

Representing Text for Joint Embedding of Text and Knowledge Bases

Kristina Toutanova@MS

Danqi Chen@Stanford

EMNLP2015

韩喆

ICSTWIP

20160107

outline

- ▶ 作者简介
- ▶ 问题定义: Knowledge Base + 文本抽取新关系
- ▶ 相关工作: KBC、文本、混合模型
- ▶ 实验
 - ▶ 模型介绍
 - ▶ 实验结果

作者简介



- ▶ 保加利亚/美国人?
- ▶ Sofia-Uni -> stanford.NLP -> MS
- ▶ 句法语法分析, MT, 摘要, ... 都做



- ▶ THU.Yao -> stanford.NLP
- ▶ DL, NLP

相关人物,
非作者:



- ▶ Phd@Edinburgh -> researcher@UTokyo -> researcher@umass -> AP@UCL.Machine Reading Lab
- ▶ 机器阅读, NLP
- ▶ 主页放了一个叫 Mika Riedel 的日本女生的绘画/雕刻作品, 貌似是他老婆?

任务简介

- ▶ 给定 RDF 知识库 $KB = \{(e_s, r, e_o), \dots\}$
- ▶ 问 KB 中没有的关系，找出最合适的实体： $(e_s, r, ?)(?, r, e_o)$
 - ▶ 类似的还有 $(e_s, ?, e_o)$ ，这里面候选 r 可能有多个？
 - ▶ 本文认为候选只有一个？
 - ▶ 本文只对第一个效果进行实验

任务简介

- ▶ 给定 RDF 知识库 $KB = \{(e_s, r, e_o), \dots\}$
- ▶ 问 KB 中没有的关系，找出最合适的实体： $(e_s, r, ?)(?, r, e_o)$
 - ▶ 类似的还有 $(e_s, ?, e_o)$ ，这里面候选 r 可能有多个？
 - ▶ 本文认为候选只有一个？
 - ▶ 本文只对第一个效果进行实验
- ▶ $e_o = \arg \max_{e_j} f((e_s, r, e_j))$
- ▶ $f((e_s, r, e_o)) = f_2(v_{e_s}, v_r, v_{e_o})$
 - ▶ 一个好的语义向量 $\mathcal{R} = \{r_1, r_2, \dots\}, \mathcal{E} = \{e_1, e_2, \dots\}$
 - ▶ 一个好的模型 f_2

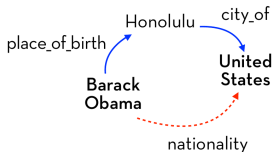
任务简介

- ▶ 基于 Riedel 13 年的文章，从 KB+text 中抽取语义向量（实体的向量，关系的向量），做关系预测

Relation extraction with matrix factorization and universal schemas

- ▶ 将 text 中的 dependency path 作为实体的“新关系”，增大图的密集程度（关系数量显著增加）

Knowledge Base



Freebase

Textual Mentions

Barack Obama is the 44th and current President of United States.

Obama was born in the United States just as he has always said.

ClueWeb

Lemur

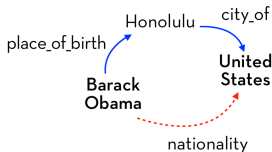
任务简介

- ▶ 基于 Riedel 13 年的文章，从 KB+text 中抽取语义向量（实体的向量，关系的向量），做关系预测

Relation extraction with matrix factorization and universal schemas

- ▶ 将 text 中的 dependency path 作为实体的“新关系”，增大图的密集程度（关系数量显著增加）

Knowledge Base



Freebase

相似的 dependency path 被看成不同的“新关系”，尽管他们只有很小的差别

Textual Mentions

Barack Obama is the 44th and current President of United States.

Obama was born in the United States just as he has always said.

ClueWeb

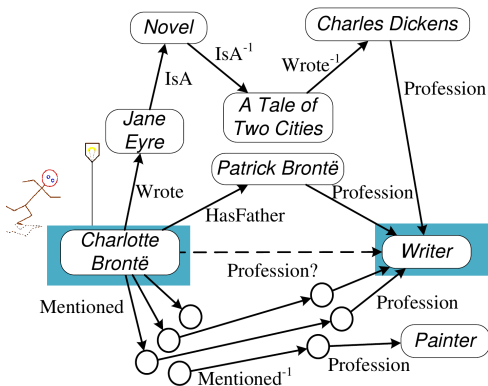
Lemur

相关工作

- Knowledge base completion
- Relation extraction using distant supervision
- Combining KB and text

Knowledge base completion

- ▶ 初始版的 path ranking algorithm (劳逆 2011)
 - ▶ 我的理解：基本想法是主语和相似实体有相同的属性/属性值，近似于双向 random walk+ 剪枝



