TransE

> 《Translating Embeddings for Modeling Multi-relational Data》

任务

- **在低维向量空间中，将多种关系的图谱中的实体和关系在一个低维空间中进行表示，获得每个实体的表征结果。**
- **提出一种易于训练的规范模型，该模型包含数量较少的参数，并且可以扩展到非常大的知识库。**
- **对知识图谱中的多元关系数据进行建模，在不引入额外知识的情况下，高效的实现知识补全，关系预测。**

方法（模型）

TransE：基于能量的模型，用于学习实体的低维嵌入。

1. 关系作为向量空间转变的桥梁：如果三元组`(h,l,t)`成立，则头实体embedding和关系embedding相加约等于尾实体的embedding。

   > $h + l \approx t$

2. 利用空间传递不变形，找到一个实体和向量空间，使得整关系三元组之间的势能差值最小。

   > $min(t - (h + l))$

3. 模型

- 给定一个训练集 S，三元组表示为 (h, l, t)，其中 $h, t \in E, l \in L$，实体和关系的嵌入维度设为 k，希望 $h + l$ 与 t能够尽可能的相似，因此定义一个能量函数：

  $d(h + l, t) = [(h + l) - t]^2 = || h ||_2^2 + || l ||_2^2 + || t ||_2^2 - 2(h^T t + l^T (t - h))$

  > 欧式距离

- 为了训练实体embedding和关系embedding，需要引入负样本。目标是尽可能对正样本中最小化 $d(h + l, t)$，负样本中则尽可能最大化$d(h' + l, t')$。h', t' 表示不属于某个三元组的实体。因此可以得出基于间距排序标准目标优化函数（**损失函数**）：

  $L = \sum_{(h,\ell,t) \in S} \sum_{(h',\ell,t') \in S'_{(h,\ell,t)}} [\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+$

  > 其中 $[x]_+$ 表示 x 中正例的部分，$\gamma > 0$ 表示距离因子。
  >
  > 通过最小化正样本的损失，最大化负样本的距离，达到优化嵌入表示的目的。

- 错误三元组生成：将正确三元组的头或者尾替换成其他的（每次只能选择头或者尾进行替换，不同时替换），得到错误的三元组。

  $S'_{(h,l,t)} = (h', l, t) \mid h' \in E \cup (h, l, t') \mid t' \in E$

测试数据集

- FreeBase
- WordNet

DATA SET	WN	FB15K	FB1M
ENTITIES	40,943	14,951	1×10^6
RELATIONSHIPS	18	1,345	23,382
TRAIN. EX.	141,442	483,142	17.5×10^6
VALID EX.	5,000	50,000	50,000
TEST EX.	5,000	59,071	177,404

性能水平

链接预测

评价方法

- 对于每个三元组，都将头部移除并依次替换为字典中的任意一个实体。

Table 3: **Link prediction results.** Test performance of the different methods.

DATASET	WN				FB15K				FB1M	
METRIC	MEAN RANK		HITS@10 (%)		MEAN RANK		HITS@10 (%)		MEAN RANK	HITS@10 (%)
Eval. setting	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Raw*
Unstructured [2]	315	304	35.3	38.2	1,074	979	4.5	6.3	15,139	2.9
RESCAL [11]	1,180	1,163	37.2	52.8	828	683	28.4	44.1	-	-
SE [3]	1,011	985	68.5	80.5	273	162	28.8	39.8	22,044	17.5
SME(LINEAR) [2]	545	533	65.1	74.1	274	154	30.7	40.8	-	-
SME(BILINEAR) [2]	526	509	54.7	61.3	284	158	31.3	41.3	-	-
LFM [6]	469	456	71.4	81.6	283	164	26.0	33.1	-	-
TransE	**263**	**251**	**75.4**	**89.2**	**243**	**125**	**34.9**	**47.1**	**14,615**	**34.0**

- raw：原始数据
- filtered：移除错误三元组

> 某些错误的三元组会变成有效的三元组。在测试中，可能会出现某些错误三元组排序比测试集三元组靠前的情况，但是这些三元组都是真实的。为了解决这个缺陷对评价指标带来的影响，从数据集中删除错误的三元组。

结论

在原始数据集和去除错误的三元组之后的数据集上，TransE均具有较低的平均排名和较高的hits@10排名。

四种类型的实体预测[1-1,1-N,N-1,N-N]

- 根据头实体和尾实体的对应关系划分。
- 给定关系和实体预测另一个实体。

Table 4: **Detailed results by category of relationship.** We compare Hits@10 (in %) on FB15k in the filtered evaluation setting for our model, TransE and baselines. (M. stands for MANY).

TASK	PREDICTING *head*				PREDICTING *tail*			
REL. CATEGORY	1-TO-1	1-TO-M.	M.-TO-1	M.-TO-M.	1-TO-1	1-TO-M.	M.-TO-1	M.-TO-M.
Unstructured [2]	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE [3]	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME(LINEAR) [2]	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME(BILINEAR) [2]	30.9	**69.6**	**19.9**	38.6	28.2	13.1	**76.0**	41.8
TransE	**43.7**	65.7	18.2	**47.2**	**43.7**	**19.7**	66.7	**50.0**

结论

TransE在1-1的情况下预测效果较好。

TransE在FB15k测试集上的样例预测

- 粗体是测试元组正确的尾部，斜体是训练集上其它正确的尾部。

Table 5: **Example predictions** on the FB15k test set using TransE. **Bold** indicates the test triplet's true tail and *italics* other true tails present in the training set.

