Representation Learning of Knowledge Graphs with Entity Descriptions

AAAI 2016

谢若冰，刘知远，贾珈, Huanbo Luan, 孙茂松

Tsinghua University

outline

- 作者简介
- 论文简介
- 相关工作：transXXX, socher.NTN
- 模型
- 实验内容、效果
- **手动**对比

作者简介

谢若冰

图: 贾珈

- 情感计算、语音交互...

图: 刘知远

- KG，语义计算，“社会计算”...

图: 孙茂松

- 计算语言学，汉语切词...

Huanbo Luan

论文简介

论文简介

1. 以前的词向量从 KB/KG 学习，但是 KG 是非常稀疏的

论文简介

1. 以前的词向量从 KB/KG 学习，但是 KG 是非常稀疏的
2. KB 中对应三元组少的、没有三元组对应（zero-shot）的实体的词向量很不好

论文简介

1. 以前的词向量从 KB/KG 学习，但是 KG 是非常稀疏的
2. KB 中对应三元组少的、没有三元组对应（zero-shot）的实体的词向量很不好
3. 实体可能没有三元组，但一般都有维基百科页面正文
4. 我要把 KB 和自然语言 text 结合，放在一起训练，生成一个**更好**的词向量

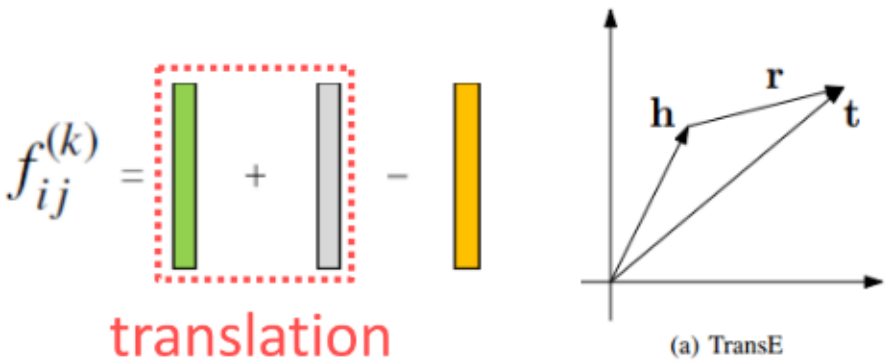
论文简介

1. 以前的词向量从 KB/KG 学习，但是 KG 是非常稀疏的
2. KB 中对应三元组少的、没有三元组对应（zero-shot）的实体的词向量很不好
3. 实体可能没有三元组，但一般都有维基百科页面正文
4. 我要把 KB 和自然语言 text 结合，放在一起训练，生成一个**更好**的词向量

什么叫更好？

比 transE 好就可以叫更好了

$$f\left(e_i, r_k, e_j\right) = \left\|\mathbf{e}_i + \mathbf{r}_k - \mathbf{e}_j\right\|_1$$



相关工作

相关工作-1

transE,transR, PTransE

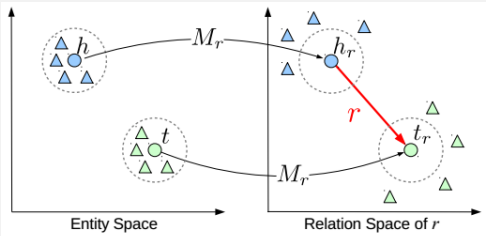


Figure 1: Simple illustration of TransR.