Knowledge base completion

- ▶ DISTMULT (后面讲, Yang 2015)
- ▶ TransE (之前讲过, 贴在下面)

- **Translating Embeddings for Modeling Multi-relational Data**

- 隐变量表示

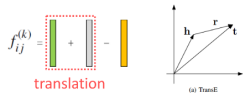


- 打分函数 $f(e_i, r_k, e_j) = \|e_i + r_k - e_j\|_1$

- 基本思想是h (头词) 经过r (关系) 迁移(translation) 之后的向量与相应的t(尾词)之间的差值, 越小表示越匹配

- 目标函数 $\min_{\{e_i\}, \{r_k\}} \sum_{r^+ \in O} \sum_{r^- \in N_{r^+}} [\gamma + f(e_i, r_k, e_j) - f(e'_i, r_k, e'_j)]_+$

- 最小化基于间隔的排序损失函数



Relation extraction using distant supervision

- ▶ 即单纯从文本中抽取实体的关系，不使用 KB
 - ▶ 使用 dependency path 建立实体间的关系
 - ▶ 没用考虑 d-path 里面的相同子结构
 - ▶ 比较老，09-11 年

Combining knowledge base and text information

同时利用 KB 和 text 来抽取新关系

1. 后来的 path ranking algorithm (劳逆 2012)
 - ▶ 从 text 中抽取 text-graph (实体是点, 关系是边), 解决边的稀疏性问题
2. (Neelakantan 2015) 使用在文本中共现的实体对增加图中边
3. 有的是训练每个实体和所有在实体中出现的单词
 - ▶ 含有相似单词的实体会训练出相似的词向量?
4. 分别在 KB 和 text 中训练不同的词向量
5. 没用考虑 d-path 里面的相同子结构

模型介绍

目标函数

目标函数 $L(\mathcal{T}_{\text{KB}}; \Theta) + \tau L(\mathcal{T}_{\text{text}}; \Theta) + \lambda \|\Theta\|^2$

$$L(\mathcal{T}; \Theta) = - \sum_{(e_s, r, e_o) \in \mathcal{T}} \log p(e_o | e_s, r; \Theta) \\ - \sum_{(e_s, r, e_o) \in \mathcal{T}} \log p(e_s | e_o, r; \Theta)$$

►

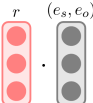
$$p(e_o | e_s, r; \Theta) = \frac{e^{f(e_s, r, e_o; \Theta)}}{\sum_{e' \in \text{Neg}(e_s, r, ?)} e^{f(e_s, r, e'; \Theta)}}$$

►

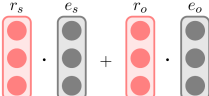
- 权重变量 τ 手动确定, $f((e_s, r, e_o))$ 采用不同的模型 (或他们的 f 函数的值简单相加), 下面具体说明这几个 f 函数

模型介绍

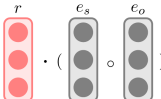
F:


$$r \cdot (e_s, e_o)$$

E:


$$r_s \cdot e_s + r_o \cdot e_o$$

DISTMULT:

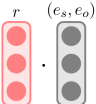

$$r \cdot (e_s \ominus e_o)$$

E 模型

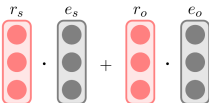
$$f(e_s, r, e_o) = v(r_s)^T v(e_s) + v(r_o)^T v(e_o)$$

模型介绍

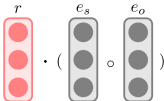
F:


$$r \cdot (e_s, e_o)$$

E:


$$r_s \cdot e_s + r_o \cdot e_o$$

DISTMULT:

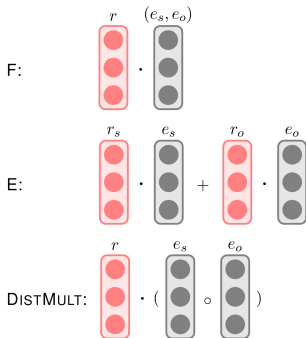

$$r \cdot (e_s \circ e_o)$$

E 模型

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

e_o 和 e_s 无关，只和 r 有关

模型介绍



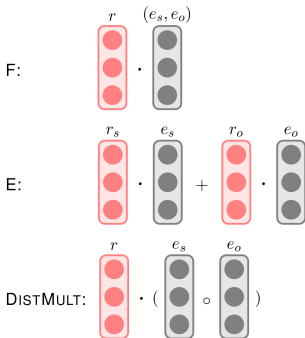
E 模型

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

DISTMULT 模型

$$f(e_s, r, e_o) = v(r)^T (v(e_s) \circ v(e_o))$$

模型介绍



E 模型

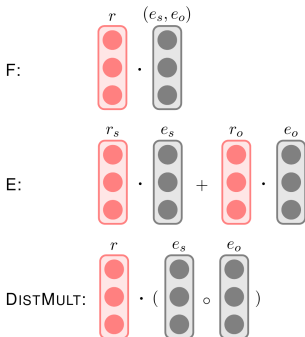
$$f(e_s, r, e_o) = v(r_s)^T v(e_s) + v(r_o)^T v(e_o)$$

