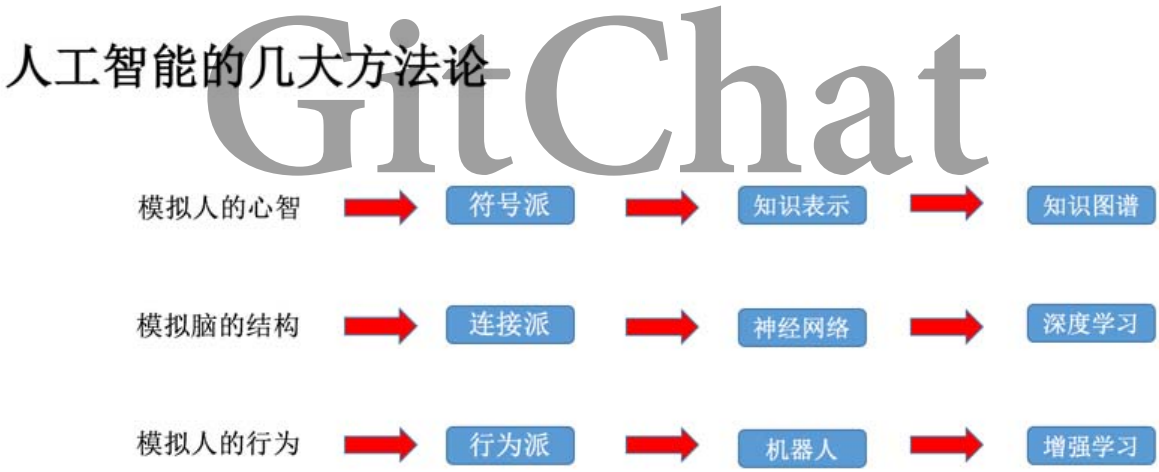
大规模知识图谱的构建、推理及应用

随着大数据的应用越来越广泛，人工智能也终于在几番沉浮后再次焕发出了活力。除了理论基础层面的发展以外，本轮发展最为瞩目的是大数据基础设施、存储和计算能力增长所带来的前所未有的数据红利。

人工智能的进展突出体现在以知识图谱为代表的知识工程以及以深度学习为代表的机器学习等相关领域。

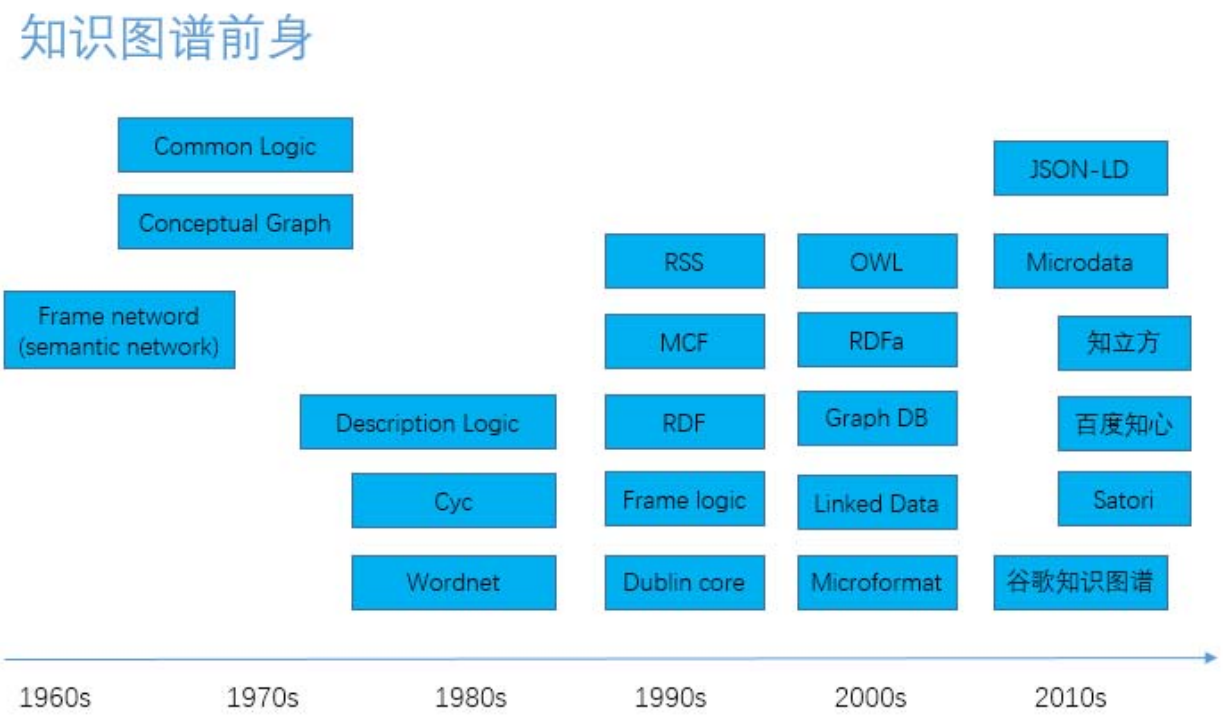
未来伴随着深度学习对于大数据的红利消耗殆尽，如果基础理论方面没有新的突破，深度学习模型效果的天花板将日益迫近。而另一方面，大量知识图谱不断涌现，这些蕴含人类大量先验知识的宝库却尚未被深度学习有效利用。

