第七讲：知识获取与知识计算

目录

7.1 知识获取

目录

7.1.1 实体抽取

实体抽取

- 实体抽取定义
- 研究方法分类
  - 基于规则与字典的方法
  - 基于机器学习的方法

实体抽取

- 实体抽取定义
  - 从原始语料中自动识别出指定类型的命名实体，主要包括实体名（如人名、地名、机构名、国家名等）、缩略词，以及一些数学表达式（如货币值、百分数、时间表达式等）

- 示例

  5月19日下午，史密斯教授做客北京大学海外名师讲堂。

  时间　　人名　　机构名

基于规则与字典的方法

- 基于规则与字典的方法
  - 根据人工构造出的规则和已知的命名实体库，将限定领域的语料进行相同的预处理并与规则集合中的规则进行匹配

- 优点
  - 准确率高

- 缺点
  - 需要大量的专家来编写规则和模板，成本大
  - 难以建立完整而准确的模式集合，不灵活
  - 实体抽取的召回率低

基于机器学习的方法

- 隐马尔科夫模型
- 条件随机场
- 基于深度学习的方法

基于机器学习的方法

- 序列标注
  - 实体标注一般使用BIO模式

(B-begin，I-inside，O-outside)

输入序列	小明	昨天	晚上	在	公园	遇到	了	小红	。
语块	B-NP	B-NP	I-NP	B-PP	B-NP	B-VP		B-NP	
标注序列	B-Agent	B-Time	I-Time	O	B-Location	B-Predicate	O	B-Patient	O
角色	Agent	Time	Time		Location	Predicate	O	Patient	

  - 还有BIOES标注模式

(B-begin，I-inside，O-outside，E-end，S-single)

基于机器学习的方法

- 隐马尔科夫模型
  - 假定分词后的文档词语序列为$W = (w_1, \dots, w_n)$，$T = (t_1, \dots, t_n)$为词序列的实体标注结果。模型旨在给定词语序列W的情况下，找出概率最大的标注序列T，即，求使$P(T|W)$最大的标注序列

$$T_{max} = arg_T maxP(T|W)$$

根据贝叶斯公式，

$$P(T|W) = P(T)P(W|T)/P(W)$$

其中，$P(W)$可以看成一个常数，则有

$$T_{max} = arg_T maxP(T)P(W|T)$$

其中，$P(T)P(W|T)$是引入隐马尔科夫模型来计算的参数。如果穷举序列W和T的所有可能情况，这个问题是NP难的

基于机器学习的方法

- 隐马尔科夫模型
  - 按照马尔科夫假设，当前状态t_i只和其前一状态t_{i-1}有关，因此有

$$P(T)P(W|T) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

其中，$P(w_i|t_i)$表示隐状态为t_i的词语集合中出现w_i的概率，$P(t_i|t_{i-1})$表示上一词语标注为t_{i-1}时，当前词语标注为t_i的转移概率。进一步

$$T_{max} = arg_T max \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

$$T_{max} = -arg_T min \sum_{i=0}^{n} \{lnP(w_i|t_i) + lnP(t_i|t_{i-1})\}$$

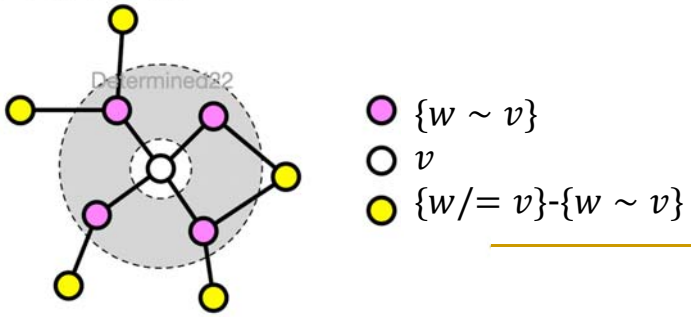
训练时，取$P(w_i|t_i) \approx Count(w_i, t_i)/Count(t_i)$，其中$Count(w_i, t_i)$表示词语$w_i$被标注为$t_i$的次数，$Count(t_i)$表示隐状态$t_i$出现的总次数

基于机器学习的方法

- 条件随机场(Conditional Random Field, CRF)
  - 设$X = (X_1, \dots, X_n)$与$Y = (X_1, \dots, X_n)$是联合随机变量。若在给定随机变量X的条件下，随机变量Y构成一个由无向图$G = (V, E)$表示的马尔科夫模型，则条件概率分布$P(Y|X)$称为条件随机场，即：

$$P(Y_v|X, Y_w, w/= v) = P(Y_v|X, Y_w, w \sim v)$$

其中，$w/= v$表示图$G = (V, E)$中节点v以外的所有节点，$w \sim v$表示与节点v有连边的所有节点



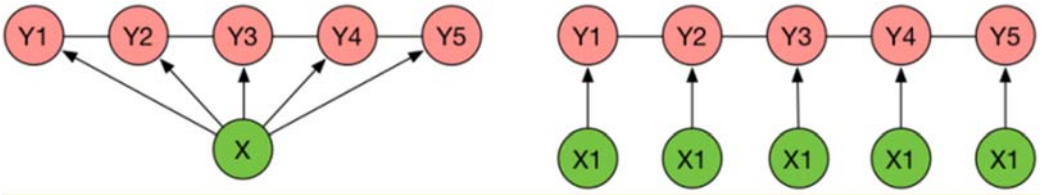
基于机器学习的方法

- 线性链条件随机场(Linear Chain CRF)
  - 设$X = (X_1, \dots, X_n)$与$Y = (Y_1, \dots, Y_n)$均为线性链表示的随机变量序列。若在给定的随机变量序列X的条件下，随机变量序列Y的条件概率分布$P(Y|X)$构成条件随机场，且满足马尔科夫性，即：

$$P(Y_i|X, Y_1, \dots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1})$$

则称$P(Y|X)$为线性链的条件随机场

  - 线性链CRF不仅考虑了上一个状态Y_{i-1}，还考虑了后续的状态Y_{i+1}



X= (X1,X2...)

X和Y具有相同的图结构

基于机器学习的方法

- 线性链条件随机场模型
  - 与隐马尔科夫模型相同，将CRF用于命名实体识别，其目标也是求$T_{max} = arg_T max P(T|W)$，但是这里

$$P(T|W) = \frac{1}{Z(W)} exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

$$Z(W) = \sum_t exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

其中$Z(W)$是归一化因子，在所有可能的输出序列上求和；λ_k和δ_l为权重因子

基于机器学习的方法

- 线性链条件随机场模型

$$P(T|W) = \frac{1}{Z(W)} exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

  - $\psi_k(t_{i-1}, t_i, W, i)$是转移函数，依赖于当前和前一位置，表示从标注序列中位置$i-1$的标记t_{i-1}转移到位置i上的标记为t_i的概率

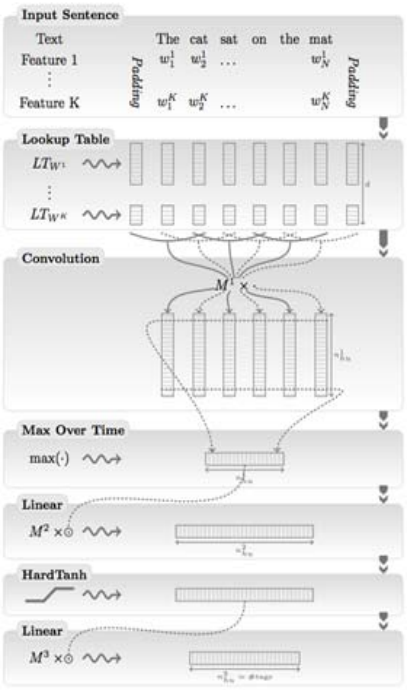
$$\varphi_k(t_{i-1}, t_i, W, i) = \begin{cases} 1, & \text{满足条件} \\ 0, & \text{其他} \end{cases}$$

  - $\phi_l(t_i, W, i)$是状态函数，表示标记序列在位置i上标记为t_i的概率

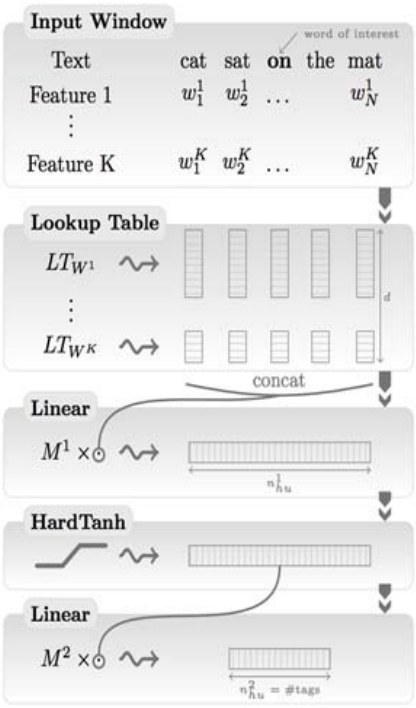
$$\phi_l(t_i, W, i) == \begin{cases} 1, & \text{满足条件} \\ 0, & \text{其他} \end{cases}$$

基于机器学习的方法

- 基于深度学习的模型



句子模型



窗口模型

基于机器学习的方法

- 基于深度学习的模型
  - 输入层：将窗口/句子中每一个词的特征按照词序拼接

$$f_\theta^1 = \langle LT_W([w]_1^T)\rangle_t^{d_{win}} = \begin{pmatrix} \langle W\rangle^1_{[w]_{t-d_{win}/2}} \\ \vdots \\ \langle W\rangle^1_{[w]_t} \\ \vdots \\ \langle W\rangle^1_{[w]_{t+d_{win}/2}} \end{pmatrix}$$

  - CNN层：通过CNN层变成大小一致的中间表示

$$\langle f_\theta^l\rangle_t^1 = W^l \langle f_\theta^{l-1}\rangle_t^{d_{win}} + b^l \quad \forall t$$

