一、知识图谱实践的系统工程观念

1. 工程观

知识图谱实践是一种典型的大规模知识工程，在实践过程中需要坚持工程观念。工程观是利用科学原理提出有效方案解决实际问题的观念。基础的自然科学以认识世界为基本使命，而工程学科旨在改造世界。知识图谱的工程观要求我们具备优化问题的求解思路，解决行业智能化升级中的一系列实际问题。

使命

工程观的使命是利用知识图谱解决行业智能化升级转型过程中涌现的一系列实际问题。实践者需要提出优化问题的求解思路，以应对有限资源（如人力、资金、数据、算力）和紧迫的实际问题。优化问题求解思路包括如何在有限的资源条件下，设计出解决问题的最优方案。

要求

能否解决这些问题以及如何解决这些问题是知识工程研究者和实践者面临的迫切问题。工程观要求实践者具备优化问题的求解思路，明确优化问题中的约束，建立合理的优化目标，提出性价比高的方案。这意味着实践者需要具备综合分析和解决问题的能力，能够在资源有限的情况下找到最佳的解决方法。

方法

在知识图谱建设中，目标知识图谱的规模、粒度、精度等都是优化目标需要考虑的因素。例如，手动构建知识图谱的成本高，而自动构建的成本较低，但仍需优化资源使用。在实际工程实践中，需要明确优化问题中的约束，建立合理的优化目标，提出性价比高的方案。例如，手动构建知识图谱时，每个三元组的成本约为2至6美元，而自动构建的成本大约为其1/250至1/15，这样每个三元组仍需耗费1至15美分。因此，在知识图谱相关的工程实践中，需要注重优化问题和实际约束。

2. 系统观

系统观认为现实世界中的大部分复杂系统由相互作用、相互依赖的若干组成部分组合而成的有机整体。知识图谱系统是一个典型的复杂系统，包含众多组件和多样的要素，需要复杂的人机协作。

特性

知识图谱系统具有涌现性、交互性、演化性等特征。涌现性是指系统整体表现出各组成部分所不具备的特性，即1+1>2的效果。交互性和演化性强调知识图谱系统与外部数据和应用的复杂交互，系统需要随着环境和需求的变化进行动态调整。例如，智能医疗系统要求高准确率，虽然每个独立的NLP模型难以达到要求，但通过人机协同和多模型组合，能够实现近乎100%的应用准确率。

复杂系统

知识图谱作为系统化整体并不是若干组件的简单组合，而是复杂策略指引下的有机组合，其效果取决于组合策略。在当前NLP技术仍不能完全有效完成抽取任务的情况下，充分利用各类资源、有效利用已有的业务知识和人力因素进行验证或标注，均对知识图谱落地效果有显著影响。例如，在知识图谱系统中，明确各组成部分及其相互关系，能够更好地协调系统的整体功能，实现复杂任务的有效执行。

二、领域知识图谱框架构建

方法

领域知识图谱框架的构建需要调研现有框架，完成目标框架的人工构建。主要参考资源包括Schema.org、百度百科和互动百科的分类体系。这些资源提供了分类和属性的基础，可以帮助构建一个包含多层次分类和属性的知识框架。

框架构建

构建的知识图谱体系包括4个一级类别、28个二级类别和238个三级类别。一级类别包括创造性工作、组织、人物和地点等。每个一级类别下包含多个二级类别和三级类别，涵盖了广泛的实体和关系。例如，创造性工作类别下可以包括文学、艺术、科学研究等二级类别，而在文学类别下可以进一步细分为小说、诗歌、戏剧等三级类别。

三、半结构化文本中的知识抽取

目标

半结构化文本中的知识抽取主要从百科普通条目半结构化网页中抽取实体属性名和属性值。这些半结构化信息块通常包含明显的结构标签，如“xxx：”，可以通过模板和规则进行抽取。