融合知识图谱与深度学习，已然成为进一步提升深度学习效果的重要思路之一。以知识图谱为代表的符号主义，和以深度学习为代表的联结主义，日益脱离原先各自独立发展的轨道，走上协同并进的新道路。



一、 大规模知识图谱的构建

知识图谱自上世纪60年代从语义网络发展起来以后，分别经历了1980年代的专家系统、1990年代的贝叶斯网络、2000年代的OWL和语义WEB，以及2010年以后的谷歌的知识图谱。谷歌目前的知识图谱已经包含了数亿个条目，并广泛应用于搜索、推荐等领域。



知识图谱的存储和查询语言也经历了历史的洗涤，从RDF到OWL以及SPARQL查询，都逐渐因为使用上的不便及高昂的成本，而被工业界主流所遗弃。图数据库逐步成为目前主要的知识图谱存储方式。

目前应用比较广泛的图数据库包括Neo4J、graphsql、spark graphx（包含图计算引擎）、基于hbase的Titan、BlazeGraph等，各家的存储语言和查询语言也不尽相同。实际应用场景下，OrientDB和postgresql也有很多的应用，主要原因是其相对低廉的实现成本和性能优势。

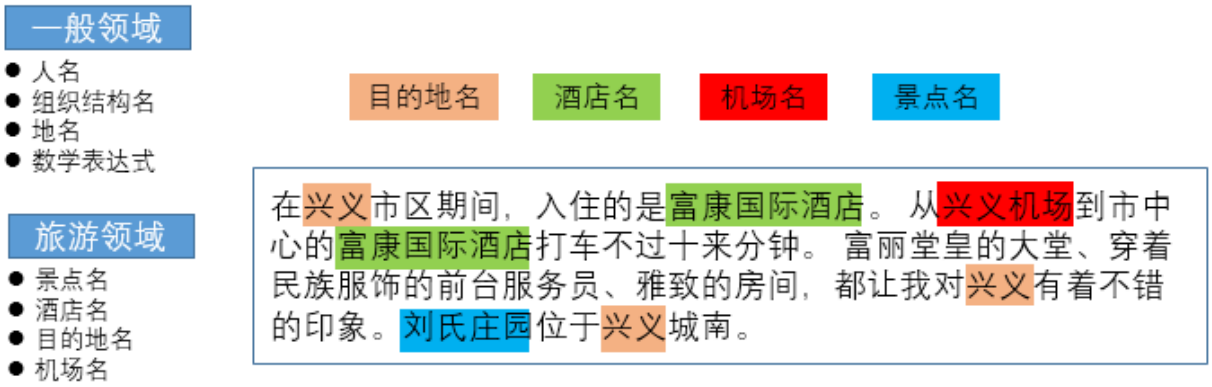
由于大规模知识图谱的构建往往会有众多的实体和关系需要从原始数据（可以是结构化也可以是非结构化）中被抽取出来，并以图的方式进行结构化存储，而我们依赖的原始数据往往存在于多源异构的环境中，所以进行海量知识抽取和融合，就成了首要的无法回避的严峻问题。