基于机器学习的方法

- 基于深度学习的模型
  - Max Pooling层：

$$\left[f_\theta^l\right]_i = \max_t \left[f_\theta^{l-1}\right]_{i,t} \qquad 1 \le i \le n_{hu}^{l-1}$$

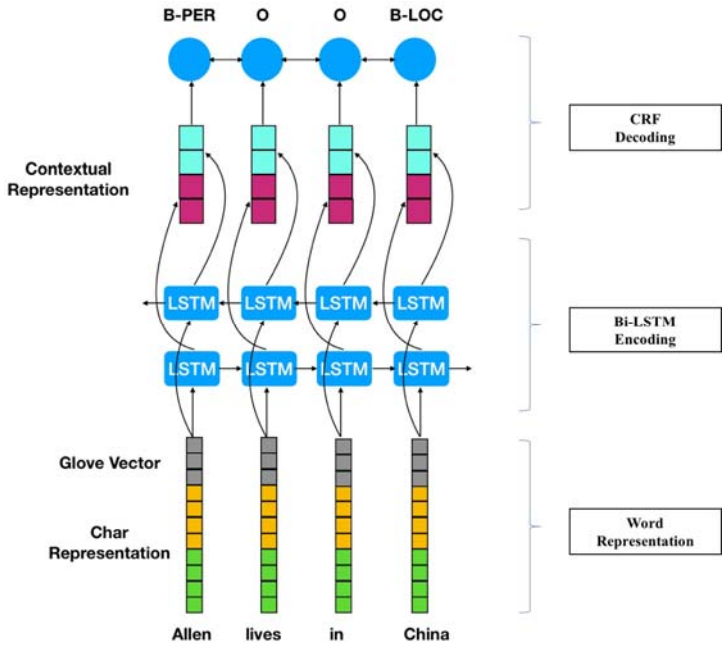
  - HardTanh层：输出各个tag上的打分

$$\left[f_\theta^l\right]_i = \text{HardTanh}(\left[f_\theta^{l-1}\right]_i),$$

$$\text{HardTanh}(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 <= x <= 1 \\ 1 & \text{if } x > 1 \end{cases}$$

基于机器学习的方法

- 基于深度学习的模型—LSTM-CRF模型



7.1.2 关系抽取

关系抽取

- 关系抽取示例

- 抽取方法分类
  - 有监督关系抽取
  - 半监督关系抽取
  - 远程监督关系抽取
  - 无监督关系抽取

有监督关系抽取

- 有监督的关系抽取通常被建模为分类问题，每个类对应一种预先定义的关系
- 在训练数据中，每个关系实例都会被打上一个关系标签，其中“无关系”被认为是一种特殊的关系标签
- 可分为三种
  - 基于特征的方法
  - 基于核的方法
  - 基于神经网络的方法

有监督关系抽取

- 基于特征的方法
  - 抽取实体对周围的词法、句法、语义特征，训练分类器并对关系未知的关系实例进行预测
  - 特征分类
    - 基于词的特征：两个实体以及它们之间的所有词、实体的中心词等
    - 基于短语的特征：两个实体是否处于同一个名词短语、动词短语、介词短语中，两个实体之间的短语中心词，两个实体之间的短语标签路径等
    - 基于语义的特征：国家名、人物相关的触发词、在WordNet的relative语义类中的词等

有监督关系抽取

- 基于核的方法
  - 设计一个核函数（Kernel Function）来计算两个关系实例的相似度，并通过支持向量机进行分类
  - 核函数：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$$

其中，$\phi()$为特征函数，核方法关键在于寻找合适的特征函数

  - 根据核函数的不同，基于核的方法大体可以分为：
    - 序列核
    - 树核：句法树核、依存树核
    - 图核：依存图核

有监督关系抽取

- 序列核方法 [Mooney and Bunescu, 2006]
  - 将句子视为词的序列，每个词都由一个特征向量表达，因而每个关系实例都可以视为一个特征向量的序列。特征包含词本身、词性标签、实体类型等
  - 特征函数定义为：$\phi_u(s) = \sum_{\mathbf{i}:u=s[\mathbf{i}]} \lambda^{l(\mathbf{i})}$

其中，$s = s_1 s_2 \dots s_{|s|}$是个序列；$\lambda \leq 1$是个对长度加权的衰减因子；$u = s[\mathbf{i}]$表示$s$的子序列，$\mathbf{i} = (i_1, i_2, \dots, i_{|u|})$是$s$的下标子序列（不一定连续，$1 \leq i_1 < \dots < i_{|u|} \leq |s|$）；$l(\mathbf{i}) = i_{|u|} - i_1 + 1$

  - 进而，定义两个长度为n的序列s, t的特征函数的积为核函数

$$K_n(s,t) = \sum_{u \in \Sigma^n} \phi_u(s) \cdot \phi_u(t) = \sum_{u \in \Sigma^n} \sum_{\mathbf{i}:u=s[\mathbf{i}]} \sum_{\mathbf{j}:u=t[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})}$$

Σ^n是长度为n的所有序列的集合

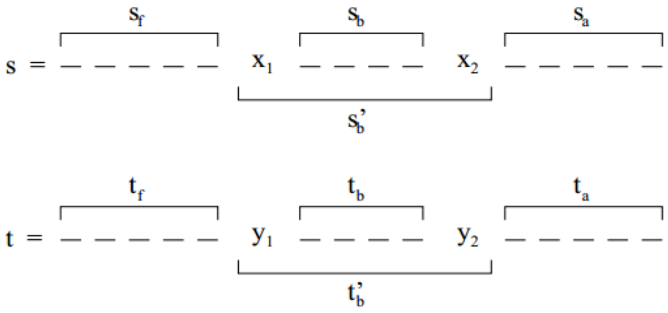
其中，u为s和t的共同子序列。该式表明了序列s和t的相似度

有监督关系抽取

- 序列核方法 [Mooney and Bunescu, 2006]
  - 若函数$c(x,y)$表示x和y两个词之间相似特征(如POS等)的个数，则核函数可以调整为：

$$K_n(s,t) = \sum_{\mathbf{i}:|\mathbf{i}|=n} \sum_{\mathbf{j}:|\mathbf{j}|=n} \prod_{k=1}^{n} c(s[\mathbf{i}_k], t[\mathbf{j}_k]) \lambda^{l(\mathbf{i})+l(\mathbf{j})}$$

  - 根据两个实体可将句子分为三段。在第一个实体前的为前段(f)，两实体之间的为中段(b)，第二个实体后的为后段(a)



有监督关系抽取

- 序列核方法 [Mooney and Bunescu, 2006]
  - 最终的关系核$rK(s,t)$由三个子核组成，分别是前-中段核$fbK(s,t)$、中段核$bK(s,t)$、中-后段核$baK(s,t)$

$$rK(s,t) = fbK(s,t) + bK(s,t) + baK(s,t)$$

$$fbK(s,t) = \sum_{1\le i,1\le j,i+j<fb_{max}} bK_i(s,t) \cdot K'_j(s_f, t_f)$$

$$bK(s,t) = \sum_{1\le i\le b_{max}} bK_i(s,t)$$

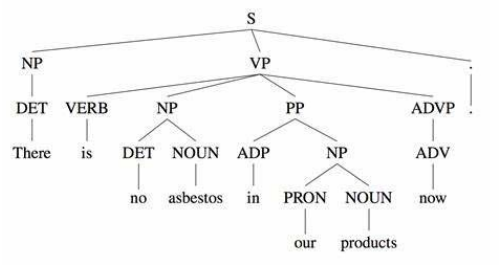
$$baK(s,t) = \sum_{1\le i,1\le j,i+j<ba_{max}} bK_i(s,t) \cdot K'_j(\bar{s}_a, \bar{t}_a)$$

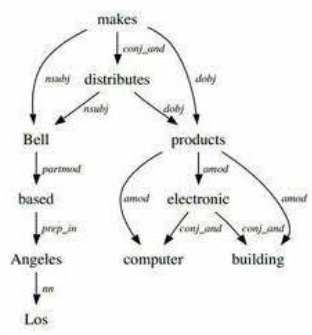
$$bK_i(s,t) = K_i(s_b, t_b) \cdot c(x_1, y_1) \cdot c(x_2, y_2) \cdot \lambda^{l(s'_b)+l(t'_b)}$$

其中$s'_b = x_1 s_b x_2$，$\bar{s}_a, \bar{t}_a$表示两序列的反转，$fb_{max}, b_{max}, fb_{max}$=4