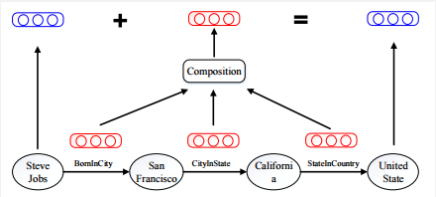
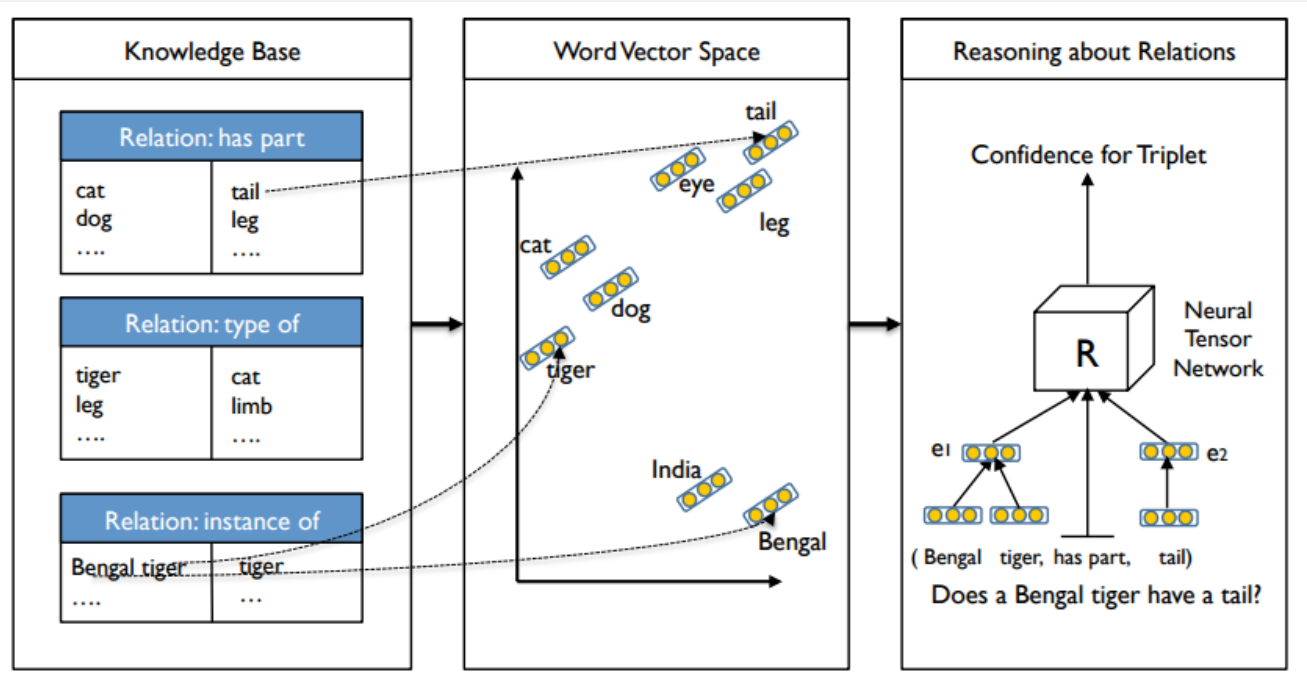


图: PTransE

NTN (neural tensor network, socher)

使用组成当前短语的单词的词向量的平均值能一定程度代表当前单词的词向量



相关工作-2

KB+WikipediaAnchor

利用 WikipediaAnchor([[迈克尔·乔丹 | 乔丹]]?) 来增大单词之间的联系

KB+Description

- 和上一个同一个作者
- 构造一个复杂的目标函数：KB 的目标（h+r-t 小）+ 文本相似（文本中距离近的单词距离小）+description 相似（一个实体的文本中的单词和他距离进）
- 本文作者认为：上述模型没有考虑文本顺序，模型没有考虑/无法避免文本的歧义

