概述：

A
人工智能的三个发展阶段：计算智能、感知智能、认知智能。知识图谱是实现认知智能的关键技术。
一个实例：IBM Watson 在问答游戏中击败人类，背后用到知识图谱。

知识图谱的表述(知识的表示)：实体和关系，关系可以分为实体内在属性和实体之间的关系。如果用图来表示，那么节点是实体，节点的属性挂在节点上（自环），节点间的关系为节点间的连边。

知识图谱可以用于检索、推理、人机交互等场景。
一个实例：Google 通过知识图谱改善搜索引擎的结果

知识图谱中存储的知识量远远超过一个人的知识量，但是目前的知识图谱无法用它已有的知识进行有效的推理等下游任务。

知识图谱天生缺乏有效的处理工具，无法直接进行推理等下游任务（少量的语义，丰富的实例）。

知识的定义(根据维基百科)：知识是通过经验、教育、探索或学习得到的关于人或事物，比如人事实、信息、描述、技能，的理解、通晓或认识。

知识图谱：实体+链接

知识表示：概念图：有数学和逻辑支撑的逻辑表示

本体：领域共享知识的描述方式。具体来说，就是定义的实体类型、属性和实体关系。

经典知识图谱：
Cyc 常识知识库 术语（概念、实体、关系的定义）+断言
WordNet 词典知识库 主要用于语义消歧 词及其语义关系
ConceptNet 常识知识库 三元组
FreeBase 三元组
WikiData 三元组
DBPedia 三元组
YAGO 三元组 是 IBM Watson 后端知识库之一
NELL 互联网自动挖掘 三元组
OPENIE
ZhiShi.me 中文知识图谱 百度百科、互动百科、维基百科
CN-DBPedia
BabelNet 多语言词典知识库

B
知识图谱应用
辅助搜索：精准检索　语义检索
辅助问答：人机互动　多轮对话
辅助数据集成：智能数据整合 大规模异构数据集成机制
辅助决策：智能推理 知识推理

知识图谱技术的核心价值
集成异构数据源　图模型对异构数据灵活集成
描述数据间的关联 实体关系 事件
实现实体链接 知识和实体间的桥梁 更理智的智能
大规模知识推理 发现隐含知识 可解释智能

微软 Concept Graph IsA 关系 用于短文本理解

C
构建知识图谱是一件复杂的系统工程

任务与问题：
数据层：多源数据、多模态数据、多媒体数据
信息层：实体识别、关系抽取、事件抽取(数据爬取、信息抽取)
知识层：实体识别、类型推断、本体构建
挖掘层：表示学习、实体链接、实体消解、链接预测

知识表示

如何存储知识，让程序能够处理。
将知识（信息）表示为机器能够处理的模式，以使机器模拟人对世界的认知和推理。
一种知识表示的分类观点：（1）基于非逻辑的知识表示（2）基于数理逻辑的知识表示（3）基于统计学习的分布式表示

表达能力和推理能力的权衡

描述逻辑：概念、关系、实例、公理（不懂）

RDF：最低限度的约束

知识建模

本体意味着存在一个外在的完美知识体系，不依赖于人的认识而存在。

本体构建：

1、确定本体的领域和范围
2、考虑重用现有本体
3、列出本体中的重要术语
4、定义类和类的继承
5、定义属性和关系
6、定义属性的限制
7、创建实例

知识图谱中包括本体层和实例层，本体通常手工构建，实例自动抽取
构建本体的目的是为了确定知识图谱能描述的知识

知识抽取

信息抽取是从半结构化或无结构化的文本中抽取结构化信息。
知识抽取是从结构、半结构化的文本中提取出信息，并把信息表示成机器可以理解的格式，并且这种表示易于进行推断。

知识抽取=信息抽取+信息表示

知识抽取的挑战：
知识的不确定性 beetles, Beetles, Beables; citizenof livedin bornin
知识的不完备性 关系缺失 属性缺失 实体缺失
知识的不一致性 互斥的标签 alive dead

