_of(tom))

在上面所有举出的例子中，需要表达的知识中的实体都是非常具体的，或者特指的，如小李、苹果等。但是，在很多情况下，需要表示的知识中的实体却是比较泛指的，如“所有人都喜欢吃苹果”，该句中“喜欢吃苹果”是一个具体的事项，而所有人则难以被抽象为一个主体，因为没有一个人能够代表所有人。这种情况下就需要引入谓词的变量与量词。

为了解决这个问题，人工智能的研究者们添加了一个变量来实现对某一类实体的这种描述，通常是用一个大写的英文字母来表示。“所有人都喜欢吃苹果”提取出谓词“喜欢吃”，就可表示为“lovetoeat(*X*,apple)”，其中的 *X* 就并非某一特指的主体，而是代表着某一类人。

通过添加变量，我们解决了特指主体的问题，但是并没能完全解决，因为“所有人都喜欢吃苹果”和“有些人喜欢吃苹果”这两句话在这种语境下依然没有办法得以明确地区分。为了表示这种区别，谓词逻辑设计了两种量词“∃*X*”和“∀*X*”。

（1）∃*X*。存在量词，意为一定存在一个 *X*，满足该命题。

（2）∀*X*。全称量词，意为对所有的 *X*，都满足该命题。

有了存在量词和全称量词，我们就可以更加细致地去区分变量的属性，上文提到的问题也就得以解决。对“所有人都喜欢吃苹果”，就设计一个变量 *X*，表示为“(∀*X*) lovetoeat(*X*,apple)”，而“有些人喜欢吃苹果”则表示为“(∃*X*) lovetoeat(*X*,apple)”。

使用带有量词的谓词语句就可以表示一些更加丰富的知识，下面即为典型的带有量词的谓词语句。

电子工业出版社版权所有 盗版必究

例 1-5

如果星期一不下雨，Tom 会去登山。

¬weather(monday, rain)→go(tom, mountains)。

所有篮球运动员都很高。

$\forall X$ (baskateball_player(X)→tall(X))。

许多人喜欢三文鱼。

$\exists X$ (person(X) ∧ likes(X, Salmon))。

1.4.4 小结

在谓词逻辑中，如果全称量词和存在量词仅仅是用来针对谓词所支配的实体，而并不是用来针对谓词本身的，则被称为一阶谓词，反之即是高阶谓词。如“(Likes) Likes (bill，cat)”，该句中的量词支配的是谓词，所以这是个高阶谓词。

1．一阶谓词逻辑

所谓的一阶谓词逻辑，是指在一阶谓词中，谓词的语义是唯一的。一阶谓词逻辑可以用来表达对客观世界的大部分陈述，包括简单的陈述、多种对象之间的关系以及推理等逻辑关系。

2．一阶谓词的优缺点

一阶谓词的优点有：

（1）自然性。一阶谓词对知识的描述非常贴近于人类的总结。

（2）精确性。一阶谓词通过谓词加上变量，加上量词，能够非常精确地表达复杂的知识。

（3）容易实现。一阶谓词的设计难度较低。

一阶谓词的缺点有：

（4）不能表示不确定性知识。

（5）形式过于自由，兼容性差。

3．一阶谓词逻辑设计过程

用一阶谓词逻辑可以将一组命题表示为一个知识库，其步骤如下：

电子工业出版社版权所有 侵权必究

（1）正确理解命题，分析原子命题，以及原子命题之间的关系。

（2）为每个原子命题定义个体、谓词。

（3）使用恰当的量词。应注意全称量词后跟条件式，存在量词后跟合取式。

（4）使用恰当的连接符。用连接符连接谓词句子，表示给定的命题。

1.5 产生式知识表示

1.5.1 产生式的概念

一阶谓词逻辑有很强的表示能力，但是，它也有很多缺点，如形式过于灵活不容易统一、无法表示不确定知识等，因此，往往无法直接用于计算。除了一阶谓词逻辑，在人工智能学科中，还有另一种知识表示形式——产生式（production rule）。产生式能够根据已知的条件产生新知识，能够弥补一阶谓词逻辑的不足。下面就具体介绍产生式知识表示的相关知识。

