翻译嵌入建模多关系数据

Antoine Bordes, Nicolas Usunier, Alberto
Garcia-Dura'n Universite' de Technologie de
Compie`gne - CNRS Heudiasyc UMR 7253
贡比涅，法国
{bordesan, nusunier, agarciad}@utc.fr

Jason Weston, Oksana Yakhnenko
Google
111第八大道
纽约，纽约州，美国
{jweston, oksana}@google.com

抽象

我们考虑在低维向量空间中嵌入多关系数据的实体和关系的问题。 我们的目标是提出一个容易训练的规范模型，包含数量减少的参数，可以扩展到非常大的数据库。 因此，我们提出TransE，一种将关系解释为对实体的低维嵌入进行翻译的模型的方法。 尽管简单，但这个假设证明是有力的，因为广泛的实验表明，TransE在两个知识库的链接预测方面明显优于最先进的方法。 此外，还可以成功训练1M个实体，25k个关系和1700多个训练样本的大规模数据集。

1 介绍

多关系数据是指有向图，其节点对应于形式（头，标签，所有）（表示为（h，f，t））的实体和边，其中每一个指示存在名称标签之间的关系实体头部和尾部。 多关系数据模型在许多领域发挥着举足轻重的作用。 示例是社交网络分析，其中实体是成员和边缘（关系）是友谊/社交关系链接，推荐系统，其中实体是用户，产品和关系是购买，评级，审查或搜索产品或知识库（KB）如Freebase[1]，Google知识图[2]或GeneOntology[3]，其中KB的每个实体代表一个抽象的概念或世界的具体实体，关系是代表涉及其中两个事实的预言。 我们的工作重点是建模来自KB的多关系数据（Wordnet [9] 和Freebase [1] 在本文中），目的是提供一个有效的工具，通过自动添加新的事实来完成它们，而不需要额外的知识。

建模多关系数据一般来说，建模过程归结为提取实体之间的局部或全局连接模式，并通过使用这些模式来推断特定实体与所有其他实体之间观察到的关系。 单一关系的地方性概念可能纯粹是结构性的，比如朋友的朋友是我的朋友

[1] freebase.com
[2] google.com/insidesearch/features/search/knowledge.html [3] geneontology.org

社交网络，但也可以依靠实体，比如那些喜欢“星球大战IV”的人也喜欢“星球大战5”，但是他们可能会或者可能不喜欢泰坦尼克号。 与单一关系数据相比，在对数据进行一些描述性分析之后，可以进行特定但简单的建模假设，而关系数据的难点在于，局部性的概念可能同时涉及不同类型的关系和实体，因此建模多关系数据需要更多的通用方法，可以同时考虑所有异构关系来选择适当的模式。

在协同过滤中用户/项目聚类或矩阵分解技术成功地表示单关系数据中实体连通性模式之间的非平凡相似性之后，在关系学习框架内设计了多关系数据的mot现有方法从潜在的属性，如指出的[6]; 即通过学习和操作成分（实体和关系）的潜在表征（或嵌入）。 从这些方法的自然扩展到多关系域，例如随机块模型的非参数贝叶斯扩展[7, 10, 17] 和基于张量分解的模型 [5] 或集体矩阵分解 [13, 11, 12], 许多最近的方法都集中在增加贝叶斯聚类框架中模型的表达性和普遍性 [15] 或基于能量的框架来学习实体在低维空间的嵌入 [3, 15, 2, 14]. 这些模型的更大的表达性是以牺牲模型复杂度的大幅度增加为代价的，导致模型假设难以解释，并且计算成本较高。 此外，由于这类高容量模型的适当正则化难以设计，或者由于需要解决许多局部最小值问题而需要解决的非凸优化问题来训练，所以这种方法可能会受到过拟合的影响。 事实上，它显示在 [2] 一个更简单的模型（线性而不是双线性）在几个具有相对大量不同关系的多关系数据集上获得与最具表现力的模型几乎一样好的性能。 这表明，即使在复杂和异构的多关系域中，简单而恰当的建模假设也可能导致准确性和可伸缩性之间更好的平衡。