对于结构化的数据转换为图结构是比较容易和相对轻松的工程 ，所以建议这一步应该首先被完成。

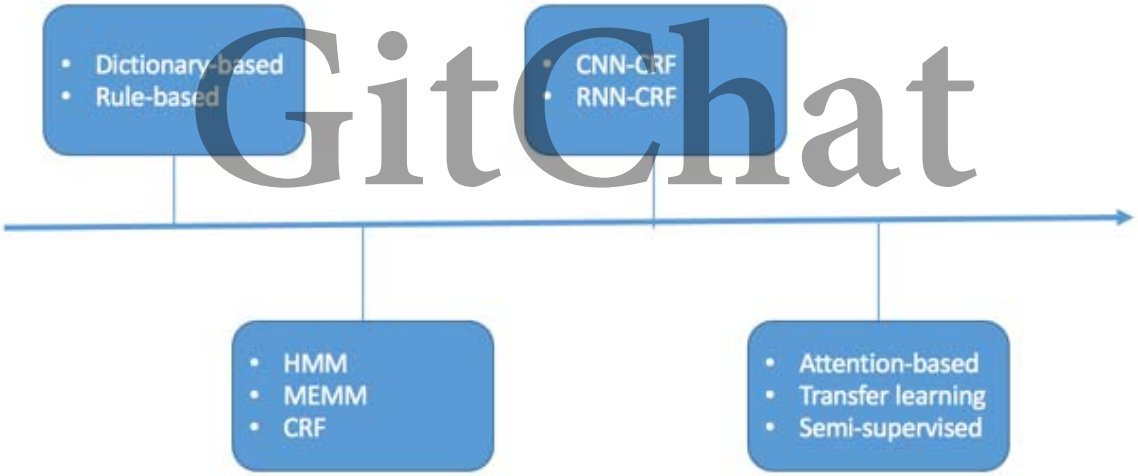
对于复杂的非结构化数据，现阶段进行知识图谱构建的主要方法有传统NLP和基于深度学习模型两类方法，而目前越来越多倾向于使用深度学习来抽取AVP（属性-值对）。

有很多深度学习模型可以用来完成端到端的包括命名实体识别NER、关系抽取和关系补全等任务，从而构建和丰富知识图谱。

命名实体识别（Named Entity Recognition, NER）是从一段非结构化文本中找出相关实体（triplet中的主词和宾词），并标注出其位置以及类型，它是NLP领域中一些复杂任务（如关系抽取、信息检索等）的基础。



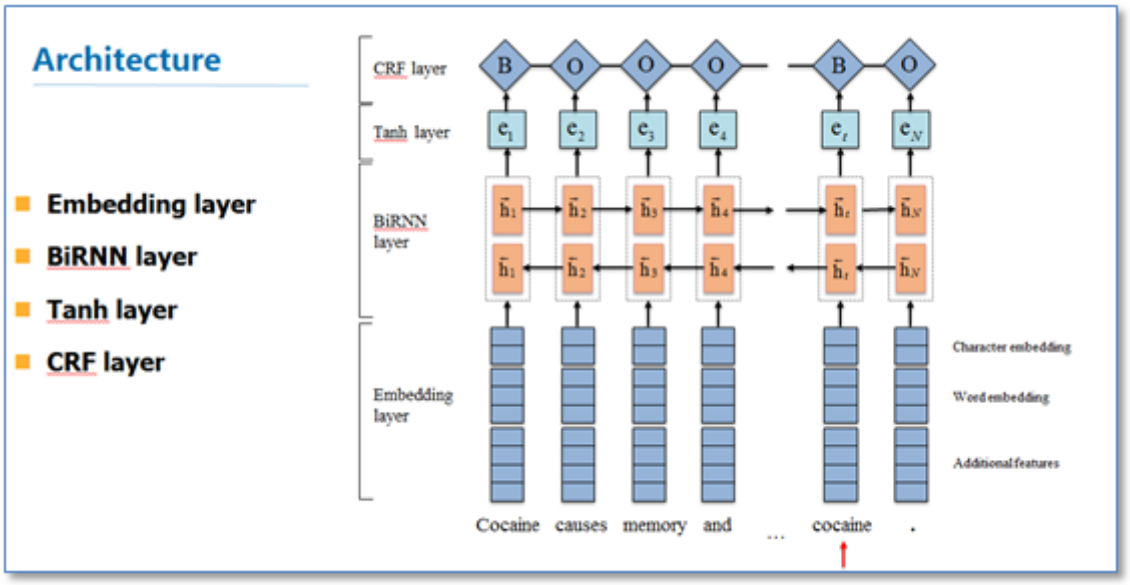
NER一直是NLP领域的热点，从早期基于字典和规则的方法，到传统机器学习的方法，再到近年来基于深度学习的方法，NER方法的大致演化如下所示。