产生式是一种能够根据已知的条件产生新知识的式子，这些式子往往以推理规则的形式来描述。产生式的概念最早由美国数学家波斯特于 1943 年提出，在 20 世纪的 60 年代与 70 年代，产生式系统被成功应用于人工智能研究中，并在自动推理定理证明和专家系统中得到了广泛的应用，成为当时人工智能学科的主流方法。之后，产生式系统这种表示形式也被应用于更多的领域，如在形式语言学中，产生式用来描述语言的一个结构；在计算语言学中，产生式被用来描述句法分析器的这种句法规则等。

产生式是用来描述规则性或者事实性知识的一种表达式，其基本形式有规则性和事实性两种，下面分别进行介绍。

1.5.2 规则性知识的产生式

规则性知识的产生式的形式如下：

电子工业出版社版权所有

IF P THEN Q

或 $P \rightarrow Q$

在该式中，P 被称为规则或者产生式的前件，Q 被称为产生式的后件，整个产生式的含义可以理解为“若前件成立，则对应的后件也成立”，这就是规则性产生式的基本形式。

1．确定性推理知识

所有能够用一阶谓词逻辑表示的知识都能够用规则性产生式来表达，且规则性产生式与一阶谓词逻辑蕴含式很像。下面是一个典型的例子。

例 1-6

已知知识：“对于任意动物 X，如果它会飞且是卵生，则它是一只鸟。”

一阶谓词逻辑表示为：“$\forall X$ 动物$(X) \wedge$会飞$(X) \wedge$卵生$(X) \rightarrow$鸟(X)。”

规则性产生式表示为：“IF 动物 AND 会飞 AND 卵生 THEN 该动物是鸟。”

2．规则知识

规则性产生式与一阶谓词逻辑蕴含式相比，不仅能够表达推理知识，还能够表达更多的知识，如“室内温度过高则将空调打开”。这样的知识，无法用一阶谓词逻辑蕴含式来表达，而可以用规则性产生式表达为“IF 室内温度>28 度 THEN 打开空调”。因此，只有一阶谓词逻辑蕴含式可以转写为规则性产生式，而规则性产生式则不一定能够转写为一阶谓词逻辑蕴含式。

3．不确定性规则知识

规则性产生式不仅能够表达确定的规则性知识，还能够表达不确定的规则性知识，如图 1-2 所示，为不确定性规则产生式。

```
IF   微生物的染色斑是革兰氏阴性   and
     微生物外形为杆状     and
     病人是中间宿主
THEN  微生物为绿脓杆菌        (0.6)
```

图 1-2　不确定性规则产生式

图 1-2 所示是一条医学领域的专家知识。在这条规则中，规则性产生式

电子工业出版社版权所有 盗版必究

的后件“微生物为绿脓杆菌”并非确实成立，而是一种有概率的推测结果，因此，在原有的规则性产生式的基础上，在后件的末尾加上了“置信度”，即该产生式中的“(0.6)”，表示在前件成立的情况下，后件有 60%的可能性成立。因此，在需要表达不确定性规则知识时，只需要在确定性规则产生式的末尾加上置信度即可。

1.5.3 事实性知识的产生式

除规则性知识外，在日常生活中，有许多知识本身就是一个固定的事实，如“篮球是圆的”“北京是中国的首都”“明天可能会下雨”“π≈3.14 159”等，这些事实性的知识无法用规则来表达，自然也不能用规则性产生式来表示。

事实性的知识需要用事实性产生式来表示。事实性产生式一般是多元组的形式，根据所表示的知识的形式涉及对象的数量不同，又可分为关系型产生式和属性型产生式。其形式分别如下：