步骤

1. **定位半结构化信息块**：从百科文本中定位半结构化信息块，假设“xxx：”的结构连续出现，可以认定该区域为半结构化信息块。例如，百度百科页面中常见的“人物信息：姓名、出生日期、职业”等信息块，可以通过识别这些特征定位出相关的信息块。
2. **学习抽取模板**：利用结构化信息中的属性名进行定位，并学习相应的抽取模板。通过分析这些信息块的格式，提取出属性名前后的固定格式，从而构建抽取模板。
3. **抽取属性名和属性值**：根据模板和属性名进行属性名和属性值的抽取，优化模板的质量。例如，在一个人物条目中，模板可以识别出“出生日期：1990年1月1日”中的“出生日期”作为属性名，“1990年1月1日”作为属性值。

四、非结构化文本中的知识抽取

方法

非结构化文本中的知识抽取涉及文本分割、清扫、训练数据挑选和模型训练。具体步骤如下：

1. **数据预处理**：将百科页面分为“百科名片、infobox、正文文本、开放分类”等部分，去除无关信息，提取纯文本。例如，一个百科页面可以分割成“概述、主要内容、补充内容”等部分，并去除广告、编辑者信息等。
2. **训练模型**：利用CRFs（条件随机场）进行实体属性值的抽取，训练句子分类器判别抽取的属性值是否准确。例如，通过对大量标注好的样本进行训练，建立模型来识别文本中的实体和属性。
3. **后处理**：处理多属性问题，汇总去重抽取结果。例如，一个实体可能有多个属性值，需要通过去重和筛选，保留最准确的属性值。

训练数据挑选

挑选训练数据时，需要根据百科的“开放分类”将不同页面分类，并进行KNN分类筛选高质量页面。通过这种方法，可以有效减少噪声数据，提高训练数据的质量。

实体属性回标

通过回标规则，从infobox中属性值匹配到文本中的句子，产生用于训练的语料。例如，从infobox中提取出“出生地：上海”，并在文本中找到包含“上海”的句子进行标注，形成训练样本。

五、知识图谱众包构建

1. 众包介入阶段

知识图谱的构建涉及元知识创建、知识获取和知识精化三个阶段，众包可以在这些阶段中发挥重要作用。

元知识创建

由专家搭建基本知识框架，涉及深层次语义理解和知识体系的设计。例如，建立一个领域的本体模型，定义相关的概念和关系，这些工作通常需要领域专家的参与。

知识获取

利用众包实现数据标注，构建知识获取模型，通过这些模型从文本或数据中自动获取知识。例如，众包工人可以标注大量的文本数据，提供实体、关系等信息，训练机器学习模型自动抽取知识。

知识精化

众包手段用于验证和纠错补漏，确保知识图谱的质量和准确性。例如，通过众包平台，让大量工人对自动抽取的知识进行验证和修正，提高知识图谱的整体质量。

2. 知识型众包特点

知识型众包任务多样性强，工人多样性强，任务质量难以评价且影响面大。

核心问题

- **筛选任务**：挑选最重要和机器最不擅长的任务进行众包。重要任务需要优先处理，而机器不擅长的任务需要人类的智慧来解决。
- **选择工人**：根据任务需求精挑细选工人。不同的任务对工人的技能要求不同，需要根据任务特点选择合适的工人。
- **设计工作流**：优化任务设计、激励和质量控制。合理设计任务流程，提供激励措施，确保任务质量。

六、知识图谱质量控制

1. 质量评估维度

知识图谱的质量评估维度包括准确性、一致性、完整性和时效性。

准确性

考察知识图谱中各类知识的准确程度，通常通过与黄金标准数据比对或领域专家抽样检查进行评估。例如，通过将知识图谱中的实体和关系与专家标注的标准数据进行对比，评估其准确性。

