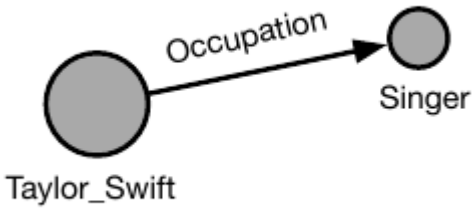
GAKE阅读记录及理解

1 背景知识

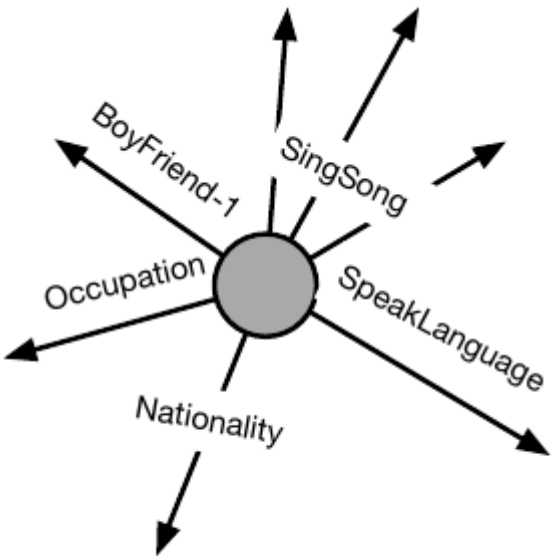
创新点：相比与大多数其它模型仅利用图的“三元组”信息这一信息，该论文提出的GAKE则利用KG的“图结构信息“，即利用图的上下文（**graph context**——Neighbor/Edge/Path context）信息进行训练，而大多数其它模型则未显示地利用到这些信息，因而训练出的结果，在发现这些信息的能力方面“可能”就比较薄弱（仅为个人推测）

Neighbor context: consists of **target entity** and its **directed linked entities** along with their **relations**



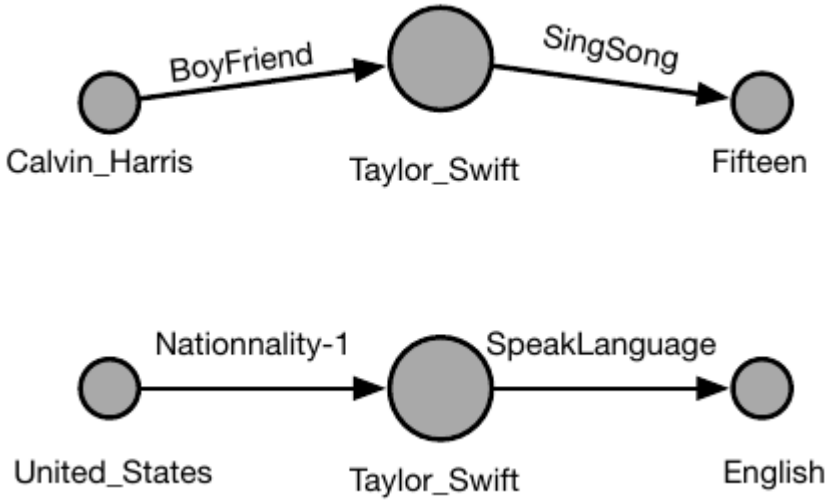
(b) Neighbor Context

Edge context: all kinds of **relations** relevant to the **target entity**



(c) Edge Context

Path context: all **relations** and **entities** in a path containing the **target entity**



2 模型框架

目标函数：

$$O = \lambda_N O_N + \lambda_P O_P + \lambda_E O_E \tag{1}$$

其中 λ_i 是 Hyperparameter，而 O_i 定义如下：

$$O_N = \sum_{s_i \in S} \sum_{c_N(s_i) \in C_N(s_i)} \log p(s_i | c_N(s_i)) \tag{2}$$

$$O_P = \sum_{s_i \in S} \sum_{c_P(s_i) \in C_P(s_i)} \log p(s_i | c_P(s_i)) \tag{3}$$