（1）关系型产生式：（对象 1，对象 2，关系）

（2）属性型产生式：（对象，属性，值）

“篮球是圆的”就是属性型知识，其事实性产生式为“(篮球，形状，圆形)”；“北京是中国的首都”则是关系型知识，其事实性产生式为“(中国，北京，首都)”；而对于“明天可能会下雨”这种不确定的属性型知识，其事实性产生式也和规则性产生式一样，在末尾加上置信度，具体表示为“(明天，天气，下雨，0.8)”；“π≈3.14 159”的事实性产生式则为“(π，近似值，3.14 159)”。

1.5.4 产生式系统

产生式一般由专家根据具体的知识、具体的要求来撰写，而为了运用产生式解决具体问题，在专家根据具体知识设计好产生式知识库之后，还需要进一步设计产生式系统，才能够利用产生式知识库进行问题求解。下面就对产生式系统的相关知识进行具体介绍。

电子工业出版社版权所有 盗版必究

1．产生式系统的结构

产生式系统主要包括 4 个主要部分，其结构如图 1-3 所示。

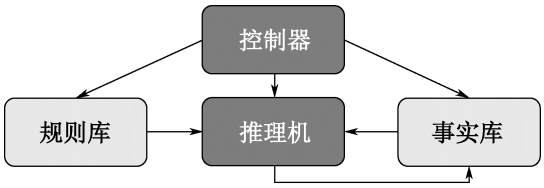


图 1-3　产生式系统结构

（1）规则库。规则库中存放相关的规则性产生式。

（2）事实库。事实库中存放已有的事实，以及通过推理得到的新的事实。

（3）控制器。控制器控制整个解决问题的流程。

（4）推理机。推理机读取事实库和规则库，将事实与规则的前件进行匹配，以产生新的事实。

2．产生式系统的推理步骤

想要通过产生式系统来判断一个断言的真伪，则需要通过以下的步骤：

（1）推理机读取事实库和规则库。

（2）推理机将事实与规则的前件进行匹配，以产生新的事实。

（3）如果新的事实中包含了待证明的断言，则推理结束。

3．应用产生式推理的例子

下面通过一个例子来进行产生式系统运用的具体介绍。图 1-4 所示是一个规则库，该规则库的作用是根据动物的特征来识别动物的种类。

该规则库一共包括 14 条关于动物分类的规则，现已全部表示为规则性产生式并进行编号。有了规则库就可以根据一些条件来判断动物的类别，现需要推理“一种动物有毛、有利齿、有爪、眼睛前视、体表有黑斑且为黄褐色。这是什么动物？”其具体步骤如下：

首先根据所给的条件，建立如图 1-5 所示的事实库，将规则库与事实库逐条匹配，产生新的事实。检测规则库，r1：“IF 有毛 THEN 哺乳动物”，其前件可以与事实库中的 r15 匹配，执行该产生式，产生“哺乳动物”的新事实。

电子工业出版社版权所有 盗版必究

r1: IF 有毛 THEN 哺乳动物
r2: IF 喂奶 THEN 哺乳动物
r3: IF 吃肉 AND 哺乳动物 THEN 食肉动物
r4: IF 有利齿 AND 有爪 AND 眼睛前视 AND 哺乳动物 THEN 食肉动物
r5: IF 食肉动物 AND 黑条纹 AND 黄褐色 THEN 虎
r6: IF 食肉动物 AND 黑斑 AND 黄褐色 THEN 豹
r7: IF 哺乳动物 AND 有蹄 THEN 有蹄动物
r8: IF 哺乳动物 AND 反刍 THEN 有蹄动物
r9: IF 有蹄动物 AND 黑条纹 AND 白色 THEN 斑马
r10: IF 有蹄动物 AND 黄褐色 AND 黑斑 AND 长颈 AND 长腿 THEN 长颈鹿
r11: IF 有羽毛 THEN 鸟
r12: IF 会飞 AND 生蛋 THEN 鸟
r13: IF 鸟 AND 不会飞 AND 黑色或白色 AND 长腿 THEN 鸵鸟
r14: IF 鸟 AND 不会飞 AND 黑色或白色 AND 会游泳 THEN 企鹅