在嵌入空间中作为翻译的关系在本文中，我们介绍TransE，这是一个用于学习实体低维嵌入的基于能量的模型。 在TransE中，关系表示为嵌入空间中的翻译：如果（h，f，t）成立，则尾部实体t的嵌入应该接近头部实体h的嵌入加上一些取决于关系的向量F。 我们的方法依赖于减少的一组参数，因为它只为每个实体和每个关系学习一个低维向量。

我们基于翻译的参数化背后的主要动机是层次关系在KB中非常常见，翻译是表示它们的自然转换。 事实上，考虑到树的自然表示（即维度2中的节点的嵌入），兄弟彼此靠近并且给定高度处的节点被组织在x轴上，父子关系对应于y轴。 由于空翻译矢量对应于实体之间的等价关系，所以模型也可以表示兄弟关系。 因此，我们选择使用我们每个关系的参数预算（一个低维向量）来表示我们认为是KB中的关键关系。 另一个次要的动机来自最近的工作[8], 作者从自由文本中学习单词嵌入，以及不同类型的实体之间的一对一关系，比如国家和城市之间的“资本”，是（巧合而不是自愿地）用该模型表示为翻译嵌入空间。 这表明可能存在嵌入空间，其中不同类型的实体之间的1对1关系也可以由翻译表示。 我们的模型的意图是强化这种嵌入空间的结构。

我们在部分的实验4 证明这个新模型，尽管其简单性和其主要为建模层次而设计的架构，最终在大多数类型的关系上都是强大的，并且可以显着超越现实世界知识库中链接预测的最新方法。 此外，其轻量化参数允许它在包含1M实体，25k关系和17M以上训练样本的Freebase的大规模分裂中成功训练。

在本文的其余部分，我们将在Section中描述我们的模型2 并在第一节讨论其与相关方法的联系 3. 我们详细介绍了一个关于Wordnet和Freebase的广泛的实验性研究4, 将TransE与文献中的许多方法进行比较。 最后我们通过勾画出一些未来的工作方向 5.

算法1学习TransE

输入训练集S = { (h, f, t) }，实体和rel。 设置E和L，余量 γ，嵌入暗淡。 k。

1: 对每个f∈L初始化£←uniform ($-\sqrt{6}/k$, $\sqrt{6}/k$)
2: 对于每个f∈L，£←££l £l
3: e←均匀 ($-\sqrt{6}/k$, $\sqrt{6}/k$) 为每个实体e∈E
4: 循环
5: 对于每个实体e∈E，e←e / 1 e 1
6: Sbatch←样品 (S, b) //样品尺寸为b的小批次
7: Tbatch←∅//初始化三元组对
8: *为 (h, f, t) ∈Sbatch做*
9: (h^I, f, t^I) ←样本 ($S_{(1H, F, T)}$) //对损坏的三元组
10: $T_{batch} \leftarrow T_{batch} \cup$ H (h, f, t), (h^I, f, t^I)
11: 结束
12: 更新embeddingsw.rt $\nabla \gamma + d(\boldsymbol{h} + £, \boldsymbol{t}) - d(\boldsymbol{h} + £, \boldsymbol{t})_+$
H (h, f, t), $(h^I, f,$
13: 结束循

(h, f, t), $(h^I, f,$
13: 结束循

2 基于翻译的模型

给定由两个实体h, t E (实体集合) 和关系f L (关系集合) 组成的三元组 (h, f, t) 的训练集S，我们的模型学习实体的矢量嵌入和关系。 嵌入取R^k中的值 (k是模型超参数)，用粗体字表示相同的字母。 我们模型背后的基本思想是由f-标记边缘引起的函数关系对应于嵌入的平移，即我们希望当 (h, f, t) 成立时 (t应该是最近的邻居h +£)，而h +£应该远离t否则。 在一个基于能量的框架之后，一个三元组的能量等于d (h +£, t)，对于一些相异度量d，我们认为它是L1或者L2范数。

为了学习这种嵌入，我们在训练集上最小化了基于边界的排序标准:

$$L = \sum_{(h, f, t) \in S} \sum_{(h^I, f, t^I) \in S^{小时}(h, f, t)的} \gamma + d(\boldsymbol{h} + £, \boldsymbol{t}) - d(\boldsymbol{h} + £, \boldsymbol{t})_+ \quad (1)$$

其中[x] +表示x的正部分，$\gamma > 0$是余量超参数

$$S_{(1H, F, T)} = (h^I, f, t) \mid h^I \in E \cup (h, f, t^I) \mid t^I \in E。 \quad (2)$$