有监督关系抽取

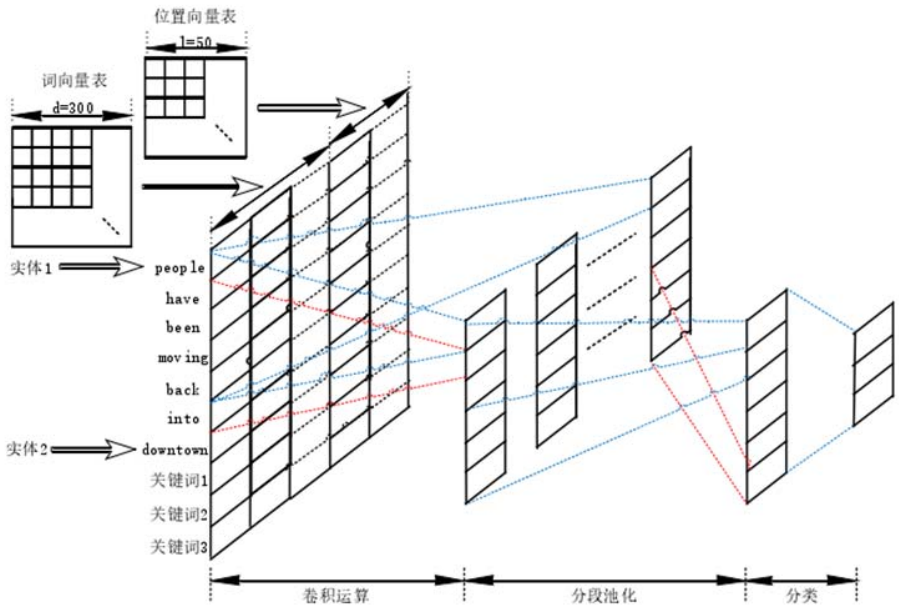
- 基于核的方法
  - 树核：句法树核、依存树核
  - 图核：依存图核





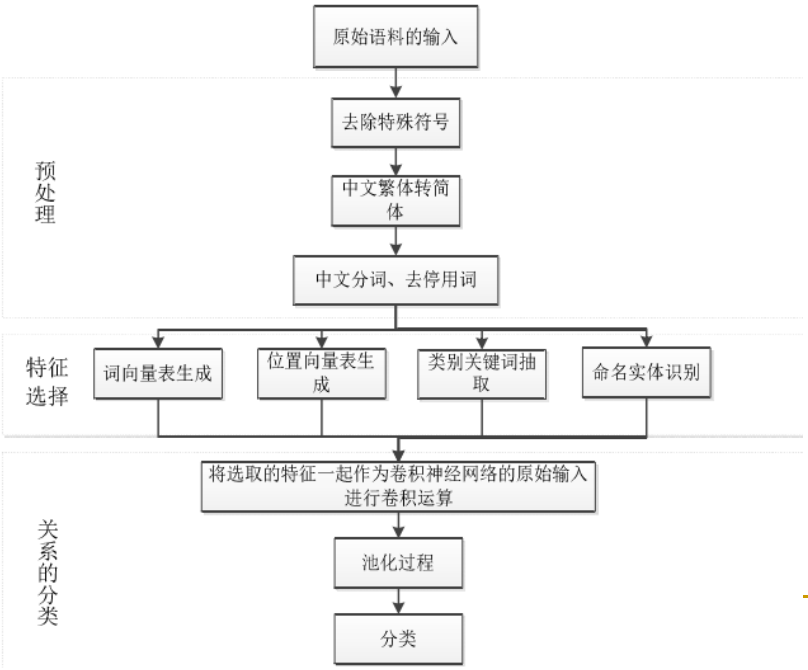
有监督关系抽取

- 基于神经网络的方法
  - 基于卷积神经网络



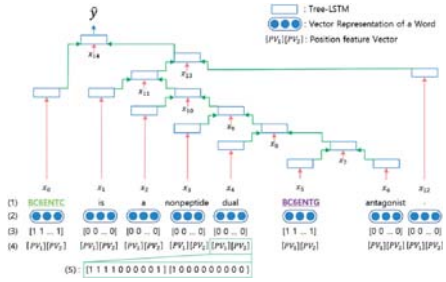
有监督关系抽取

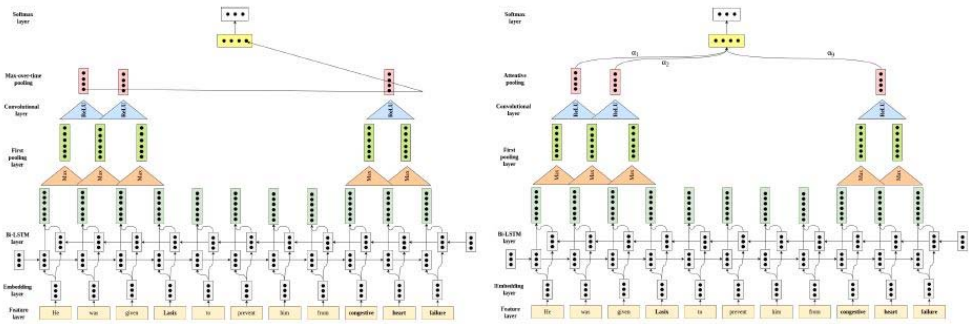
- 基于神经网络的方法
  - 基于卷积神经网络



有监督关系抽取

- 基于神经网络的方法
  - 基于循环神经网络
  - 基于递归神经网络





半监督关系抽取

- 半监督关系抽取
  - 主要动机：节省有监督关系抽取中人工标注的成本
  - 可以分为三类
    - 基于自举的方法
    - 基于标签传播的方法
    - 基于主动学习的方法

半监督关系抽取

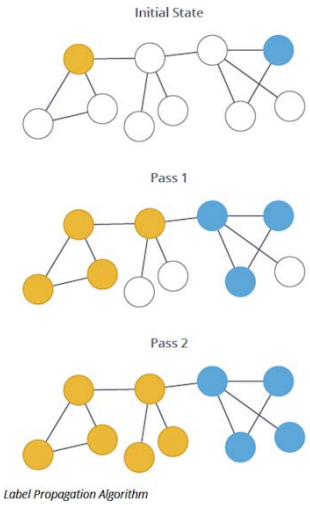
- 基于自举的方法
  - DIPRE（Dual Iterative Pattern Relation Extraction）[Brin, 1998]
  - 核心假设：模式-关系对偶性，即：通过高质量的模式可以找到高质量的关系实例，通过高质量的关系实例可以学习高质量的模式
  - 从少量人工挑选的种子实例出发，用学习到的模式匹配大量无标注数据，将匹配上的数据放入关系实例集合，然后重新学习模式，迭代地扩展关系实例集合，直到获取到足够规模的关系实例
  - 示例：
    - 关系实例：“作者-书名”
    - 关系模式：$(order, urlprefix, prefix, middle, suffix)$
    - 若一个URL能够匹配正则表达式$urlprefix*$，且该URL对应的网页中存在一段文本可以匹配正则表达式$*\ prefix, author, middle, title, suffix\ *$，则称 $(author, title)$匹配上该模式

半监督关系抽取

- 基于标签传播的方法
  - 基本思想
    - 有标签和无标签的关系实例都被视为图中的节点并表达为一个特征向量，节点间的相似度作为边的权值
      - 关系实例R_i和R_j之间的相似度（边权）为

        $$W_{ij} = exp\left(\frac{s_{ij}}{\sigma^2}\right)$$

        其中，s_{ij}表示两个实例对应特征向量之间的相似度；σ^2用以缩放权重，具体设为有标签的实例之间的相似度的平均值
    - 每个有标签的节点其标签信息会迭代地通过带权的边传播到相邻的无标签节点，直到最终传播过程收敛



半监督关系抽取

- 基于主动学习的方法
  - 核心思想：允许学习算法询问一些无标签数据的真实标签

远程监督关系抽取

- 初始动机：通过外部知识库代替人对语料进行标注，从而低成本地获取大量有标注数据 [Mintz et al., 2009]
- 核心思想：如果知识库中存在三元组$\langle e_1, R, e_2 \rangle$，那么语料中所有出现实体对$\langle e_1, e_2 \rangle$的语句，都标注为表达了关系R
- 根据这一假设，对每个三元组$\langle e_1, R, e_2 \rangle$，将所有$\langle e_1, e_2 \rangle$共现的句子都标注标签$R$，用分类方法解决关系抽取问题

远程监督关系抽取

- Riedel等[Riedel et al., 2010]认为Mintz的假设过强，可能引入噪声模式，因而提出“at-least-once”假设：
  - 如果存在三元组$\langle e_1, R, e_2 \rangle$，那么所有$\langle e_1, e_2 \rangle$实体对共现的语句中，至少有一句体现了关系R在这两个实体上成立的事实
- 引入了多实例学习机制，将所有$\langle e_1, e_2 \rangle$共现的句子聚成一个句袋，并将任务由对句子分类变为对句袋分类

无监督关系抽取

- 无监督关系抽取方法将实体对之间的字符串抽取出来，通过聚类和化简得到其关系标签
  - 不依赖标注数据
  - 标签（关系类型）未归一化
- 基于层次聚类的关系抽取方法[Hasegawa et al., 2004]
  - 标注命名实体，通过实体对得到关系实例
  - 关系实例由TF-IDF表示，以余弦相似度衡量两关系实例的相似度
  - 两个簇的相似度为两簇中实例的最大距离（最小相似度）
  - 用频繁出现的共有词作为簇的标签（关系类型）

7.2 知识融合

目录

7.2.1 实体对齐

实体对齐

- 实体对齐的定义
  - 实体对齐也称为实体匹配或实体解析，指的是判断相同或不同的知识库中的两个实体是否指向同一个对象的过程
    - 如：“诗仙”和“李太白” 两个指称词都应指向同一个实体“李白”
- 实体对齐方法分类
  - 成对实体对齐
  - 集体实体对齐

基于属性相似度的成对实体对齐

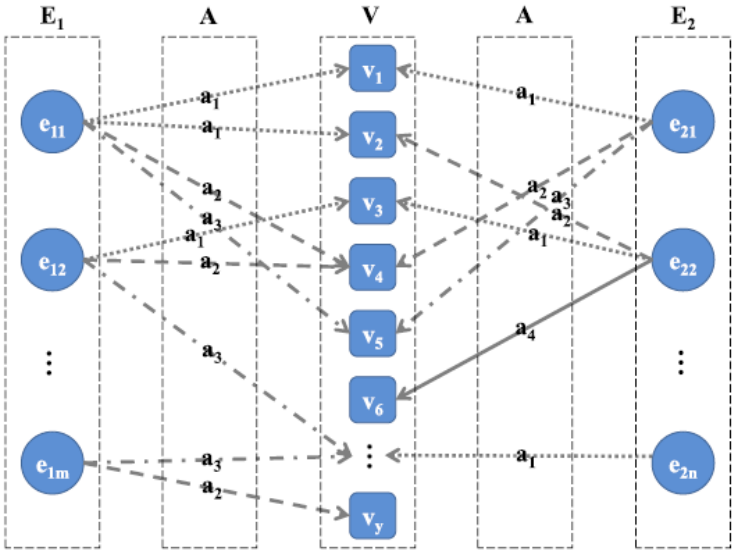
- 成对实体对齐
  - 将一个实体对基于属性相似性进行成对比较

$$\begin{cases} sim_{Attr}^{sum}(e_1, e_2) \geq t_1 \Rightarrow e_1, e_2\text{匹配} \\ t_2 \leq sim_{Attr}^{sum}(e_1, e_2) < t_1 \Rightarrow e_1, e_2\text{可能匹配} \\ sim_{Attr}^{sum}(e_1, e_2) < t_2 \Rightarrow e_1, e_2\text{不匹配} \end{cases}$$