模型

模型

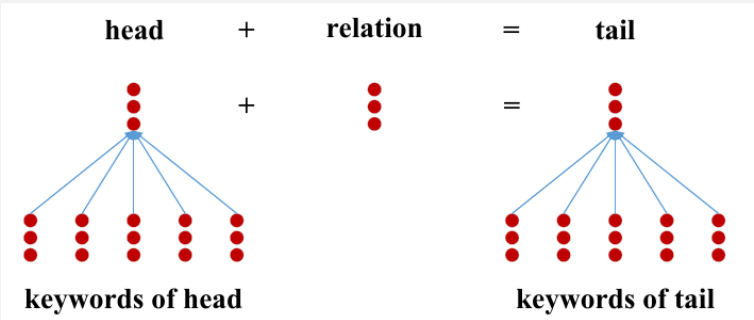
- $(h, r, t) \in T, h, t \in E, r \in R$
- 每个实体（h 或 t）同时训练 2 个向量 h_s(从triple中学习的向量), h_d(从正文中学习的向量)
- 目标函数 $E = E_S + E_D,\ E_D = E_{DD} + E_{DS} + E_{SD},$
  - $E_{DD} = ||\mathbf{h_d} + \mathbf{r} - \mathbf{t_d}||$
  - $E_{DS} = ||\mathbf{h_d} + \mathbf{r} - \mathbf{t_s}||$ and $E_{SD} = ||\mathbf{h_s} + \mathbf{r} - \mathbf{t_d}||,$

模型-Encoder

- 2 种 Encoder 生成从正文中学习的文本向量：CBOW、CNN

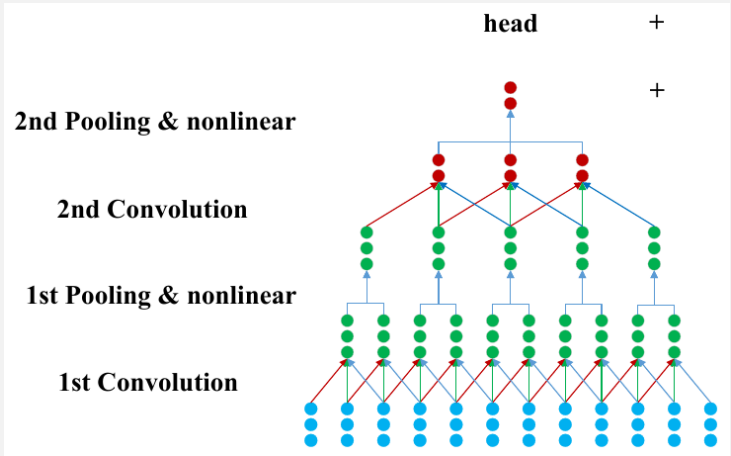
CBOW

1. 根据 tf-idf 值，在每个实体页面中取最重要的 20 个单词作为他的文本特征
2. 将这 20 个单词的词向量的平均值作为该实体的文本向量的值



CNN

1. (1->2).1 对将连续的 2 个单词的词向量拼接得到一个长的词向量 x_i'
2. (1->2).2 卷一下：$\mathbf{z}_i^{(l)} = \sigma(\mathbf{W}^{(l)}\mathbf{x}_i'^{(l)} + \mathbf{b}_i^{(l)})$,
3. (2->3) max-pooling: $\mathbf{x}_i^{(2)} = \max(\mathbf{z}_{n\cdot i}^{(1)}, \cdots, \mathbf{z}_{n\cdot(i+1)-1}^{(1)})$.
4. (3->4) 卷一下：同 $z_i^{(1)}$, 得到 $z_i^{(2)}$
5. (4->5) mean-pooling: $\mathbf{x}^{(3)} = \sum_{i=1,\cdots,m} \frac{\mathbf{z}_i^{(2)}}{m}$,



模型-训练

目标函数

$$L = \sum_{(h,r,t)\in T} \sum_{(h',r',t')\in T'} \max(\gamma + d(h+r,t) - d(h'+r',t'), 0),$$

负样本：

$$T' = \{(h',r,t)|h' \in E\} \cup \{(h,r,t')|t' \in E\} \cup \{(h,r',t)|r' \in R\}$$