根据等式构造的一组损坏的三元组2, 是由训练三胞胎组成的，头部或尾部被一个随机的实体取代 (但不是同时)。 亏损功能 (1) 有利于训练三胞胎的能量的值低于被破坏的三胞胎，因此是预期标准的自然实现。 注意，对于一个给定的实体，当实体出现在头部或者三元组的尾部时，其嵌入向量是相同的。

通过随机梯度下降 (在小批量模式下)，在可能的h，£和t上进行优化，并且附加的约束条件是实体嵌入的L2范数为1 (不给定正则化或范数约束标签嵌入£)。 这个约束对我们的模型是重要的，因为它是以前的基于嵌入的方法[3, 6, 2], 因为它阻止了
训练过程通过人为地增加实体嵌入规范来使L最小化。

算法中描述了详细的优化过程1. 所有对实体和关系的嵌入首先按照中提出的随机过程进行初始化[4]. 在算法的每个主要迭代中，实体的嵌入向量首先被归一化。 然后，从训练集中采样一小组三联体，并将作为小批量的训练三联体。 对于每个这样的三元组，我们然后采样一个单一的损坏的三元组。 然后通过采用具有恒定学习速率的梯度步骤来更新参数。 该算法基于验证集上的性能停止。

3 相关工作

部分1 描述了嵌入KB的大量工作。 我们在这里详细介绍我们的模型和那些模型之间的联系[3] (结构化嵌入或SE) 和 [14].

表1：FB15k的参数数量及其值（以百万为单位）。 ne和nr是nb。 实体和关系； k嵌入维度。

方法	NB。 参数	在FB15K上
非结构化[2]	0 (NEK)	0.75
RESCAL [11]	$O(nek + nrk^2)$	87.80
SE [3]	$O(nek + 2nrk^2)$	7.47
中小企业（线性）[2]	$O(nek + nrk + 4k^2)$	0.82
中小型企业（BILINEAR）[2]	$O(nek + nrk + 2k^3)$	1.06
LFM [6]	$O(nek + nrk + 10k^2)$	0.84

表2：本文中使用的数据集的统计，并从两个知识库，Wordnet和Freebase中提取。

数据集	WN	FB15K	FB1M
实体	40,943	14,951	1×10^6
关系	18	1,345	23,382
培养。 EX。 有效的	141,442	483,142	17.5×10^6
	5,000	50,000	50,000
TEST EX。	5,000	59,071	177,404

SE [3] 将实体嵌入到R^k中，并且将关系嵌入到两个矩阵$L1R^{k\times k}$和$L2R^{k\times k}$中，使得d（L1h，L2t）对于损坏的三元组（h，f，t）很大（否则小）。 其基本思想是当两个实体属于同一个三元组时，他们的嵌入应该在一些依赖于关系的子空间中彼此靠近。 用头部和尾部两个不同的投影矩阵来解释关系f的可能的不对称性。 当相异函数的形式为d（x，y）= g（xy）时，对于某个g：R^kR（例如g是一个范数），则具有k + 1大小嵌入的SE严格地更具表现力因为k + 1维上的线性算子可以在k维的子空间（通过约束所有嵌入的第k + 1维等于1）再现仿射变换。 SE，以L2作为单位矩阵，并且为了重现翻译而采用的L1则相当于TransE。 尽管我们模型的表现力较低，但是在我们的实验中，我们的性能仍然比SE好。 我们认为这是因为（1）我们的模型是一种更直接的方式来表示关系的真实属性，（2）在嵌入模型中优化是困难的。 对于SE来说，更大的表现力似乎更多的是不足之处，而不是更好的表现。 训练错误（在Section 4.3）倾向于证实这一点。

另一个相关的方法是神经张量模型[14]. 这个模型的一个特殊情况对应于以下形式的学习成绩s（h，f，t）（损坏的三胞胎的较低分数）：

$$s(h, f, t) = h^T_1 Lt + £^T_2 h + £^T t \qquad (3)$$

其中$L\in R^{k\times k}$，$L1\in R^k$和$L2\in R^k$，都依赖于f。

如果我们将平方欧氏距离视为相异函数，我们有：

$$d(h + £, t) = \tfrac{2}{2} h\ 1 + \tfrac{2}{2} 1\ 1 + \tfrac{2}{2} t\ 1 - 2^{T} ht^{H} + £(t-h)。$$