其中，e_1, e_2为待匹配的实体

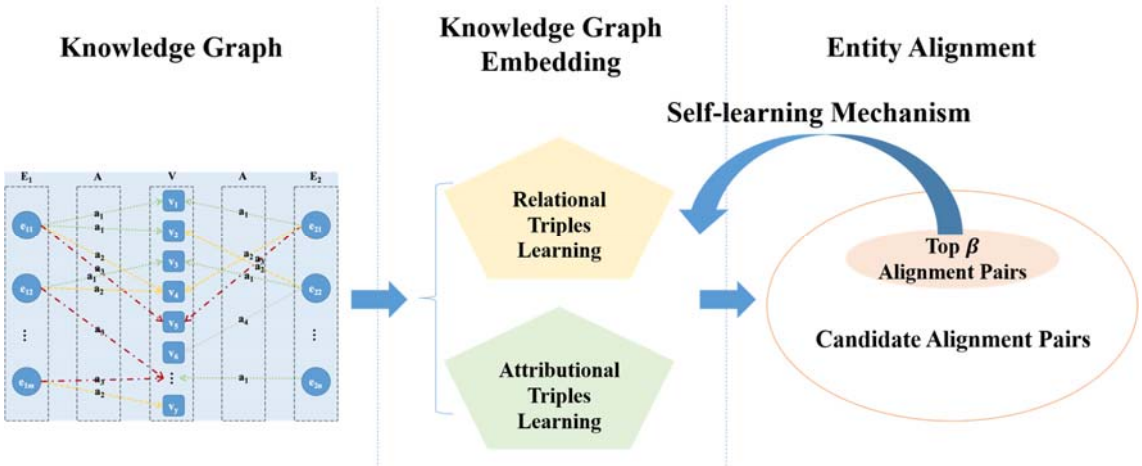
基于神经网络的成对实体对齐

- SEEA (Self-learning and Embedding based method for Entity Alignment)

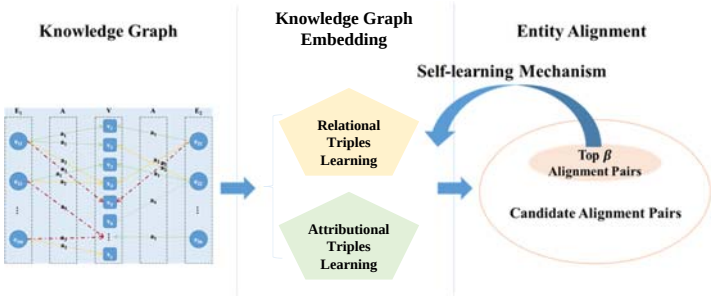


SEEA模型框架

- 两个主要模块
  - 知识图谱表示学习和实体对齐
- 一个机制
  - 自学习机制

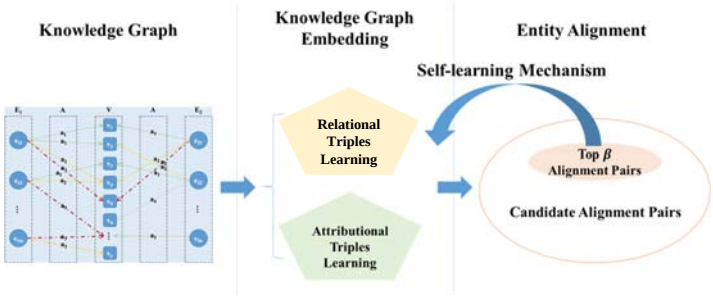


知识图谱表示学习



- 区分关系和属性
  - 开始时没有关系三元组，后续通过学习逐步加入
- 学习时给予重要的属性三元组更高的权重
- 既是实体又是属性值的元素只用一个向量表示，而不像现有工作用两个向量表示

知识图谱表示学习



- 关系三元组
  - 优化概率$P(h|r,t,\mathbf{X})$和$P(t|h,r,\mathbf{X})$
  - 忽略$P(r|h,t,\mathbf{X})$，由于关系r为固定的对齐关系
  - 通过负采样学习

$$P(h|r,t,\mathbf{X}) = \prod_{(h,r,t)\in RT}\left[\sigma(g_r(h,t))\prod_{i=1}^{c_{1h}}\mathbb{E}_{(h_i,r,t)\sim P(RT_h^-)}\sigma(g_r(h_i,t))\right]$$

头实体对应的负采样数

从RT_h^-中随机采样实例

头实体对应的负例集

sigmoid函数

得分函数

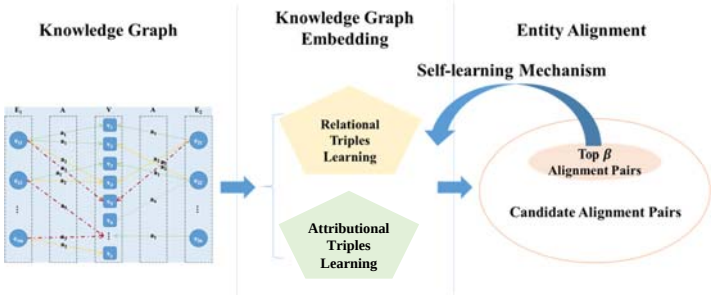
$$g_r(h,t) = -\|\mathbf{h}-\mathbf{t}\|_{L_1}$$

负样本生成方法

- 负样本生成策略
  - 1. 在实体集合中随机选择实体h' (t')，替换(h,r,t)中的$h(t)$ ，生成负样本(h',r,t)或者(h,r,t')
  - 2. 在选择替换实体的时候，不是完全随机在实体集合中选择，而是在适合关系r的实体集合中随机选取。
  - 例如：进行尾部实体替换时，只是用其他的地名替换“上海”，如“成都”，而不会使用人名进行替换：

（姚明，出生于，上海）

知识图谱表示学习



- 属性三元组
  - 优化$P(v|h,a,\mathbf{X})$，实体和属性之间的关联用一个分类模型捕捉
  - 用负采样信息

$$P(v|h,a,\mathbf{X}) = \prod_{(h,a,v)\in AT}\left[\sigma(\varphi_a(h,v))\prod_{i=1}^{c_2}\mathbb{E}_{(h,a,v_i)\sim P(AT^-)}\sigma(\varphi_a(h,v_i))\right]$$

属性值对应的负采样数

得分函数 $\varphi_a(h,v) = \|f(\mathbf{h}\mathbf{W}_\mathbf{a}+\mathbf{a})-\mathbf{v}\|_{L_1}+b$

知识图谱表示学习

- 损失函数
  - 有效三元组的概率为1，负例的概率为0

$$L'(\mathbf{X}) = \prod_{(h,r,t)\in RT}\left[|1-\sigma(g_r(h,t))|\prod_{i=1}^{c_1}\mathbb{E}_{(h_i,r,t_i)\sim P(RT^-)}|0-\sigma(g_r(h_i,t_i))|\right]$$
$$\prod_{(h,a,v)\in AT}\left[|1-\sigma(\varphi_a(h,v))|\prod_{i=1}^{c_2}\mathbb{E}_{(h,a,v_i)\sim P(AT^-)}|0-\sigma(\varphi_a(h,v_i))|\right]$$

c_1：头尾实体对应的负例数

知识图谱表示学习

- 最终损失函数
  - 取对数并加上正则项

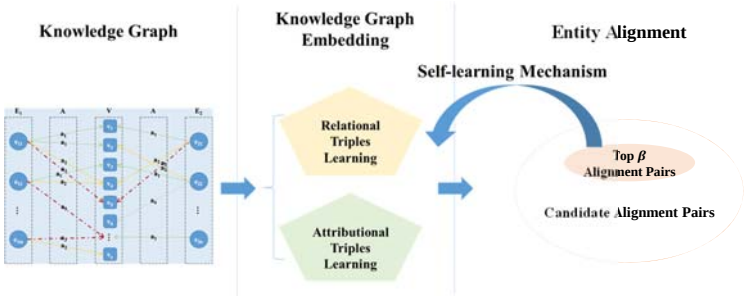
$$L(\mathbf{X}) = \sum_{(h,r,t)\in RT}\left[|1-\sigma(g_r(h,t))|+\sum_{i=1}^{c_1}\mathbb{E}_{(h_i,r,t_i)\sim P(RT^-)}|0-\sigma(g_r(h_i,t_i))|\right]$$
$$+\sum_{(h,a,v)\in AT}\left[|1-\sigma(\varphi_a(h,v))|+\sum_{i=1}^{c_2}\mathbb{E}_{(h,a,v_i)\sim P(AT^-)}|0-\sigma(\varphi_a(h,v_i))|\right]$$
$$+\lambda C(\mathbf{X})$$

  - 用L_2范数定义正则项

$$C(\mathbf{X}) = \sum_{h\in E}\left[|\mathbf{h}|_{L_2}-1\right]_+ + \sum_{v\in V}\left[|\mathbf{v}|_{L_2}-1\right]_+ + \sum_{h\in E}\sum_{a\in A}\left[|f(\mathbf{hW_a}+\mathbf{A_a})|_{L_2}-1\right]_+$$

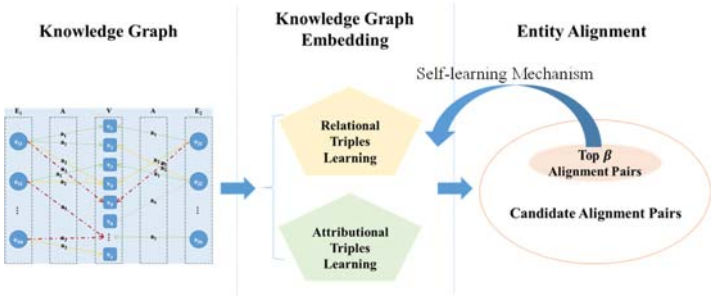
其中，$[z]_+ = \max(0,z)$

实体对齐



- 衡量指标
  - 通过所学到的实体向量的cosine相似度确定对齐的实体对
- 选与头实体A最相似的尾实体B构成(A, B)对齐对

自学习机制



- 反馈
  - 选择cosine相似度前β的对齐实体对构成关系三元组再训练模型
- 迭代直至收敛
  - 知识图谱表示学习 → 实体对齐
  - 实体对齐 → 知识图谱表示学习