一致性

检测知识图谱中的知识表达是否一致，避免存在互相矛盾的知识。例如，确保同一实体在不同条目中的描述一致，避免出现同一个实体被描述为两个不同的对象。

完整性

考察知识图谱对某领域知识的覆盖程度，评估相关领域中的所有知识是否被包含。例如，通过分析知识图谱中的实体和关系，评估其是否涵盖了领域中的主要概念和关系。

时效性

考察知识图谱中的知识是否为最新知识，确保动态变化的知识及时更新。例如，对于时效性要求较高的信息，如最新的研究成果、市场动态等，需要及时更新以保持知识图谱的准确性。

2. 质量评估方法

- **人工抽样检测法**：领域专家进行抽样质量检测与评估。专家通过抽取一定比例的知识图谱内容进行检查，评估其质量和准确性。
- **一致性检测法**：通过预定义的一致性检测规则发现知识冲突。例如，制定规则检查同一实体在不同条目中的描述是否一致，发现并解决矛盾。
- **基于外部知识的对比评估法**：利用高质量外部知识源进行质量检测。例如，将知识图谱中的知识与可信的外部知识库进行对比，评估其一致性和准确性。

七、实践案例

背景与意义

基于前期"百科在线工程"成果，解决细粒度知识自动抽取和百科知识服务问题，构建知识图谱。例如，利用现有的百科数据，建立一个结构化的知识图谱，提供更高效的知识检索和服务。

项目框架

包括知识框架构建、实体分类、属性抽取、实体消歧和知识显示等环节，构建全面的知识图谱体系。例如，项目中首先构建一个知识框架，定义各类实体和关系，然后通过自动抽取和人工标注，完善知识图谱内容，最后通过可视化平台展示知识图谱，实现知识的检索和应用。

八、实践经验与基本原则

合理定位

设定合理目标，让机器代替专家助理工作，而非完全代替领域专家。例如，在医学领域，知识图谱可以帮助医生进行初步诊断和推荐，但最终的决策仍需医生来做。

应用牵引

从应用出发，明确技术需求，适配应用，推动知识图谱技术的发展。例如，在金融领域，知识图谱可以用于风险评估和投资决策，技术开发需要满足这些具体应用需求。

循序渐进

技术体系发展呈现出部分技术先成熟再逐步带动相关技术发展的特点，整体发展需要经历漫长周期。例如，先实现基础的实体识别和关系抽取，再逐步引入高级的知识推理和动态更新技术。

先易后难

从结构化程度高的数据中抽取知识，逐步处理复杂的知识，降低知识获取的难度。例如，先从表格数据、数据库等结构化数据中抽取知识，再处理文本、图像等复杂数据。

由粗到细

知识表示有粒度粗细之分，应遵循由粗到细、逐步求精的原则，逐步完善知识表示。例如，先构建一个粗粒度的知识框架，定义主要的实体和关系，再逐步细化，增加详细的属性和子关系。

人机协同

当前知识图谱的构建需要机器和人类协同工作，充分利用各自优势。例如，机器可以处理大量的基础知识抽取和整理工作，而人类可以进行高层次的知识分析和验证。

快速启动

利用已有的知识资源，快速构建相关领域的知识图谱，降低构建成本。例如，利用现有的领域本体、叙词表等资源，快速建立基础的知识图谱，再通过自动抽取和人工标注进行完善。

九、大模型时代知识图谱构建实战

大模型在知识图谱构建中的应用

利用大模型的知识能力进行本体学习、事件本体构建和信息抽取任务。大模型能够胜任本体构建任务，通过指令学习和ICL（In-Context Learning）等技术，提升知识图谱构建的效率和准确性。例如，利用大模型的预训练能力，可以快速学习和理解大量的领域知识，应用于知识图谱的构建。

实例

CogIE工具包的使用，通过多样化的指令数据微调大语言模型，使其适应任务形式，具备域外泛化性，支撑CogNet的构建和更新。例如，通过CogIE工具包，可以实现命名实体识别、细粒度实体分类、实体链接、关系抽取、事件抽取和框架语义解析等功能，有效提升知识图谱的构建效率和质量。