考虑到我们的范数约束（1 h 1^2= 1 t 2 1^2= 1 2）和排序标准 (1)，其中1 £1^2 2 在比较损坏的三元组中没有起到任何作用，因此我们的模型涉及用$h^Tt + ^T$（th）对三元组进行评分，因此对应于 [14]（方程 (3)）其中L是单位矩阵，£= £1 = £2。 我们不能用这个模型来运行实验（因为它已经和我们一样同时发表了），但是TransE的参数又少得多了：这样可以简化训练并防止不合适，并且可以弥补较低的表现力。

尽管如此，TransE的简单表述可以被看作编码一系列双向交互（例如通过开发L2版本），但是也存在缺陷。 对于h，f和t之间的3way依赖性至关重要的建模数据，我们的模型可能会失败。 例如，在小规模的亲属数据集上[7]，TransE在交叉验证（在精度 - 回忆曲线下的面积测量）方面没有达到与最先进的技术相竞争的性能[11, 6]，因为这种三元互动在这种情况下是至关重要的 [2]). 不过，我们的部分的实验4 为了处理像Freebase这样的通用大型KB，我们首先应该正确模拟最常见的连接模式，就像TransE那样。

4 实验

我们的方法TransE是从Wordnet和Freebase提取的数据进行评估的（他们的统计数据在表格中给出2），反对文献中最近的几种方法，这些方法被证明可以在各种基准测试中达到最佳性能，并且可以扩展到相对较大的数据集。

4.1 数据集

Wordnet该KB旨在生成直观可用的词典和辞典，并支持自动文本分析。 它的实体（称为同义词）对应于单词的意义，而关系定义它们之间的词汇关系。 我们考虑过使用的数据版本[2],我们在下面表示WN。 三元组的例子是（分数NN 1，上位词，评价NN 1）或（分数NN 2，具有部分音乐符号NN 1）。[4]

Freebase Freebase是一个巨大且日益增长的一般事实KB；目前有三十二亿三胞胎和八千多万个实体。 我们用Freebase创建了两个数据集。 首先，为了做一个小的数据集，我们选择了在Wikilinks数据库中也存在的实体子集[5]而且在Freebase中至少有100个提及（对于实体和关系）。 与“人/人/国籍”关系相比，我们还消除了“人/人/人/国家”这样的关系。这导致了592,213三胞胎与14,951实体和1,345关系，如表所示随机分裂2. 这个数据集在本节的其余部分用FB15k表示。 我们也想要有大规模的数据来测试TransE的规模。 因此，我们通过选择最频繁出现的100万个实体，从Freebase创建另一个数据集。 这导致了25K关系的分裂和1700多万的训练三胞胎，我们称之为FB1M。

4.2 实验装置

评估协议为了评估，我们使用与在中相同的排名程序[3].对于每个测试三元组，头部被删除并依次被字典中的每个实体替换。 那些被破坏的三联体的不同（或能量）首先由模型计算，然后按升序排序；最终存储正确实体的等级。 整个过程是重复的，同时去除尾巴而不是头部。 我们报告那些预测的排名和命中@ 10的平均值，即排名前十的正确实体的比例。

这些指标是指示性的，但是当一些损坏的三元组最终成为有效的三元组时，可能是有缺陷的，例如来自训练集。 在这种情况下，那些可能被排在测试三元组之上，但是这不应该被视为错误，因为三元组都是真的。 为了避免这种误导行为，我们建议从损坏的三元组列表中删除出现在训练集，验证集或测试集中的所有三元组（除了测试三元组之外）。 这确保了所有损坏的三元组不属于数据集。 在下面，我们根据两种设置报告平均等级和匹配：原始的（可能有缺陷的）称为原始的，而我们将新的称为过滤的（或者过滤的）。 我们只提供FB1M实验的原始结果。