集体实体对齐

- 集体实体对齐
  - 基于相似性传播的集体实体对齐
    - 利用初始的匹配，以“bootstraping”方式迭代地产生新的匹配。在此过程中，实体间的相似度会随着算法的迭代不断变化，直到算法收敛或达到指定的停止条件

7.2.2 实体链接

实体链接

- 实体链接的定义
  - 实体链接指的是利用知识库中的实体对知识抽取阶段所获得的实体指称词进行消歧的过程
- NIL实体
  - 如果实体指称在知识库中找不到对应的实体，则称其为“NIL实体”，实体链接还需要对NIL实体进行预测
- 实体链接为每一个实体指称词在知识库中找到对应的映射或者给出NIL实体的标签

实体链接

- 实体链接系统
  - 候选实体生成
  - 候选实体排序
  - NIL实体预测

候选实体生成

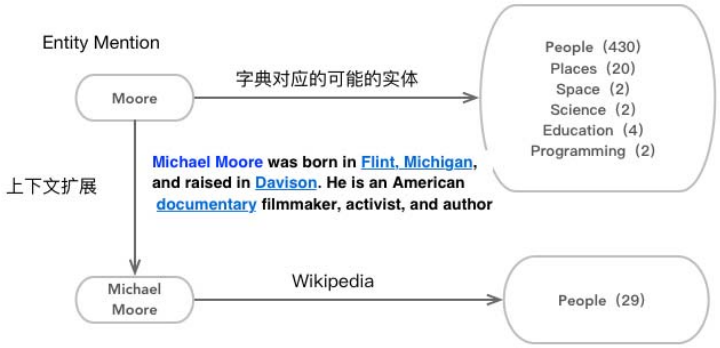
- 候选实体生成
  - 为了降低实体链接的复杂度，对于每一个实体指称词，实体链接系统首先需要确定一个实体指称词有可能指向的实体集合。这一阶段可以过滤掉大部分实体指称项不可能指向的实体，仅仅保留少量的候选实体
- 主要方法
  - 基于字典的方法
  - 基于文本扩展的方法
  - 基于搜索的方法

候选实体生成

- 基于字典的方法
  - 通过一个以实体为键，其所有可能的指称词为值的字典来发现实体指称词对应的实体
- 基于文本扩展的方法
  - 在实体指称词出现的上下文中，通过启发式或指代消解的方法对实体进行扩展，然后再利用其他方法对扩展后的指称词进行候选实体生成

候选实体生成

- 基于搜索的方法
  - 利用在知识库上构建的索引去获得候选实体，也可以直接使用搜索引擎去获得候选实体，构建索引可以使用Lucene和Elastic Search



候选实体排序

- 候选实体排序
  - 将实体指称词和生成的候选实体按照匹配度进行排序，然后将匹配度最高的实体作为实体链接的结果

- 方法分类
  - 基于相似度计算的方法
  - 基于机器学习的方法
  - 基于图的方法

候选实体排序

- 基于相似度计算
  - 利用编辑距离、Jaccard相似度、Dice系数，TF-IDF相似度等，计算实体指称词和知识库中实体的相似度来完成实体链接
  - Wikipedia Link-based Measure

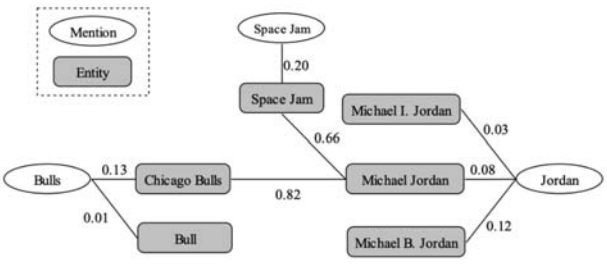
$$WLM = \frac{log max(|A|,|B|) - log|A \cap B|}{log|W| - log min(|A|,|B|)}$$

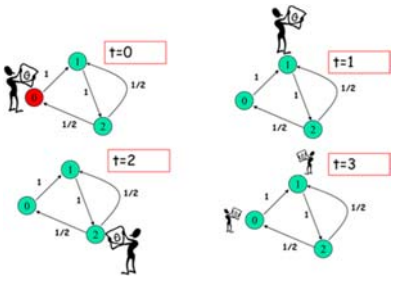
A 和B 是链向实体a和实体b的文章集合，W 是Wikipedia中所有文章的集合。两个实体a和b的相似度可以用定义如下：

$$Sim(a,b) = 1 - WLM$$

候选实体排序

- 基于图的方法
  - 构造一个节点为所有实体指称词与其候选实体，边表示实体与实体之间的相关性的图
    - 每一个实体指称词与对应的候选实体有一条边相连，边的权重通过余弦相似度计算
    - 在不同的候选实体集合之间也通过一条边相连，通过WLM计算得到边上的权重
  - 候选实体的得分通过随机游走模型得到





候选实体排序

- 基于机器学习的方法
  - 在实体对上提取特征构成特征向量，转化为一般机器学习中的二值分类问题

NIL实体预测

- NIL实体聚类
  - 指称项特征表示
    - 词袋模型
    - 语义特征
    - 社会化网络
    - 维基百科的知识
    - 多元异构语义知识融合
  - 聚类
    - 层次聚类
    - KNN
    - K-means

7.3 知识推理与计算

目录

7.3.1 知识推理简介

什么是知识推理？

- 人类视角
  - 人们从已知的事实出发，通过运用已掌握的知识，找出其中蕴含的事实或归纳出新的事实的过程
  - 按照某种策略由已知判断推出新的判断的思维过程
  - 基于特定的规则和约束，从存在的知识获得新的知识
- 计算机视角
  - 在计算机或智能系统中，模拟人类的智能推理方式，依据一定的推理控制策略，利用形式化的知识进行机器思维和求解问题的过程

利用已知的知识推出新知识的过程

知识推理分类

- 按新知识推出的途径分为演绎推理、归纳推理和默认推理
  - 演绎推理：从一般到个别的推理；发展历史悠久（1935年Gentzen提出自然演绎推理），涵盖自然演绎、归结原理、表演算等广泛使用的方法
    - 例子：（假言三段论）
    - 大前提：任何三角形只可能是锐角三角形、直角三角形和钝角三角形
    - 小前提：这个三角形既不是锐角三角形，也不是钝角三角形
    - 结　　论：它是一个直角三角形
  - 归纳推理：从足够多的事例归纳出一般性结论的推理过程；可以追溯到1964年Solomonof创立的普遍归纳推理理论，是一个基于观察的预测理论
    - 例子：所有观察到的乌鸦都是黑的，所以所有乌鸦都是黑的
  - 默认推理又称为缺省推理，是在知识不完全的情况下，通过假设某些条件已经具备而进行的推理；1980年Reiter正式提出缺省推理逻辑
    - 例子：如果没有足够的证据能证明条件A不成立，则默认A是成立的，并在此默认的前提下进行推理，推导出某个结论

知识推理分类

- 按所用知识的确定性分为确定性推理和不确定性推理
  - 确定性推理所用的知识是精确的并且推出结论也是确定的
  - 在不确定性推理中，知识和证据都具有某种程度的不确定性。不确定性推理又分为似然推理和近似推理(模糊推理)，前者基于概率论(如贝叶斯推理)，后者基于模糊逻辑（1973年由Zadeh提出）
- 按推理过程中推出的结论是否单调增加分为单调推理和非单调推理
  - 单调推理中，随着推理的推进和新知识的加入，推出的结论单调递增，越来越接近最终目标
  - 非单调推理由Minsky提出；在推理过程中，随着新知识的加入，可能需要否定已推出的结论，使得推理退回到前面的某一步甚至是重新开始
- 按是否运用与问题有关的启发性知识分为启发式推理和非启发式推理
  - 启发式推理在推理过程中运用解决问题的策略、技巧和经验，加快推理
  - 非启发式推理只按照一般的控制逻辑进行推理

知识推理分类

- 从方法论的角度划分为基于知识的推理、统计推理和直觉推理
  - 基于知识的推理根据已掌握的事实，通过运用知识进行推理
  - 统计推理根据对事物的统计信息进行推理
  - 直觉推理又称为常识推理，是根据常识进行的推理
- 按推理的繁简不同分为简单推理和复合推理
- 根据结论是否具有必然性分为必然性推理和或然性推理
- 根据推理控制方向划分为正向推理、逆向推理、混合推理和双向推理
- 此外还有时间推理、空间推理和案例推理等推理方法
  - 时间推理是对与时间有关的知识进行的推理
  - 空间推理是对空间对象之间的空间关系进行定性/定量分析和处理的过程
  - 案例推理通过使用或调整老问题的的解决方案推理新问题的解决方案

知识表示与推理

- 自然语言
- 基于符号的表示方法
  - Predicate Logic（谓词逻辑）
  - Semantic Net
  - Frame（框架）
  - Script（脚本）
  - Semantic Web（语义网）

知识推理

- 基于分布式表达的表示方法
  - 张量分解
  - 基于翻译的模型
  - 神经网络模型

知识计算(推理)