图 1-4 产生式系统的规则库

r15:有毛
r16:有利齿
r17:有爪
r18:眼睛前视
r19:黑斑
r20:黄褐色

图 1-5 产生式系统的事实库

向事实库中添加“r21：哺乳动物”。

再次检测规则库，r4：“IF 有利齿 AND 有爪 AND 眼睛前视 AND 哺乳动物 THEN 食肉动物”，其前件可以与事实库中的 r16、r17、r18、r21 匹配，执行该产生式，产生“食肉动物”的新事实。

向事实库中添加“r22：食肉动物”。

再次检测规则库，r6：“IF 食肉动物 AND 黑斑 AND 黄褐色 THEN 豹”

向事实库中添加“r23：豹”。

最后，得到明确分类结论，推理结束。

由以上使用产生式系统推理的过程可见，规则库主要用来匹配以及产生新的事实，事实库则需要存放现有的所有事实。如果需要的结果已经在事实库中，则推理完成。

产生式在人工智能发展过程中起到了非常重要的作用，具体体现在以下 3 个方面。

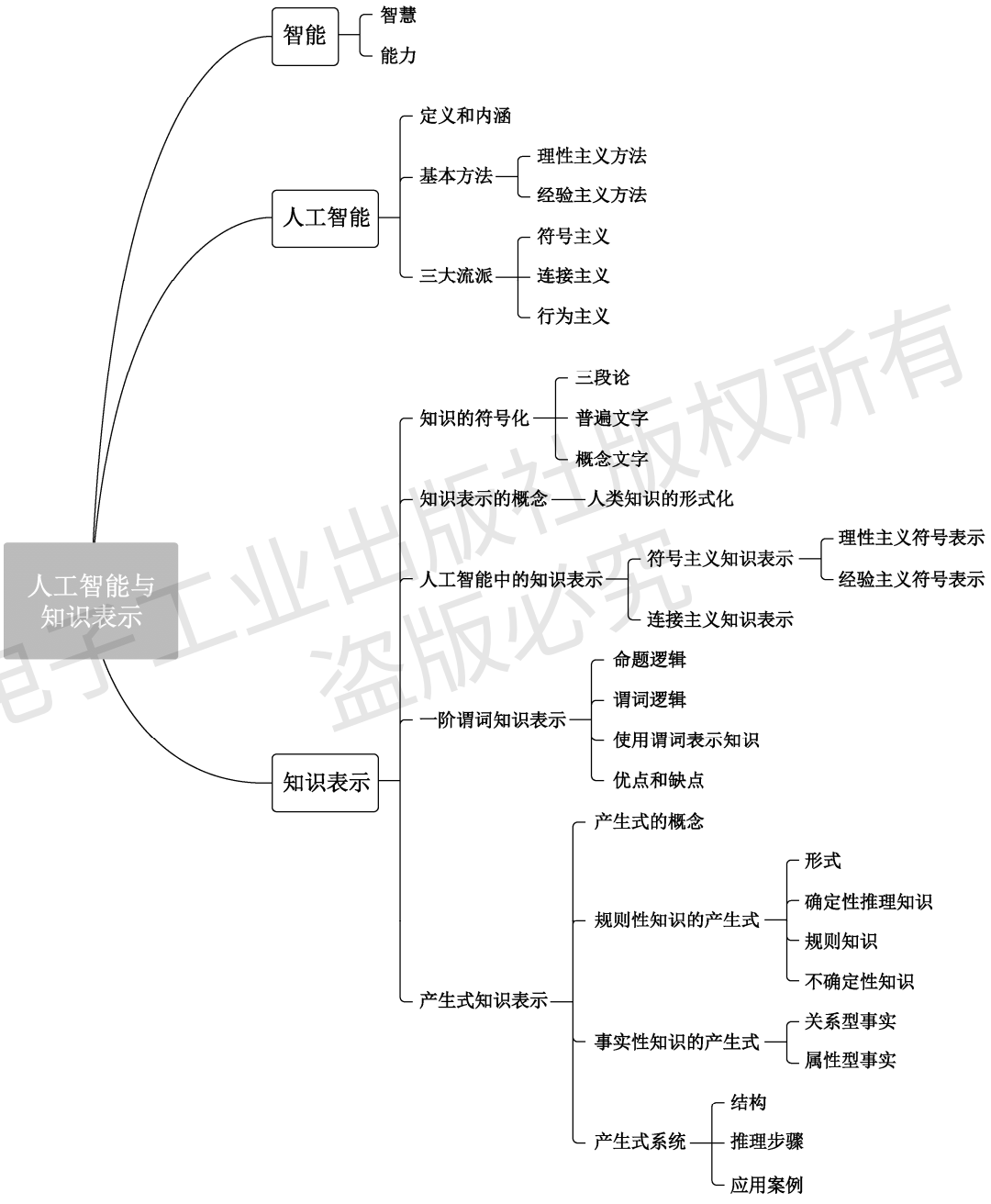
（1）在人工智能学科发展的早期，产生式类型的知识表示方法，配合推理器，成为 AI 的主流，推动了定理证明、自动推理、专家系统等方法的发展。