在机器学习中，NER被定义为序列标注问题。不同于分类问题，序列标注问题中的预测标签不仅与输入特征有关，还与之前的预测标签有关，也就是预测标签之间存在相互依赖和影响。

条件随机场（Conditional Random Field，CRF）是序列标注的主流模型。它的目标函数不仅考虑输入的状态特征函数，还包含了标签转移特征函数。在训练的时候可以使用SGD学习参数。在预测时，可以使用Vertibi算法求解使目标函数最大化的最优序列。

目前常见的基于深度学习的序列标注模型有BiLSTM-CNN-CRF。它主要由Embedding层（词向量、字向量等）、BiLSTM、tanh隐藏层以及CRF层组成（对于中文可以不需要CNN）。我们的实验表明BiLSTM-CRF可以获得较好的效果。在特征方面，由于秉承了深度学习的优点，所以无需特征工作的铺垫，使用词向量及字向量就可以得到不错的效果。



近几个月来，我们也在尝试使用Attention机制，以及仅需少量标注样本的半监督来进行相应的工作。

在BiLSTM-CRF的基础上，使用Attention机制将原来的字向量和词向量的拼接改进为按权重求和，使用两个隐藏层来学习Attention的权值，这样使得模型可以动态地利用词向量和字向量的信息。同时加入NE种类的特征，并在字向量上使用Attention来学习关注更有效的字符。实验效果优于BiLSTM-CRF的方法。

这里之所以用了大量篇幅来说NER的深度学习模型，是因为关系抽取模型也是采用同样的模型实现的，其本质也是一个序列标注问题。所以这里不再赘述。

知识图谱构建中的另外一个难点就是知识融合，即多源数据融合。融合包括了实体对齐、属性对齐、冲突消解、规范化等。对于Open-domain这几乎是一个举步维艰的过程，但是对于我们特定旅游领域，可以通过别名举证、领域知识等方法进行对齐和消解，从技术角度来看，这里会涉及较多的逻辑，所以偏传统机器学习方法，甚至利用业务逻辑即可覆盖大部分场景。

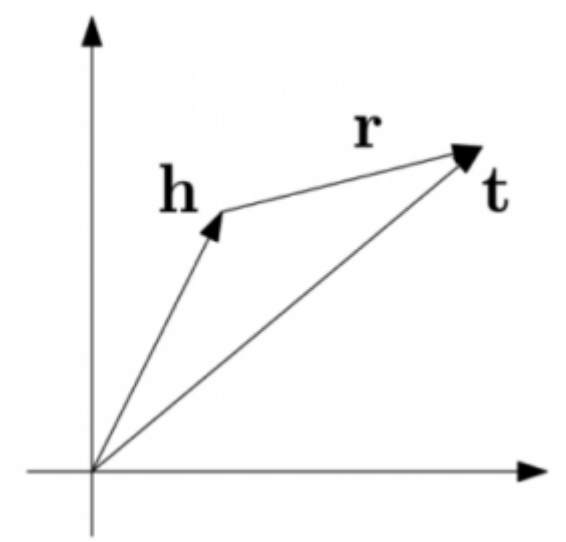
知识图谱schema是知识的分类体系的表现，还可以用于逻辑推理，也是用于冲突检测的方法之一，从而提高知识图谱的质量。