- 距离函数 d 为 L1 范式（绝对值相加，又称曼哈顿距离）
- 负样本中的新实体向量既可以使用 triple 向量，也可以使用文本向量
- 待训练的参数集合为 $(\mathbf{X}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{E}, \mathbf{R})$：文本向量、2 个卷积矩阵、实体向量、关系向量
- X 由 word2vec 在维基百科上先跑好，E, R 可以随机初始化，也可以使用 transE 的结果

实验

实验-数据集、参数

DATA SET	WN	FB15K
ENTITIES	40,943	14,951
RELATIONSHIPS	18	1,345
TRAIN. EX.	141,442	483,142
VALID EX.	5,000	50,000
TEST EX.	5,000	59,071

图: 原始的 FB15k

Dataset	#Rel	#Ent	#Train	#Valid	#Test
FB15K	1,341	14,904	472,860	48,991	57,803
Dataset	#Ent	$\#e-e$	$\#d-e$	$\#e-d$	$\#d-d$
FB20K	19,923	57,803	18,753	11,586	151

图: 本实验的数据集

- 作者去除了正文过短的实体（<3 个单词）
- 为了验证 Zero-shot，作者将 fb15k 扩充成 fb20k：
  - 从 freebase 中随机选和 fb15k 有 triple 关系的新点加进来，最后把新点和集合中其他实体的 triple 边加进来
- 参数：学习率 $\lambda = 0.001$ margin:$\gamma = 1$ 卷积初始合并 $k = 2$ 个连续单词词向量 $n = 100$ 维，$n_w = 100, n_f = 100$

实验-KBC-1

补全三元组，预测实体：(h, r, t): 给定 r, t, 求 h；或者给定 h, r, 求 t

Metric	Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter
TransE	210	119	48.5	66.1
DKRL(CBOW)	236	151	38.3	51.8
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN)+TransE	**181**	**91**	**49.6**	**67.4**

- filter 指的是把预测结果中在训练集、测试集中出现的三元组删掉之后的效果
  - 理论上在训练集中出现的三元组不会在测试集中出现，不可能是答案
- transE 原论文提供的 HIT@10 数据是 34.9, 47.1
  - 作者强调他自己实现了 transE，比原论文效果好（实验结果没问题）
  - 结论说自己的方法比 transE 有明显提升

实验-KBC-1

补全三元组，预测实体：(h, r, t): 给定 r, t, 求 h；或者给定 h, r, 求 t

Metric	Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter
TransE	210	119	48.5	66.1
DKRL(CBOW)	236	151	38.3	51.8
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN)+TransE	**181**	**91**	**49.6**	**67.4**

- filter 指的是把预测结果中在训练集、测试集中出现的三元组删掉之后的效果
  - 理论上在训练集中出现的三元组不会在测试集中出现，不可能是答案
- transE 原论文提供的 HIT@10 数据是 34.9, 47.1
  - 作者强调他自己实现了 transE，比原论文效果好（实验结果没问题）
  - 结论说自己的方法比 transE 有明显提升
  - 有点自相矛盾

实验-KBC-2

补全三元组，预测关系：(h, r, t): 给定 h, t, 求 r

Metric	Mean Rank		Hits@1(%)	
	Raw	Filter	Raw	Filter
TransE	2.91	2.53	69.5	90.2
DKRL(CBOW)	2.85	2.51	65.3	82.7
DKRL(CNN)	2.91	2.55	**69.8**	89.0
DKRL(CNN)+TransE	**2.41**	**2.03**	**69.8**	**90.8**

实验-分类

1. 取 FB15K 中实体的所有类别并统计频率，取出现频率最高的 50 个作为分类的候选

Metric	FB15K	FB20K
TransE	87.9	-
BOW	86.3	57.5
DKRL(CBOW)	89.3	52.0
DKRL(CNN)	**90.1**	**61.9**