DISTMULT 模型

$$f(e_s, r, e_o) = v(r)^T (v(e_s) \circ v(e_o))$$

(e_s, r, e_o) 和 (e_s, r, e_o) 一样，学出来的 r 向量不能区分反向关系

模型介绍



E 模型

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

DISTMULT 模型

$$f(e_s, r, e_o) = v(r)^T (v(e_s) \circ v(e_o))$$

复杂度, $N_e = |\mathcal{E}|$, $N_r = |\mathcal{R}|$, K 为维度

- ▶ E: $KN_e + 2KN_r$
- ▶ DISTMULT: $KN_e + KN_r$
- ▶ F: $KN_e^2 + KN_r$

模型介绍

E 模型

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

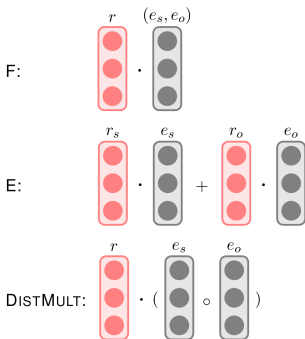
DISTMULT 模型

$$f(e_s, r, e_o) = v(r)^T (v(e_s) \circ v(e_o))$$

复杂度, $N_e = |\mathcal{E}|$, $N_r = |\mathcal{R}|$, K 为维度

- ▶ E: $KN_e + 2KN_r$
- ▶ DISTMULT: $KN_e + KN_r$
- ▶ F: $KN_e^2 + KN_r$

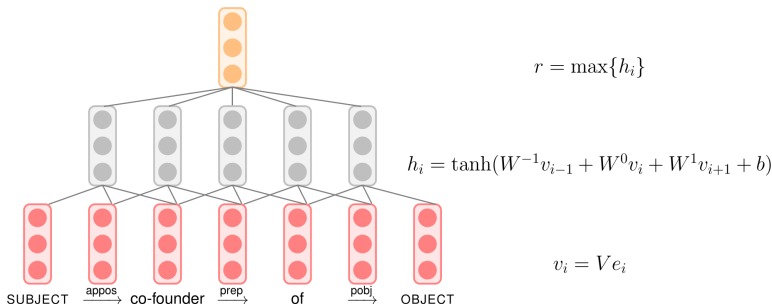
E、F 模型由 Riedel 提出, DISTMULT 由 Yang 提出, 作者直接使用



模型介绍

CONV: Compositional Representations of Textual Relations

- 为了解决从句子中抽取关系时两个很相似的实体对所使用的 d-path 被当成不同的关系
 - 过去 $r_{co-founder_{of}}$ 和 KB 中的关系类似, 只和 KB 中的实体有关系
 - 现在 $v_{r_{co-founder_{of}}}$ 由 CNN 的输出层决定, 和 d-path 里面的每个节点都有关系



实验

实验

KB:FB15k-237

FB15k 的子集, 237 种关系

text:ClueWeb12

覆盖了 13.9k/14.5k 个实体

	# Relations	# Entities	# Triples in Train / Validation / Test
KB	237	14,541	272,115 / 17,535 / 20,466
Text	2,740k	13,937	3,978k / 0 / 0

- ▶ 可以极大提高 F 模型学习实体对向量的效果
 - ▶ 降低稀疏度
 - ▶ 训练集/开发集/测试集中 40%/26%/28% 的（主语，客体）实体对在 ClueWeb 中出现
 - ▶ 上面开发集的 26% 的实体对中有 18% 在训练集中出现，所以有 5% 的可能实体对在训练集看过，比不用文本的概率提高了 50 倍

实验

KB:FB15k-237

FB15k 的子集, 237 种关系

text:ClueWeb12

覆盖了 13.9k/14.5k 个实体

	# Relations	# Entities	# Triples in Train / Validation / Test
KB	237	14,541	272,115 / 17,535 / 20,466
Text	2,740k	13,937	3,978k / 0 / 0

► 剪枝策略

- 只有满足 relation 客体类别的实体才会被纳入候选 (去除完全不可能的实体)
- 如果实体在训练、开发、测试集中出现, 则不会被纳入候选
 - 这些实体可能是正确的, 会导致想要的答案没有排在第一位
- 训练时每个三元组选取 200 个负样本, 权重变量 $\tau = 0.25$, 向量维度为 10 时, 效果最好