总而言之，构建知识图谱没有统一的方法，因为其构建需要一整套知识工程的方法，需要用到NLP、ML、DL技术，用到图数据库技术，用到知识表示推理技术等。知识图谱的构建就是一个系统工程，而且知识的更新也是不可避免的，所以一定要重视快速迭代和快速产出检验。

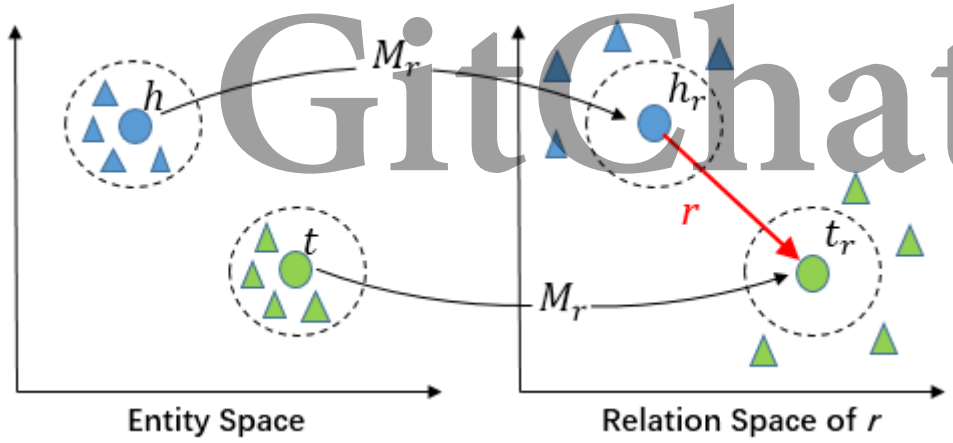
二、 知识图谱的推理

在知识图谱构建过程中，还存在很多关系补全问题。虽然一个普通的知识图谱可能存在数百万的实体和数亿的关系事实，但相距补全还差很远。知识图谱的补全是通过现有知识图谱来预测实体之间的关系，是对关系抽取的重要补充。传统方法TransE和TransH通过把关系作为从实体A到实体B的翻译来建立实体和关系嵌入，但是这些模型仅仅简单地

假设实体和关系处于相同的语义空间。而事实上，一个实体是由多种属性组成的综合体，不同关系关注实体的不同属性，所以仅仅在一个空间内对他们进行建模是不够的。



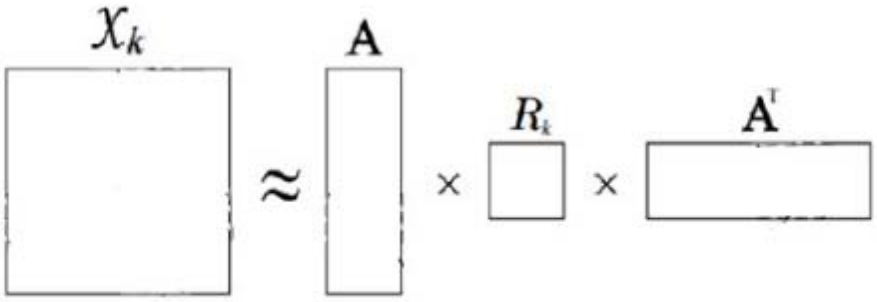
因此我们尝试用TransR来分别将实体和关系投影到不同的空间中，在实体空间和关系空间构建实体和关系嵌入。对于每个元组（h,r,t），首先将实体空间中的实体通过Mr向关系r投影得到hr和tr，然后是hr+r ≈tr。特定的关系投影能够使得两个实体在这个关系下真实地靠近彼此，使得不具有此关系的实体彼此远离。



知识图谱推理中还经常将知识图谱表示为张量tensor形式，通过张量分解（tensor factorization）来实现对未知事实的判定。常用于链接预测（判断两个实体之间是否存在某种特定关系）、实体分类（判断实体所属语义类别）、实体解析（识别并合并指代同一实体的不同名称）。