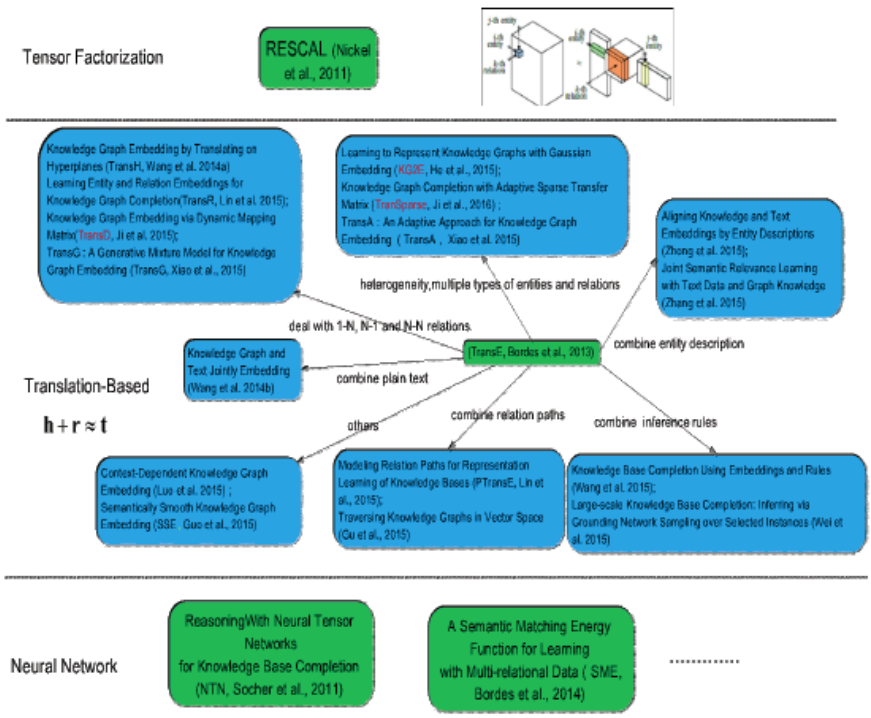
7.3.2 基于分布式表达的知识计算

基于分布式表达的知识计算

张量分解方法

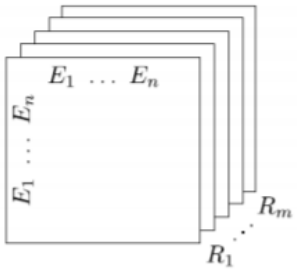
基于翻译的方法

神经网络方法



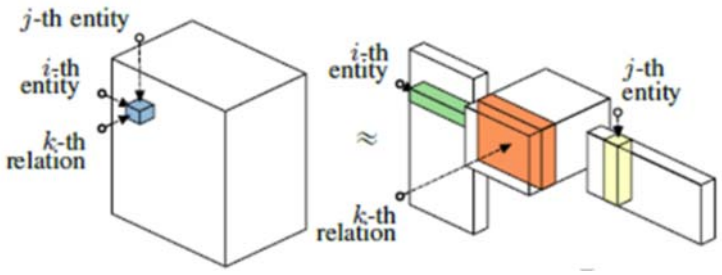
用张量表示知识图谱

知识图谱中三元组的结构是（头部实体h，关系r，尾部实体t），其中r连接头尾实体。以E_1，E_2，⋯ E_n表示知识图谱中的实体，以R_1，R_2，⋯ R_N表示知识图谱中的关系，则可以使用一个三维矩阵（张量）表示知识图谱



Nickel et al. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11).

张量分解得到实体、关系表示

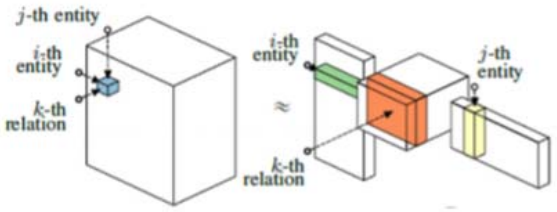


张量分解的目标函数

- 表示知识图谱的张量记为$\mathcal{Y}$，其第k个矩阵记为Y_k，是第k种关系的矩阵，表示该种关系在向量空间中与头尾部实体相互作用
- 对Y_k可以进行如下的低秩分解：

$$\mathrm{Y_k} = \mathrm{AR_kA^T} \qquad k = 1,2,\cdots m$$

其中，$Y_k \in R^{n*n}$，$A \in R^{n*r}$，$R_k \in R^{r*r}$，r表示矩阵A的秩；A是实体向量矩阵，每一行表示一个实体的向量，转置后其每一列表示一个实体的向量



张量分解的目标函数

- 由上述内容可知，A 和R_k 均是待求解的变量。因此目标函数是：

$$\min_{A,R_k}(f(\boldsymbol{A},\boldsymbol{R}_k)+g(\boldsymbol{A},\boldsymbol{R}_k))$$

其中 $f(\boldsymbol{A},\boldsymbol{R}_k)$ 是目标函数

$$f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\left(\sum_k\left\|\boldsymbol{Y}_k-\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^{\mathrm{T}}\right\|_F^2\right)$$

$g(\boldsymbol{A},\boldsymbol{R}_k)$ 是正则化项

$$g(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\gamma\left(\|\boldsymbol{A}\|_F^2+\sum_k\|\boldsymbol{R}_k\|_F^2\right)$$

张量分解的目标函数

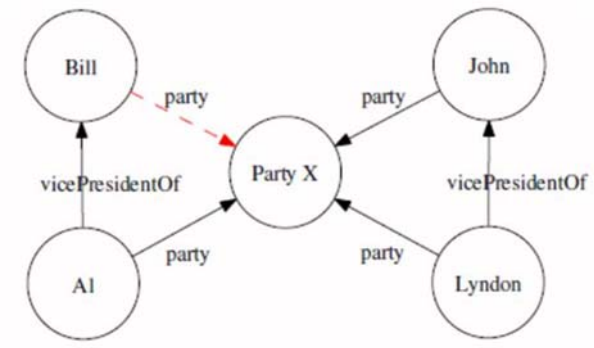
- 将目标函数写成分量形式

$$f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\left(\sum_k\left\|\boldsymbol{Y}_k-\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^{\mathrm{T}}\right\|_F^2\right)\Rightarrow f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\sum_{i,j,k}\left(y_{ijk}-\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j\right)^2$$

其中，y_{ijk}是张量中的一个元素，$\boldsymbol{a}_i$表示$\boldsymbol{A}$的第i行，即

$$[\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_n]=\boldsymbol{A}$$

张量分解模型的解释



$$\boldsymbol{a}_{A1}^T\boldsymbol{R}_{party}\boldsymbol{a}_{partyX}\approx\boldsymbol{a}_{Lyndon}^T\boldsymbol{R}_{party}\boldsymbol{a}_{partyX}\ \Rightarrow\ \boldsymbol{a}_{A1}^T\approx\boldsymbol{a}_{Lyndon}^T$$

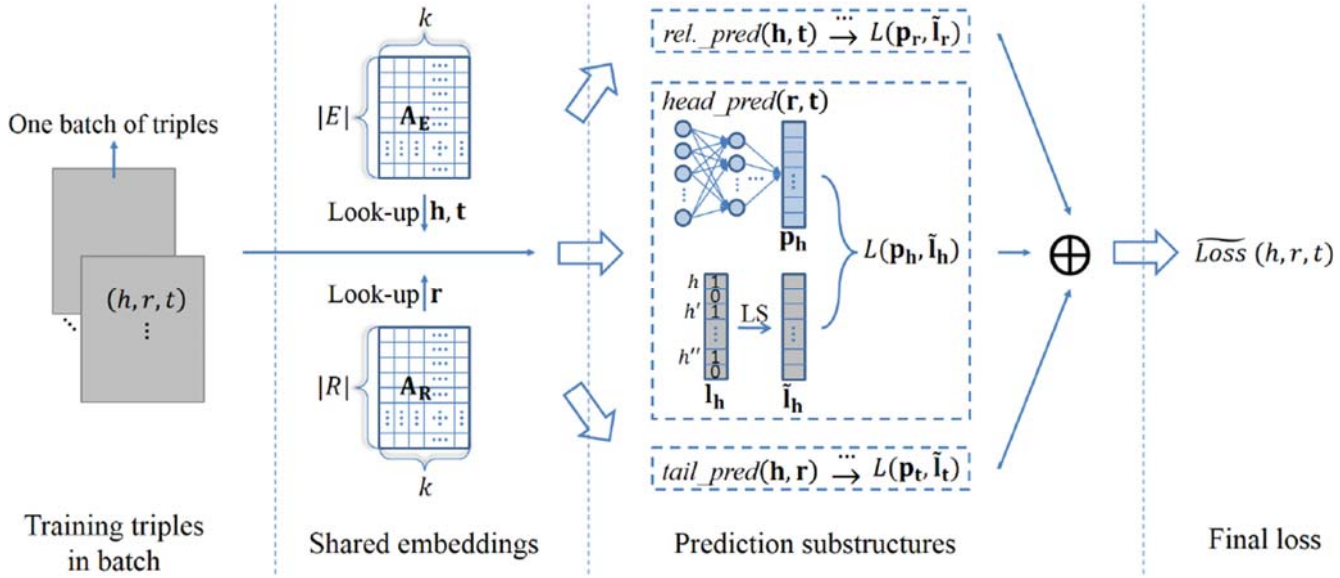
$$\boldsymbol{a}_{A1}^T\boldsymbol{R}_{vicePresidentOf}\boldsymbol{a}_{Bill}\approx\boldsymbol{a}_{Lyndon}^T\boldsymbol{R}_{vicePresidentOf}\boldsymbol{a}_{John}\Rightarrow\boldsymbol{a}_{Bill}\approx\boldsymbol{a}_{John}$$

基于神经网络的知识计算: SENN的动机

- 现有关系推断方法没有区分建模具体任务$(?,r,t)$、$(h,?,t)$和$(h,r,?)$
- 实际上，这些任务的推断性能差异显著，关系预测的效果远好于实体预测
- SENN (Shared Embedding based Neural Network)
  - 分开建模关系推断的三个任务
  - 通过共享向量表示，在统一的框架中集成这三个任务