实验——参数选择

Model	Overall		With mentions		Without mentions	
	MRR	HITS@10	MRR	HITS@10	MRR	HITS@10
KB only						
F	16.9	24.5	26.4	49.1	13.3	15.5
E	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT	35.7	52.3	26.0	39.0	39.3	57.2
E+DISTMULT	37.3	55.2	28.6	42.9	40.5	59.8
F+E+DISTMULT	33.8	50.1	15.0	26.1	40.7	59.0
KB and text						
F ($\tau = 1$)	19.4	27.9	35.4	61.6	13.4	15.5
CONV-F ($\tau = 1$)	19.2	28.4	34.9	63.7	13.3	15.4
E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
CONV-E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT ($\tau = 0.01$)	36.1	52.7	26.5	39.5	39.6	57.5
CONV-DISTMULT ($\tau = 0.25$)	36.6	53.5	28.3	43.4	39.7	57.2
E + DISTMULT ($\tau = 0.01$)	37.7	55.7	28.9	43.4	40.9	60.2
CONV-E + CONV-DISTMULT ($\tau = 0.25$)	40.1	58.1	33.9	49.9	42.4	61.1

- ▶ with mentions: 测试集中的实体对在 text 中出现过
- ▶ 非 F 模型的第三、第二列好像弄反了
- ▶ 对于 F 模型, 只训练在 text 中出现的实体对 (减小复杂度)

实验——参数选择

Model	Overall		With mentions		Without mentions	
	MRR	HITS@10	MRR	HITS@10	MRR	HITS@10
KB only						
F	16.9	24.5	26.4	49.1	13.3	15.5
E	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT	35.7	52.3	26.0	39.0	39.3	57.2
E+DISTMULT	37.3	55.2	28.6	42.9	40.5	59.8
F+E+DISTMULT	33.8	50.1	15.0	26.1	40.7	59.0
KB and text						
F ($\tau = 1$)	19.4	27.9	35.4	61.6	13.4	15.5
CONV-F ($\tau = 1$)	19.2	28.4	34.9	63.7	13.3	15.4
E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
CONV-E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT ($\tau = 0.01$)	36.1	52.7	26.5	39.5	39.6	57.5
CONV-DISTMULT ($\tau = 0.25$)	36.6	53.5	28.3	43.4	39.7	57.2
E + DISTMULT ($\tau = 0.01$)	37.7	55.7	28.9	43.4	40.9	60.2
CONV-E + CONV-DISTMULT ($\tau = 0.25$)	40.1	58.1	33.9	49.9	42.4	61.1

- ▶ F 的实体对太稀疏，表现不好
- ▶ E 的效果比较好，但是实际上和主语没关系，所以效果不如 DISTMULT
- ▶ E+DISTMULT 最好

实验——参数选择

Model	Overall		With mentions		Without mentions	
	MRR	HITS@10	MRR	HITS@10	MRR	HITS@10
KB only						
F	16.9	24.5	26.4	49.1	13.3	15.5
E	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT	35.7	52.3	26.0	39.0	39.3	57.2
E+DISTMULT	37.3	55.2	28.6	42.9	40.5	59.8
F+E+DISTMULT	33.8	50.1	15.0	26.1	40.7	59.0
KB and text						
F ($\tau = 1$)	19.4	27.9	35.4	61.6	13.4	15.5
CONV-F ($\tau = 1$)	19.2	28.4	34.9	63.7	13.3	15.4
E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
CONV-E ($\tau = 0$)	33.2	47.6	25.5	37.8	36.0	51.2
DISTMULT ($\tau = 0.01$)	36.1	52.7	26.5	39.5	39.6	57.5
CONV-DISTMULT ($\tau = 0.25$)	36.6	53.5	28.3	43.4	39.7	57.2
E + DISTMULT ($\tau = 0.01$)	37.7	55.7	28.9	43.4	40.9	60.2
CONV-E + CONV-DISTMULT ($\tau = 0.25$)	40.1	58.1	33.9	49.9	42.4	61.1

- ▶ 如果使用 text 信息
- ▶ CONV 对 E 没有提升, 对 DISTMULT 有小提升
- ▶ 随机初始化的词向量 (40.3%) 和使用从 KB-only 训练出的词向量效果 (38.7%) 差别较大
- ▶ CNN 的窗口大小影响不大

相关研究

- ▶ F 模型和 E、DISTMULT 模型不是一回事
- ▶ F 太稀疏了，应该填得更满一些
 - ▶ NACCL2015 Injecting Logical Background for Relation Extraction
 - ▶ 通过自动的挖掘 KB 里面的高可信度的一阶逻辑（比如 $professorAt(x, y) \Rightarrow employeeAt(x, y)$ 来“填充”稀疏的矩阵）