常见的模型有RESCAL模型和TRESCAL模型。

RESCAL模型的核心思想，是将整个知识图谱编码为一个三维张量，由这个张量分解出一个核心张量和一个因子矩阵，核心张量中每个二维矩阵切片代表一种关系，因子矩阵中每一行代表一个实体。由核心张量和因子矩阵还原的结果被看作对应三元组成立的概率，如果概率大于某个阈值，则对应三元组正确；否则不正确。



$$\min_{\mathbf{A},\{\mathbf{R}_k\}} \sum_{k=1}^{K} \left\| \mathcal{X}_k - \mathbf{A}\mathbf{R}_k\mathbf{A}^T \right\|_F^2 + \lambda_1 \left\| \mathbf{A} \right\|_F^2 + \lambda_2 \sum_{k=1}^{K} \left\| \mathbf{R}_k \right\|_F^2$$

而TRESCAL则是解决在输入张量高度稀疏时所带来的过拟合问题。

路径排序算法也常用来判断两个实体之间可能存在的关系，比如PRA算法。本文不展开描述。

三、 大规模知识图谱的应用

知识图谱的应用场景非常广泛，比如搜索、问答、推荐系统、反欺诈、不一致性验证、异常分析、客户管理等。由于以上场景在应用中出现越来越多的深度学习模型，因此本文主要讨论知识图谱在深度学习模型中的应用。

目前将知识图谱用于深度学习主要有两种方式，一种是将知识图谱的语义信息输入到深度学习模型中，将离散化的知识表示为连续化的向量，从而使得知识图谱的先验知识能够称为深度学习的输入；另外一种是利用知识作为优化目标的约束，指导深度学习模型的学习过程，通常是将知识图谱中的知识表示为优化目标的后验正则项。

知识图谱的表示学习用于学习实体和关系的向量化表示，其关键是合理定义知识图谱中关于事实（三元组h,r,t）的损失函数fr(h,t)，其总和是三元组的两个实体h和t的向量化表示。通常情况下，当事实h,r,t成立时，期望最小化fr(h,t)。

$$\sum_{<h,r,t>\in O} f_r(\boldsymbol{h},\boldsymbol{t})$$

常见的有基于距离和翻译的模型。

基于距离的模型，比如SE模型，其基本思想是当两个实体属于同一个三元组时，它们的向量表示在投影后的空间中也应该彼此靠近。所以损失函数定义为向量投影后的距离

$$f_r(\boldsymbol{h},\boldsymbol{t}) - \| \boldsymbol{W}_{r,1}\boldsymbol{h} - \boldsymbol{W}_{r,2}\boldsymbol{t} \|_{l_1}$$

其中矩阵Wr1和Wr2用于三元组中头实体h和尾实体t的投影操作。

基于翻译的模型可以参考前述的TransE, TransH和TransR模型。其通过向量空间的向量翻译来描述实体与关系之间的相关性。

$$f_r(h,t) = \| h + r - t \|_{l_1/l_2}$$

当前的知识图谱表示学习方法都还存在各种问题，这个领域的发展也非常迅速，值得期待。

知识图谱的表示转换后，根据不同领域的应用，就可以和各种深度学习模型相结合，比如在自动问答领域，可以和encoder-decoder相结合，将问题和三元组进行匹配，即计算其向量相似度，从而为某个特定问题找到来自知识图谱的最佳三元组匹配。也有案例在推荐系统中，通过网络嵌入（network embedding）获取结构化知识的向量化表示，然后分别用SDAE（Stacked Denoising Auto-Encoder）和层叠卷积自编码器（Stacked Convolutional Auto-Encoder）来抽取文本知识特征和图片知识特征，并将三类特征融合进协同集成学习框架，利用三类知识特征的整合来实现个性化推荐。

随着深度学习的广泛应用，如何有效利用大量先验知识，来大大降低模型对大规模标注语料的依赖，也逐渐成为主要的研究方向之一。在深度学习模型中融合常识知识和领域知识，将是又一大机遇和挑战。