结构化信息：从关系数据库中抽取知识
类、属性、实例、实例间的关系

半结构化信息：百科知识库 对特定百科使用特定模板

无结构化信息：
是当前知识图谱构建的技术瓶颈
关键技术：实体识别、关系抽取、事件抽取。
Pipeline 的抽取过程会迅速降低知识的质量

事件抽取：从数据中抽取事件信息，并以结构化和语义化形式展示。例如：事件的时间、地点、人物、原因

数据获取：网络爬虫

知识抽取：命名实体识别
基于规则和词典：划分句子、分词+词性标注、构建词典 识别实体边界 命名实体分类
词典：辅助分词、匹配实体、对实体分类

基于机器学习、神经网络

半监督方法：无标注语料用于获取好的 embedding

迁移学习：迁移学习的关键在于找到新问题和原问题的相似性。跨域、跨应用、跨语言。
T-A T-B T-C 共享的组件不同。T-C 只共享 embedding。

预训练模型：

知识抽取：关系抽取
隐藏在句法结构后面，由词语的语义范畴建立起来的关系

句法关系：位置关系、替代关系、同现关系

知识抽取：事件抽取
事件描述、事件触发词、事件元素、事件角色
强化学习进行事件抽取：教会机器识别错误的 label

知识融合

本体匹配、实例对齐

知识工程的理想是构建一个统一的知识库，但人类知识体系复杂、知识含有某些主观看法、有些知识随时间变化、不同组织在构建知识库。

知识图谱优先考虑重用现有知识库

结论：知识具有共享性的同时还有自治性和动态性，因此不可避免的遇到知识异构的现象。

知识异构的两个层次：语言层异构（语法、逻辑、表达能力不匹配）、模型层异构（概念解释不匹配）

知识图谱：一定的抽象层知识+大量的实例层事实

重用现有的大量知识

匹配器：
字符串匹配器： 计算编辑距离 动态规划； 最大公共子序列；Dice 系数；Jaccard 相似度；N-gram；
文本匹配器：tf-idf 将文档转化为向量；
将待匹配对象的相关文本组织成文档形式，再转化为文档向量。实际中非常有效。

虚拟文档

结构匹配器：利用本体的结构信息弥补文本信息量不足的情况
不采用图匹配的技术，复杂度太高；采用相似度传播模型很有效
对文本信息量少、文本信息不规范有帮助

结构匹配是解决若信息场景的有效方法。

匹配调谐

实例匹配：
真实图谱中可用于匹配的语义信息非常有限。

匹配器不是知识融合的关键，不同匹配器的效果差异不大

知识图谱表示学习

表示学习：将讲究对象（文字、图像、语音等）的语义信息表示为稠密低维的实值向量。
知识表示学习：将知识库中的实体和关系表示为稠密低维的实值向量。
知识通常用三元组表示(head, relation, tail)

One-hot 表示：假设所有对象都是独立的，每个对象只在自己的那一维为 1，其余维度为 0。

多源异质信息表示形式统一，便于迁移和融合。

翻译模型 基本出发点是向量空间 h+r=t 有诸多变形
语义匹配模型
矩阵分解

文本描述

知识表示学习的挑战：
在线学习、融合异构数据进行知识表示

知识存储

关系数据库、图数据库、RDF 三元组

知识问答

对一个输入：实体识别、指代消歧、省略补全
情感向量
70%回答依靠从语料中检索。关键词检索、语义检索
回答生成：seq2seq
深度参与、任务完成

实体链接

命名实体链接+词义消歧

基于篇章主题

跨语言实体链接
利用实体链接促进自然语言处理任务

知识推理

知识图谱补全

面向多元关系的知识推理
融合多源信息与多种方法的知识推理
基于小样本学习的知识推理
动态知识推理