（2）产生式作为一种通用的人类知识的表示形式，也成功应用于许多领域，如早期的基于规则的句法分析器、机器翻译器等。

（3）事实性知识的产生式表示方法，发展为知识图谱，成为近几年人工智能领域知识表示的核心方法。

电子工业出版社版权所有 盗版必究

【本章思维导图】



电子工业出版社版权所有 盗版必究

【本章习题】

【单选题】

1. 下列不属于能将问题表达为计算机能接受的方式的流派的是：(　　)。

A. 连接主义　　B. 语言主义

C. 行为主义　　D. 符号主义

2. 以下哪种应用是在理性主义的基础上建立的：(　　)。

A. 神经网络　　B. 专家系统

C. 知识图谱　　D. 推理系统

3. 三段论推理指的是："如果所有的 B 都满足 A 且(　　)则 C 必定满足 A。"

A. C 属于 B　　B. B 属于 C

C. A 属于 C　　D. C 属于 A

4. 谓词逻辑是(　　)的知识表示方法。

A. 连接主义　　B. 联想主义

C. 经验主义　　D. 符号主义

5. 下列可以用于连接谓词表达的符号是：(　　)。

A. *　　B. /

C. !　　D. ↔

6. "世界上可能有外星人"用事实性产生式来表示可以为：(　　)。

A.（世界，存在，外星人）

B.（外星人，存在，世界）

C.（世界，存在，外星人，0.003）

D.（世界，存在，外星人，-0.003）

7. 在产生式系统中，负责存放相关的规则性产生式的是(　　)。

A. 事实库　　B. 规则库

C. 推理机　　D. 控制器

【判断题】

8. 知识库是一种知识表示的形式。　　(　　)

电子工业出版社版权所有 盗版必究

9．命题是指对客观世界的带有真假意义的确定性的疑问句。（　　）

10．命题可以是复合的论述。（　　）

11．“香蕉是一种水果”“苹果是一种水果”这两种表述在用符号表示命题逻辑时可以有相似的表示。（　　）

12．将命题形式化为以主语为中心，谓词为主语附属成分的形式称为谓词逻辑。（　　）

13．在谓词逻辑的表示形式中，语句中可以出现多个谓词。（　　）

【填空题】

14．人工智能区别于自然智能，是一种由_____产生的智能现象。

15．人工智能研究的三大流派是符号主义、_____、行为主义。

16．理性主义和经验主义的区别在于，智能机器获得_____的方式不同。

17．理性主义主要采用_____来表示知识。

18．连接主义中的知识表示形式通常是语义向量以及_____。

19．最简单的命题至少由一个主语和一个_____组成。

20．产生式根据表示的知识的形式涉及对象的数量不同可分为_____和属性型产生式。

21．事实性知识的产生式表示方法，发展为_____，成为近几年人工智能领域知识表示的核心方法。

电子工业出版社版权所有 盗版必究