SENN的框架

- 2个共享表示矩阵
- 3个子结构：*head_pred*、*rel._pred*和*tail_pred*

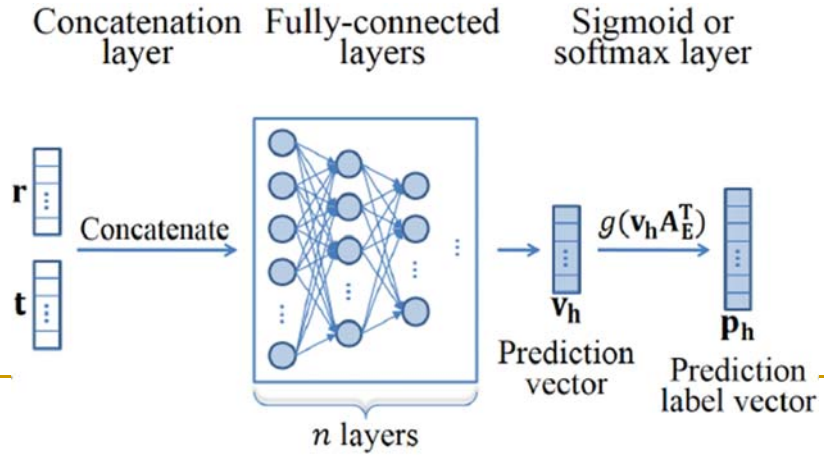


头实体预测子结构

- 得分函数

$$s(\mathbf{r},\mathbf{t}) = \mathbf{v_h}\mathbf{A_E^T}$$
$$= f\left(f\left(\cdots f\left([\mathbf{r};\mathbf{t}]\mathbf{W_{h,1}} + \mathbf{b_{h,1}}\right)\cdots\right)\mathbf{W_{h,n}} + \mathbf{b_{h,n}}\right)\mathbf{A_E^T}$$

其中f是ReLU函数

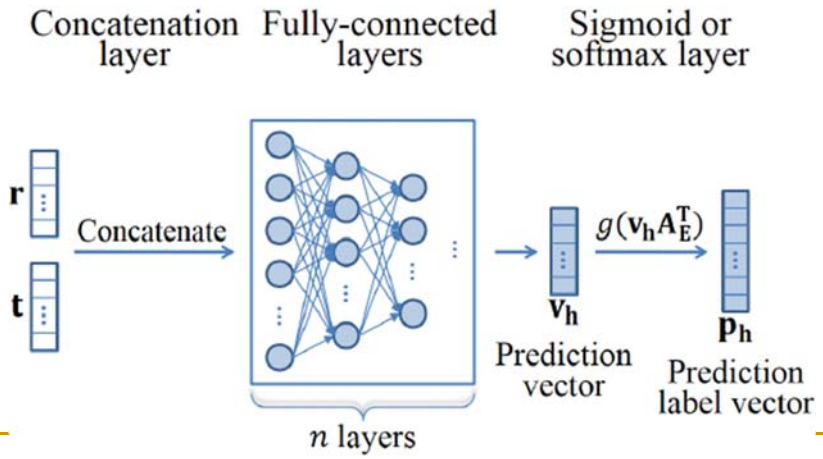


头实体预测子结构

- 预测标签向量 $\mathbf{p_h}$

$$\mathbf{p_h} = g(s(\mathbf{r},\mathbf{t}))$$

其中g 是 sigmoid 或 softmax函数



关系与尾实体预测子结构

- 关系预测子结构*rel._pred*和尾实体预测子结构*tail_pred*类似
  - 得分函数

$$s(\mathbf{h},\mathbf{t}) = f\left(f\left(\cdots f\left([\mathbf{h};\mathbf{t}]\mathbf{W}_{\mathbf{r},1}+\mathbf{b}_{\mathbf{r},1}\right)\cdots\right)\mathbf{W}_{\mathbf{r},\mathbf{n}}+\mathbf{b}_{\mathbf{r},\mathbf{n}}\right)\mathbf{A}_{\mathbf{R}}^{\mathrm{T}}$$

$$s(\mathbf{h},\mathbf{r}) = f\left(f\left(\cdots f\left([\mathbf{h};\mathbf{r}]\mathbf{W}_{\mathbf{t},1}+\mathbf{b}_{\mathbf{t},1}\right)\cdots\right)\mathbf{W}_{\mathbf{t},\mathbf{n}}+\mathbf{b}_{\mathbf{t},\mathbf{n}}\right)\mathbf{A}_{\mathbf{E}}^{\mathrm{T}}$$

  - 预测标签向量

$$\mathbf{p}_{\mathbf{r}} = g(s(\mathbf{h},\mathbf{t}))$$

$$\mathbf{p}_{\mathbf{t}} = g(s(\mathbf{h},\mathbf{r}))$$

SENN模型训练

- 标准损失函数
  - (h,r,t)的这些预测任务有它们的目标标签向量$\mathbf{l}_{\mathbf{h}}$、$\mathbf{l}_{\mathbf{r}}$ 和$\mathbf{l}_{\mathbf{t}}$

$$[\mathbf{l}_{\mathbf{h}}]_i = \begin{cases} 1, & e_i \in I_h \\ 0, & otherwise \end{cases}$$

    其中I_h 是给定r和 t训练集中的有效头实体集
  - 用标签平滑正则化目标标签向量

$$\tilde{\mathbf{l}}_{\mathbf{h}} = (1-\alpha_h)\mathbf{l}_{\mathbf{h}} + \frac{\alpha_h}{|E|}$$

    其中α_h是标签平滑超参

SENN模型训练

- 标准损失函数
  - 三个预测任务的二值交叉熵损失

$$L\left(\mathbf{p}_{\mathbf{h}},\tilde{\mathbf{l}}_{\mathbf{h}}\right)=-\frac{1}{|E|}\sum_{i=1}^{|E|}\left(\left[\tilde{\mathbf{l}}_{\mathbf{h}}\right]_i\log\left[\mathbf{p}_{\mathbf{h}}\right]_i+\left(1-\left[\tilde{\mathbf{l}}_{\mathbf{h}}\right]_i\right)\log\left(1-\left[\mathbf{p}_{\mathbf{h}}\right]_i\right)\right)$$

$$L\left(\mathbf{p}_{\mathbf{r}},\tilde{\mathbf{l}}_{\mathbf{r}}\right)=-\frac{1}{|R|}\sum_{i=1}^{|R|}\left(\left[\tilde{\mathbf{l}}_{\mathbf{r}}\right]_i\log\left[\mathbf{p}_{\mathbf{r}}\right]_i+\left(1-\left[\tilde{\mathbf{l}}_{\mathbf{r}}\right]_i\right)\log\left(1-\left[\mathbf{p}_{\mathbf{r}}\right]_i\right)\right)$$

$$L\left(\mathbf{p}_{\mathbf{t}},\tilde{\mathbf{l}}_{\mathbf{t}}\right)=-\frac{1}{|E|}\sum_{i=1}^{|E|}\left(\left[\tilde{\mathbf{l}}_{\mathbf{t}}\right]_i\log\left[\mathbf{p}_{\mathbf{t}}\right]_i+\left(1-\left[\tilde{\mathbf{l}}_{\mathbf{t}}\right]_i\right)\log\left(1-\left[\mathbf{p}_{\mathbf{t}}\right]_i\right)\right)$$

  - 三元组 (h,r,t) 的标准损失函数

$$Loss(h,r,t)=L\left(\mathbf{p}_{\mathbf{h}},\tilde{\mathbf{l}}_{\mathbf{h}}\right)+L\left(\mathbf{p}_{\mathbf{r}},\tilde{\mathbf{l}}_{\mathbf{r}}\right)+L\left(\mathbf{p}_{\mathbf{t}},\tilde{\mathbf{l}}_{\mathbf{t}}\right)$$

SENN模型训练

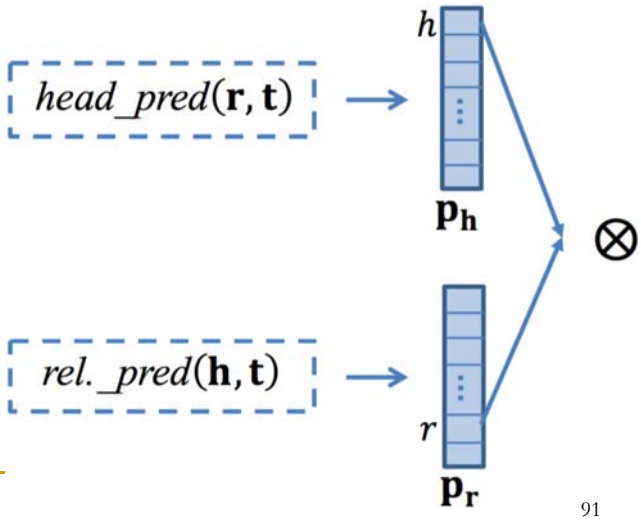
- 自适应加权损失机制
  - 根据正确答案数的不同，自适应地指定推断权重
    - 给予模型推断错确定性高的任务更重的惩罚
  - 根据任务难易程度的不同，指定不同的权重
    - 给予实体预测相对于关系预测更大的权重，使模型偏向于学好难的任务
- 三元组 (h,r,t) 的最终损失函数

$$\widetilde{Loss}(h,r,t)=\frac{w}{|I_h|}L\left(\mathbf{p}_{\mathbf{h}},\tilde{\mathbf{l}}_{\mathbf{h}}\right)+\frac{1}{|I_r|}L\left(\mathbf{p}_{\mathbf{r}},\tilde{\mathbf{l}}_{\mathbf{r}}\right)+\frac{w}{|I_t|}L\left(\mathbf{p}_{\mathbf{t}},\tilde{\mathbf{l}}_{\mathbf{t}}\right)$$

SENN+ 模型

- 在测试过程中利用 $(h,?,t)$ 辅助 $(?,r,t)$和 $(h,r,?)$

- 关系辅助的实体推理
  - 给定 $(?,r,t)$ 并假定 h 是一个有效的头实体
  - 如果进行关系预测 $(h,?,t)$，则 r 最有可能具有更高的预测标签，排在前面