- BOW：词袋模型，一个 one-hot 的超大 vector，长度为词典大小
- 文本信息对于预测类别有相对明显的提升

实验-zero-shot

- 如果测试集中的实体在训练集中没有出现，则不可能训练出词向量，更不可能找到答案
- 在 FB20K 上看看本文效果（FB15K 的测试集中大多数实体都出现过了）
- 使用 FB15K 的训练集，测试集改成 FB20K 多的那 5K 个实体引入的新三元组

1. 预测实体

Metric	$d-e$	$e-d$	$d-d$	Total
Partial-CBOW	26.5	20.9	67.2	24.6
CBOW	27.1	21.7	66.6	25.3
Partial-CNN	26.8	20.8	69.5	24.8
CNN	**31.2**	**26.1**	**72.5**	**29.5**

- $d-e$ 表示 head 是新的实体，tail 是原来 FB15K 训练集中的实体
- Partial-XX 表示测试数据的时候在训练集中出现的实体用 triple 向量表示；否则，所有数据都用文本向量表示

实验-zero-shot

- 如果测试集中的实体在训练集中没有出现，则不可能训练出词向量，更不可能找到答案
- 在 FB20K 上看看本文效果（FB15K 的测试集中大多数实体都出现过了）
- 使用 FB15K 的训练集，测试集改成 FB20K 多的那 5K 个实体引入的新三元组

1. 预测关系

Metric	$d-e$	$e-d$	$d-d$	Total
Partial-CBOW	49.0	42.2	0.0	46.2
CBOW	52.2	47.9	0.0	50.3
Partial-CNN	56.6	52.4	4.0	54.8
CNN	**60.4**	**55.5**	**7.3**	**58.2**

手工对比

手工对比

- 原始的文章到 zero-shot 实验之后就结束了，开始总结
- 没有对比

Aligning knowledge and text embeddings by entity descriptions

- EMNLP 2015
- 构造一个复杂的目标函数：KB 的目标（h+r-t 小）+ 文本相似（文本中距离近的单词距离小）+description 相似（一个实体的文本中的单词和他距离进）

Representing Text for Joint Embedding of Text and Knowledge Bases

- EMNLP 2015
- 用 CNN encode 文本中的 d-path

手工对比-Aligning knowledge and text embeddings by entity descriptions

- 中山大学 + 微软
- 目标函数 3 方面：$\mathcal{L}\left(\{\mathbf{e}_i\},\{\mathbf{r}_j\},\{\mathbf{w}_l\}\right)=\mathcal{L}_K+\mathcal{L}_T+\mathcal{L}_A$
  - KB 的优化目标（h+r-t 小）+ 文本相似（文本中距离近的单词距离小）+description 相似（一个实体的文本中的单词和他距离进）
  - 以第一方面为例：$z(h,r,t)=b-0.5\cdot\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2^2$,

$$\Pr(h|r,t)=\frac{\exp\{z(h,r,t)\}}{\sum_{\tilde{h}\in\mathcal{I}}\exp\{z(\tilde{h},r,t)\}}$$

$$\mathcal{L}_K=-\sum_{(h,r,t)}\big[\log\Pr(h|r,t)+\log\Pr(t|h,r)+\log\Pr(r|h,t)\big]$$

- 对数据没有做过清洗

Metric	MEAN		HITS@10	
	Raw	Filtered	Raw	Filtered
TransE	243	125	34.9	47.1
Jointly(anchor)	**166**	47	49.9	72.0
Jointly(desp)	167	**39**	**51.7**	**77.3**

Metric	Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter
TransE	210	119	48.5	66.1
DKRL(CBOW)	236	151	38.3	51.8
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN)+TransE	**181**	**91**	**49.6**	**67.4**

手工对比-Aligning knowledge and text embeddings by entity descriptions

- 中山大学 + 微软
- 目标函数 3 方面：$\mathcal{L}\left(\{\mathbf{e}_i\},\{\mathbf{r}_j\},\{\mathbf{w}_l\}\right)=\mathcal{L}_K+\mathcal{L}_T+\mathcal{L}_A$
  - KB 的优化目标（h+r-t 小）+ 文本相似（文本中距离近的单词距离小）+description 相似（一个实体的文本中的单词和他距离进）
  - 以第一方面为例：$z(h,r,t)=b-0.5\cdot\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2^2$,

