This the note for 知识图谱技术综述_徐增林(2016)

知识图谱构建的方式

自顶向下：先定义本体和模式，再将实体加入到知识库（知识图谱是知识库的整合）

自底向上：先从开放数据库中抽取实体，选择置信度较高的加入到知识库，在通过聚类等方式构建顶层的本体模式。

本体：概念的抽象。比如人是本体，警察、医生是具体的概念。

知识图谱四大技术：知识获取、知识表示、知识融合、知识推理

知识获取

知识要素：实体、关系（外在关系）、属性（内在属性）

实体抽取：监督学习（CRF）与规则结合的效果比较好。

关系抽取：目前研究的大多是二元关系，但英文语句中有40%是n元（多余二元）的关系。

属性抽取：规则+启发式算法效果较好

知识表示

除了三元组，还可以用在向量空间表示：分布式表示。用一个综合的向量表示实体对象的语义信息。

距离模型

将实体向量投影到同一维度的向量空间，再计算距离判断关系。（一个关系处于一个子空间中）

单层神经网络

用一个网络实体到关系的映射。对一个关系，评价函数为

$$f(h,t)=\mu^T g(M_h h+M_t t)$$

h为head，t为tail，g为激活函数，M为线性映射矩阵，u为关系的向量表示（也是参数）。

双线性模型

$f(h,t)=h^T Mt$ 当M为对称矩阵时，相当于往关系空间上投影，拉伸后再内积。 M为对角阵是效果会提升？

神经张量模型

结合了双线性模型和单层神经网络。将神经网络内层的线性变换改成双线性模型（需要n个双线性变换以保证维度一致）。

矩阵分解模型

[h,r,t]的关系构成的矩阵：1为有关系，0为无关系。用双线性模型拟合这个张量。

翻译模型

TransE

希望关系向量r=t-h，由此构建损失函数|h+r-t|

TransH

希望h,t在于r正交的方向上无差异。有归一化的意味。

TransR

希望在关系空间中满足transE的关系

TransD

分别定义h，t在r中的投影空间

TransG

一种关系可能对应多种语义，每个语义用一个高斯分布表示，由此得到了高斯混合模型。

KG2E

用高斯分布刻画r向量和h-t向量，通过衡量分布间的相似度来判断是否有关系。

知识融合

主要解决的问题是实体对齐、本体构建、质量评估

知识推理

基于一阶谓词逻辑（目前不太懂）

基于图，通过图中两个实体间的多步路径来预测它们之间的语义关系。

知识图谱应用场景

智能搜索、深度问答（这两个都是先理解，再检索）、在社交网络中查询、特定邻域应用（金融反欺诈、医疗、电商提供服务、体验）

挑战

由于目前的方法一般只在某些领域效果好，大规模开放知识抽取的研究处于起步阶段。

跨语言知识抽取方法。

除了三元组外的知识表示方法。

浮复杂知识的表示（N-to-N的关系）。

多源信息融合。

并行分布式算法。

众包算法。

跨语言知识库对齐。