如何协调准确率和效率、平衡自动化方法和人工参与,以最高效的方式构建出最准确的知识图谱,是目前构建知识图谱需要解决的一大难题.

四步法：

步骤 1:领域本体构建; 本体构建是指构建出知识图谱的本体结构,本体结构可以理解为知识图谱的框架.

步骤 2:众包半自动语义标注; 众包半自动语义标注指的是将文本页面众包给多个标注者,根据步骤1构建好的本体,利用语义标注工具标注得到高质量的标注数据.

步骤 3:外源数据补全; 外源数据补全是指将其他来源的结构化程度较好的数据按照本体结构处理后,与标注数据整合在一起.

步骤 4:信息抽取；针对知识图谱中较为稀疏的实体或者关系,从文本中进行大规模的抽取和补充

1.本体的概念：共享概念模型的明确形式化规范说明，具有共享化、明确化、概念化和形式化这 4 个基本特征。

**2**.**本体构建**有人工、自动和半自动三种构建方法，本文采用统计方法和无监督方法（半自动）

1. 人工构建：由大量领域专家协作完成
2. 自动构建：也称为本体学习，利用知识获取技术、机器学习技术、统计技术自动获取本体知识。（A survey on ontology learning research）。
3. 半自动构建：自动构建本体的过程中，在用户指导下进行。

**3.语义标注**

根据标注结果的存储方式分为嵌入式存储和独立存储；根据自动化程度分为手工标注、半自动标注和自动标注。

本文采取基于pundit开发的众包半自动语义标注工具。

4.**信息抽取**

1. 实体抽取：bootstrapping
2. 关系抽取：基于规则和方法和基于机器学习的方法（有监督、半监督和无监督）
3. 属性抽取可以视为关系抽取

**5.本体构建具体过程**

1）归纳领域概念

领域的核心概念对应本体中的类（owl:class），三种方法得到核心概念：

1. 统计方法得到领域术语，再从领域术语中得到领域核心概念；通过关键词提取算法得到重要术语（TF-IDF算法和 TextRank算法）
2. 参考质量较高的知识图谱或数据源，（Schema.org(https://schema.org.cn)、DBpedia(http://mappings.dbpedia.org/server/ontology/classes/) 和 Geonames.Schema.org）
3. 众包半自动语义标注步骤中进行完善补充。

2）定义领域关系及其约束

主要可以通过四种方法来定义关系：

1. 利用OpenIE工具对文本进行无监督的开放关系抽取
2. (https://stanfordnlp.github.io/CoreNLP/openie.html）
3. 参考质量较高的知识图谱或数据源
4. Wikidata(https://www.wikidata.org/wiki/Wikidata:Main\_Page)和

Schema.org.Wikidata。

1. 根据核心概念和百科的信息框来确定关系。核心概念下的实例再百科中都有对应的信息框，信息框中相同的关系就比较重要。
2. 众包半自动语义标注过程中补充新的关系.

3）领域专家对本体进行检查和评估