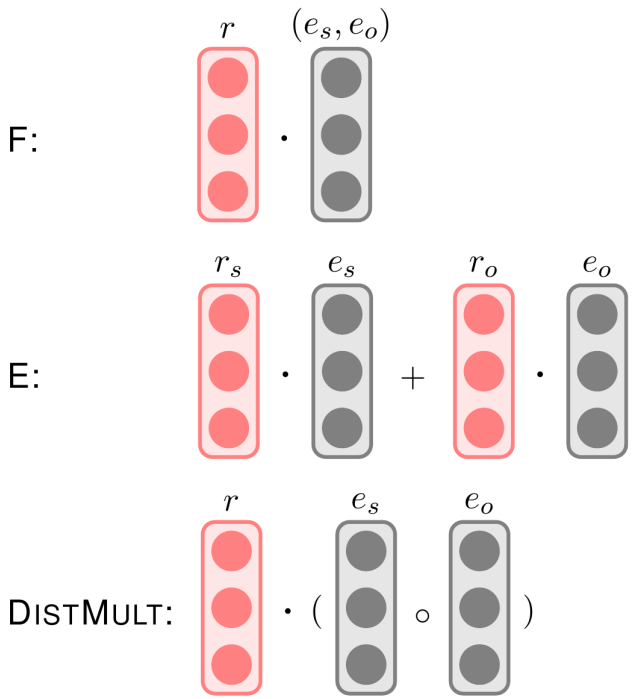
$$\Pr(h|r,t)=\frac{\exp\{z(h,r,t)\}}{\sum_{\tilde{h}\in\mathcal{I}}\exp\{z(\tilde{h},r,t)\}}$$

$$\mathcal{L}_K=-\sum_{(h,r,t)}\big[\log\Pr(h|r,t)+\log\Pr(t|h,r)+\log\Pr(r|h,t)\big]$$

- 对数据没有做过清洗
- 优点：效果好
- 缺点：归一化的概率计算复杂度大？

手工对比-Representing Text for Joint Embedding of Text and Knowledge Bases

- Stanford+ 微软
- 不同于 transE，提出 E 模型,DISTMULT 模型来衡量三元组的可靠性
- 作者提取文本中的依存关系作为 relation 和 entity 的关联关系来训练 relation 向量



E 模型

$$f(e_s, r, e_o) = v(r_s)^\intercal v(e_s) + v(r_o)^\intercal v(e_o)$$

DISTMULT 模型

$$f(e_s, r, e_o) = v(r)^\intercal \left(v(e_s) \circ v(e_o)\right)$$

手工对比-Representing Text for Joint Embedding of Text and Knowledge Bases

- Stanford+ 微软
- 不同于 transE，提出 E 模型,DISTMULT 模型来衡量三元组的可靠性
- 作者提取文本中的依存关系作为 relation 和 entity 的关联关系来训练 relation 向量

E + DISTMULT ($\tau = 0.01$)	37.7	55.7
CONV-E + CONV-DISTMULT ($\tau = 0.25$)	**40.1**	**58.1**

Metric	Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter
TransE	210	119	48.5	66.1
DKRL(CBOW)	236	151	38.3	51.8
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN)+TransE	**181**	**91**	**49.6**	**67.4**

- 这篇文章作者使用 FB15K-237，更小/好的一个 FB15K 的子集
- 最后的效果并不比本文好
- TransE 其实已经是一个很好的模型了 (?)

总结

- text + KB 训练词向量，有提升，但很依赖怎么从文本中提取信息
  - CNN 模型一般有小量提升，CBOW 模型一般没有提升
  - 怎么表示 text 中的信息很重要
- TransE 模型已经很好了，工程使用应该够？

问题？