$$O_E = \sum_{s_i \in S} \log p(s_i | c_E(s_i)) \tag{4}$$

其中，$s = (t, k)$ 表示 KG 中的 subject(i.e.,a vertex or an edge),而 t 表示 subject 的类型（如 $t = 1$ 表示 s 为 vectex, $t = 0$ 则为 edge）,k 则是对应该 subject 的下标值，$S = \{s_i\}$；$c(s_i) = \{s_w | s_w \in S \bigwedge s_w \text{ relevant to } s_i\}$。$c_N(s_i) = (s_e, s_v)$、$c_P(s_i) =< s_{e1}, s_{v1}, s_{e2}, s_{v2}, \cdots, s_{eL}, s_{vL} >$ 、$c_E(s_i) = \{s_{e1}, s_{e2}, \cdots\}$依次表示 s_i 的一个 neighbor、一条 path、所有直连 edge。而函数 p 计算方式如下：

$$p(s_i | c(s_i)) = \frac{\exp\left(\phi(s_i)^T \pi c(s_i)\right)}{\sum_{j=1}^{|S|} \exp\left(\phi(s_j)^T \pi(c(s_j))\right)} \tag{5}$$

其中，$\phi(s_i) : s_i \in S \longmapsto R^{d \times 1}$，即该函数返回 s_i 的 embedding；$\pi(s_i) : s_i \in S \longmapsto R^{d \times 1}$ 是用于”编码“ s_i 的 graph context，其计算方式如下：

$$\pi(c(s_i)) = \alpha(s_i) \sum_{s_j \in c(s_i)} \phi(s_j) \tag{6}$$

若

$$\alpha(s_i) = \frac{1}{|c(s_i)|} \tag{7}$$

则仅是对邻居 embedding 相加求均运算；若使用 attention 机制，则应为：

$$\alpha(s_i) = \frac{\exp(\theta_i)}{\sum_{s_j \in C(s_i)} \exp(\theta_j)} \tag{8}$$

模型训练时，是以”最大化“目标函数 O 为目的的

3 模型理解

公式（5）的直观理解便是，在实体 s_i 的 graph context 出现的情况下，该实体出现的概率。因而公式（1）算的其实就是，在已知所有实体的 graph context 出现的前提下，它们各自对应的实体出现的总概率。显而意见，对于训练集而言，所有实体的 graph context 是已知的，且各自对应的实体也是必定会出现的，所以在这种情况下，公式（1）所算的总概率必然就是最大的（理想下应为1），这也是为何 GAKE 算法的目标函数是最大化公式（1）的原因。

在训练过程中，由公式（5）~（8）可知，该模型训练的参数其实就是各个实体和关系的 embedding，同时，如果采用 attention 机制的话，每个实体和关系还会一一具有一个参数 $\theta \in R$。所训练出参数是能够使得公式（1）最大化的。

相比于其它模型直接将三元组作为输入进行训练，在进行训练时，该模型的输入就是所有实体（可能会随机采样）以及其对应的 graph context 信息，这样做的一个好处便是可以利用 KG 的“图结果”信息（具体信息可看上文的“截图”处），使得训练数据“更有效化、更明确化”（此好处为个人理解，不能保证正确），训练的目的便是最大化公式（1），训练的结果便是各个实体和关系对应的 embedding。该模型提出的目的主要是解决 close world 的链接预测问题。在进行预测时，模型的输入便是预测实体对应的 embedding 以及其对应可能存在的 graph context，经过该模型运行后会得到一个概率，该概率表示在给定上下文的前提下，该实体能“拥有”该上下文的概率，若该概率大于一个给定的阈值，则能人为该实体与给定的上下文存在“链接关系”。

在阅读代码的过程中发现，它的实现还使用到了 word2vec 中 CBOW 模型，目前已经理解了 CBOW 模型，但对于如何使用到 GAKE 中还不是很清除，需要进一步阅读代码。

4 疑惑

目前我有一个疑惑，经过一个模型所学习到的 embedding，可以直接作为其它模型的“输入”吗？其它模型能发现这些 embedding 背后所蕴藏的特性信息吗？