INPUT (HEAD AND LABEL)	PREDICTED TAILS
J. K. Rowling `influenced by`	*G. K. Chesterton*, J. R. R. Tolkien, *C. S. Lewis*, **Lloyd Alexander**, Terry Pratchett, Roald Dahl, Jorge Luis Borges, *Stephen King*, Ian Fleming
Anthony LaPaglia `performed in`	*Lantana*, *Summer of Sam*, *Happy Feet*, *The House of Mirth*, Unfaithful, **Legend of the Guardians**, Naked Lunch, X-Men, The Namesake
Camden County `adjoins`	**Burlington County**, *Atlantic County*, *Gloucester County*, Union County, Essex County, New Jersey, Passaic County, Ocean County, Bucks County
The 40-Year-Old Virgin `nominated for`	*MTV Movie Award for Best Comedic Performance*, *BFCA Critics' Choice Award for Best Comedy*, *MTV Movie Award for Best On-Screen Duo*, MTV Movie Award for Best Breakthrough Performance, **MTV Movie Award for Best Movie**, MTV Movie Award for Best Kiss, D. F. Zanuck Producer of the Year Award in Theatrical Motion Pictures, Screen Actors Guild Award for Best Actor - Motion Picture
Costa Rica football team `has position`	*Forward*, *Defender*, *Midfielder*, **Goalkeepers**, Pitchers, Infielder, Outfielder, Center, Defenseman
Lil Wayne `born in`	**New Orleans**, Atlanta, Austin, St. Louis, Toronto, New York City, Wellington, Dallas, Puerto Rico
WALL-E `has the genre`	Animations, Computer Animation, *Comedy film*, *Adventure film*, *Science Fiction*, **Fantasy**, Stop motion, *Satire*, Drama

结论

给定一个头部和一个标签，排在最高位的尾部被预测出来。

不同模型在不同样本数量下的性能

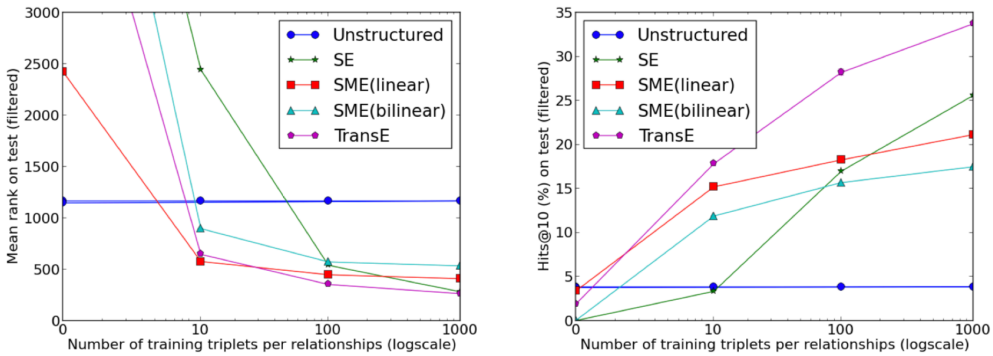

Figure 1: **Learning new relationships with few examples**. Comparative experiments on FB15k data evaluated in mean rank (left) and hits@10 (right). More details in the text.

- 左图表示测试集中平均排名：当训练集越大，TransE的平均排名下降的最快。
- 右图表示hits@10中正确的比例：当训练集越大，hits@10占比上升的最快。
- 结果表明TransE对样本预测的性能最优。

结论

TransE模型可以使用最小的参数量得到知识图谱的实体和关系向量表示。

TransE模型的参数较少，计算的复杂度显著降低，并且在大规模稀疏知识库上也同样具有较好的性能与可扩展性。

不足和改进

不足

- 在处理复杂关系`[1-N,N-1,N-N]`时，性能显著下降，比较适合处理`1-1`的关系。
- 不能够很好的处理更复杂的知识网络。

改进

- TransH模型：为了解决TransE模型在处理一对多 、 多对一 、多对多复杂关系时的局限性，TransH模型提出让一个实体在不同的关系下拥有不同的表示。
- TransR模型：一个实体是多种属性的综合体，不同关系关注实体的不同属性。不同的关系拥有不同的语义空间。
- TransD模型：给定三元组(h, r, t) , TransD模型设置了2个分别将头实体和尾实体投影到关系空间的投影矩阵。
- TranSparse模型：TranSparse是通过在投影矩阵上强化稀疏性来简化TransR的工作。通过引入稀疏投影矩阵，TransSparse模型减少了参数个数。
- TransM模型：除了允许实体在涉及不同关系时具有不同的嵌入之外，提高TransE模型性能可以从降低h+r≈t的要求研究开始。TransM模型将为每个事实（h,r,t）分配特定的关系权重theta_r。
- TransF模型：TransF只需要t与h+r位于同一个方向，同时h与t-r也位于同一个方向。
- ManifoldE模型：ManifoldE模型对于每个事实三元组 (h, r, t) 将$h + r \approx t$转换为(h+r-t)的L2范式约等于theta_r的平方。
- TransA模型：TransA模型为每个关系r引入一个对称的非负矩阵Mr，并使用自适应马氏距离定义评分函数。通过学习距离度量Mr, TransA在处理复杂关系时更加灵活。

思考

1. **Mean Rank 和 hit@10**

在测试过程中，对于一个三元组，我们将头实体或尾实体替换成任意一种其他的实体，得到（n-1）个新的关系三元组，然后对这些三元组计算实体关系距离，将这n-1个三元组按照距离从小到大排列。

- 对Mean Rank的理解

  在测试集里，求真实的实体在n-1个元素中的排名，得出平均到第多少个才能匹配到正确的结果。

- 对hit@10的理解

  在这个排好序的n-1元素中，从第一个开始遍历，看从第一个到第十个是否能够遇到真实的实体，如果遇到了就将 $hit@10+1$。