SENN+ 模型

- 关系辅助的实体推理
  - 两个额外的关系辅助向量

$$[\mathbf{q_h}]_i = \mathrm{Value}(g(s(\mathbf{e}_i,\mathbf{t})),r)$$

$$[\hat{\mathbf{q}}_\mathbf{h}]_i = \frac{1}{\mathrm{Rank}(g(s(\mathbf{e}_i,\mathbf{t})),r)}$$

  - 实体预测的最终预测标签向量

$$\tilde{\mathbf{p}}_\mathbf{h} = \mathbf{p_h} \odot \hat{\mathbf{p}}_\mathbf{h} \odot \mathbf{q_h} \odot \hat{\mathbf{q}}_\mathbf{h}$$

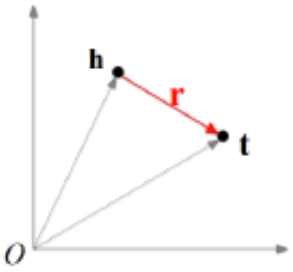
$$\tilde{\mathbf{p}}_\mathbf{t} = \mathbf{p_t} \odot \hat{\mathbf{p}}_\mathbf{t} \odot \mathbf{q_t} \odot \hat{\mathbf{q}}_\mathbf{t}$$

$$[\hat{\mathbf{p}}_\mathbf{h}]_i = \frac{1}{\mathrm{Rank}(\mathbf{p_h},e_i)}$$

$$[\hat{\mathbf{p}}_\mathbf{t}]_i = \frac{1}{\mathrm{Rank}(\mathbf{p_t},e_i)}$$

基于翻译的模型：TransE

- 关系事实=(head, relation, tail) 简写为(h, r, t)，其对应的向量表示为$(\mathbf{h}, \mathbf{r}, \mathbf{t})$



$\mathbf{h}+\mathbf{r}=\mathbf{t}$

中国 + 首都 = 北京

法国 + 首都 = 巴黎

俄罗斯 + 首都 = 莫斯科

Bordes, et al. Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, 2013 (pp. 2787-2795).

翻译模型的学习

- 势能函数
  - 对于真实事实的三元组(h, r, t)，要求$\mathbf{h}+\mathbf{r}=\mathbf{t}$；而对于错误的三元组则不满足该条件

$$f(h,r,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$$

$$f(\text{姚明},\text{出生于},\text{北京}) > f(\text{姚明},\text{出生于},\text{上海})$$

- 损失函数

$$L=\sum_{(h,r,t)\in\triangle}\sum_{(h',r,t')\in\triangle'}\max(0,f_r(h,t)+M_{opt}-f_r(h',t'))$$

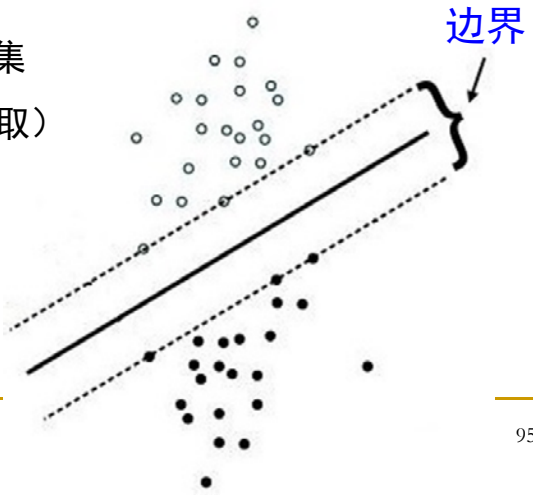
正例三元组集　负例三元组集　最优Margin超参

TransE的损失函数

- 最小化基于边界的损失函数

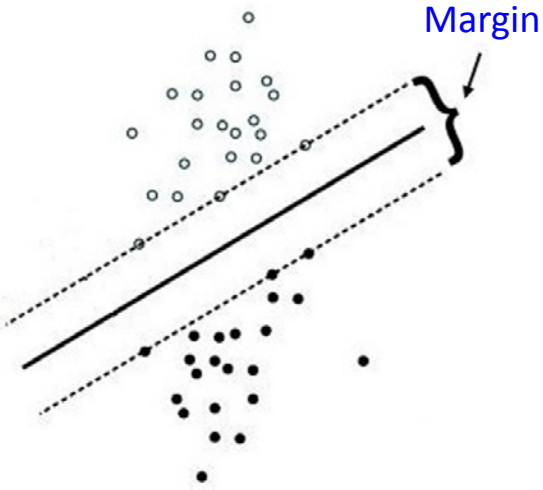
$$L = \sum_{(h,r,t) \in \triangle} \sum_{(h',r,t') \in \triangle'} \max(0, f_r(h,t) + M_{opt} - f_r(h',t'))$$

  - $f_r(h,t)$：得分函数
  - M_{opt}：最优边界，从一个候选闭集（如 $\{0.25, 0.5, 1, 2, \ldots\}$ 中选取）



损失函数的不足

- 最优边界 M_{opt} 在一个候选集中选取
- 不同的知识图谱共享相同的候选集



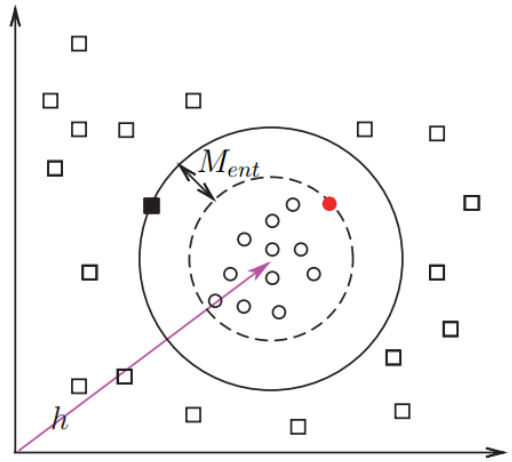
TransA的动机

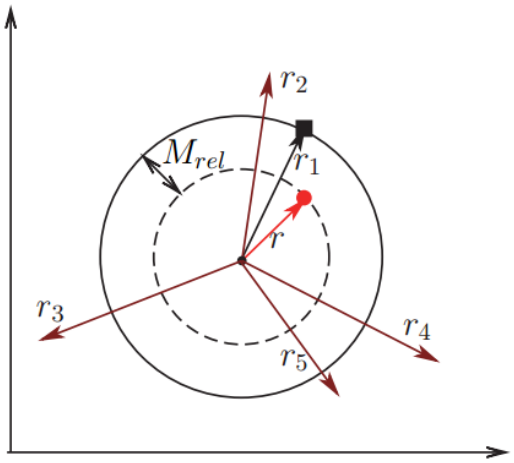
- 知识图谱有不同的局部性，具体表现在有不同的最优边界超参
  - Subset1和Subset2：将FB15K划分为关系数相等的五份，取其中两份作为Subset1和Subset2

Data sets	Optimal loss function	Mean Rank	
		Raw	Filter
Subset1	$f_r(h,t)+3-f_r(h',t')$	339	240
Subset2	$f_r(h,t)+2-f_r(h',t')$	500	365
FB15K	$f_r(h,t)+1-f_r(h',t')$	243	125

- TransA (locally Adaptive Translation method)
  - 自适应地决定不同知识图谱的边界超参

TransA的边界超参





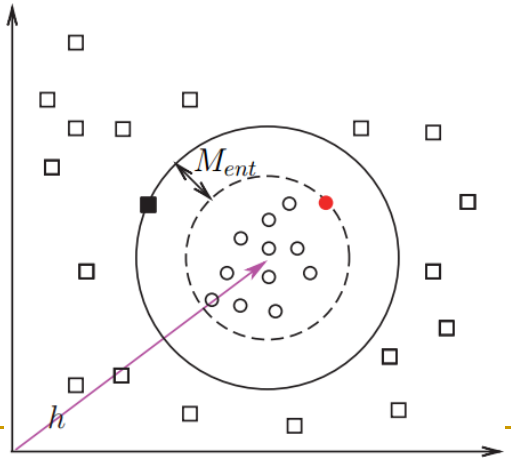
实体之间的局部距离

关系之间的相似性

$$M_{opt}=\mu M_{ent}+(1-\mu)M_{rel}$$

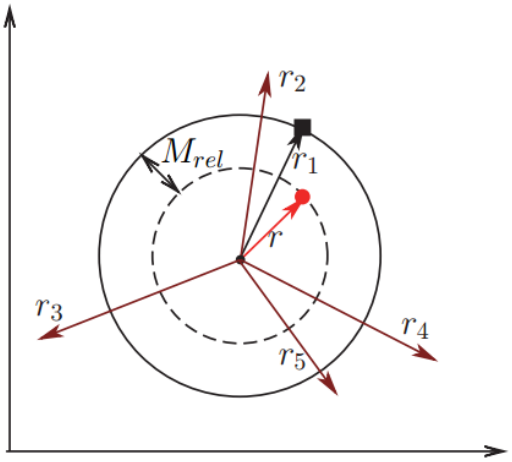
实体依赖的边界超参

- $M_{ent} = \dfrac{\sum_{r \in R_h} \min_{t,t'} \sigma(\|h - t'\| - \|h - t\|)}{|R_h|}$
  - R_h：与 h 有关的关系集
  - $|R_h|$：R_h 的大小，用于区分不同映射属性的关系
  - $\sigma(x) = \begin{cases} x & if\ x \geqslant 0, \\ -x & otherwise. \end{cases}$
  - t 为正例，t'为负例



关系依赖的边界超参

- $M_{rel} = \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|)$
  - $R_{h,r}$：h 对应的除了r 以外的关系集并且 $\|r_i\| \geqslant \|r\|$
  - M_{rel} 由与 r 最相似的关系决定



思考

- 知识抽取
  - 复杂实体抽取（如华为Mate-30手机）
  - 多元关系（如事件）抽取
  - 开放知识抽取
- 知识融合
  - 精度、效率
- 知识推理
  - 多元关系推理
  - 增量式推理

谢谢聆听！