基线第一种方法是非结构化，TransE的一个版本，它将数据视为单关系，并将所有翻译设置为0（它已被用作基线[2]）。 我们还与RESCAL（集成矩阵分解模型）进行比较[11,12],和能源型SE[3],SME（线性）/ SME（双线性）[2]和LFM[6].RESCAL是通过交替最小二乘法进行训练的，而其他则是通过随机梯度下降进行训练，如TransE。 表1将基线的理论参数数量与我们的模型进行比较，并给出FB15k的数量级。 对于低维嵌入，SME（线性），SME（双线性），LFM和TransE具有与非结构参数相同数量的参数，其他算法SE和RESCAL对于每个关系至少学习一个kk矩阵需要快速学习参数。 RESCAL在FB15k上需要大约87倍的参数，因为它需要比其他模型大得多的嵌入空间来实现良好的性能。× 由于可扩展性的原因，我们没有在RESCAL，SME（双线性）和LFM上对FB1M进行实验，参数数量或训练持续时间。

我们使用作者提供的代码来训练所有的基线方法。 对于RESCAL，由于可扩展性原因，我们必须将正则化参数设置为0，如其中所示[11],并选择50,250,500,1000,2000中的潜在维数k，从而导致验证组（使用原始设置）上的最低平均预测等级。 对于非结构化，SE，SME（线性）和SME（双线性），我们

[4]WN由语义组成，其实体由一个单词，它的词性标记和一个数字表示它指的是哪一个意义，即得分NN1编码名词“score”的第一个含义，。

[5]code.google.com/p/wiki-links

表3：链接预测结果。 测试不同方法的性能。

DATASET	WN				FB15K				FB1M	
公制	平均排名		HITS @ 10 (%)		平均排名		HITS @ 10 (%)		平均排名	HITS @ 10 (%)
EVAL。 设置	*生的*	*FILT。*	*生的*	*FILT。*	*生的*	*FILT。*	*生的*	*FILT。*	*生的*	*生的*
非结构化[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
中小企业（线性）[2]中	545	533	65.1	74.1	274	154	30.7	40.8	-	-
小型企业（BILINEAR）	526	509	54.7	61.3	284	158	31.3	41.3	-	-
[2] 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**

在20,50之间选择0.001,0.01,0.1，k之间的学习率，并通过使用验证集上的平均等级（在训练数据中总共至多1000个历元）提前停止来选择最佳模型。 对于LFM，我们也使用了平均验证等级来选择模型和
选择{25，50，75}中的潜在维度，{50,100,200,500}
学习率在{0.01,0.1,0.5}之间。

实现对于使用TransE的实验，我们选择了0.001,0.01,0.1中的随机梯度下降的学习率λ，1,2,10中的余量γ和20{，50中的潜在维度k在每个数据集的验证集上}。 根据验证性能，将相异度量d设为L1或L2距离。 最佳配置是：在Wordnet上的k = 20，λ = 0.01，γ = 2和d = L1；在FB15k上k = 50，λ = 0.01，γ = 1，d = L1；在FB1M上k = 50，λ = 0.01，γ = 1和d = L2。 对于所有的数据集，训练集的训练时间被限制在最多1,000个时期。 最好的模型是通过使用验证集上的平均预测等级（原始设置）提早停止来选择的。 项目网页提供TransE的开源实现[6].

4.3 链接预测

总体结果表3 显示所有比较方法的所有数据集上的结果。 正如预期的那样，过滤后的设置提供较低的平均等级和较高的命中@ 10，我们认为这是对链接预测中方法性能的更清晰的评估。 但是，原始和过滤之间的趋势通常是相同的。

我们的方法TransE在所有指标上都优于所有指标，并且通常具有很高的利润率，并且达到了一些有前途的绝对性能指标，如WN上超过10万（超过4万个实体）的89%，FB1M上的34%（超过1M实体）。 TransE和亚军之间的所有区别都很重要。

我们认为，TransE的良好表现是根据数据对模型进行了适当的设计，也因为其相对简单。这意味着它可以随机梯度有效地进行优化。 我们在节中展示3 SE比我们的建议更有表现力。然而，它的复杂性可能会使它很难学习，导致性能下降。 在FB15k上，SE达到165的平均等级，在训练集的50k三联体子集中达到35.5%的10分，而TransE达到127和42.7%，表明TransE确实不太适应不足，并且这可以解释其更好的表现。 中小企业（双线性）和线性调频（LFM）遭受同样的培训问题：我们从来没有设法好好培训他们，以便他们能够充分利用他们的能力。 LFM的不良结果也可能由我们的评估设置来解释，基于排名实体，而LFM最初是为了预测关系而提出的。 RESCAL在FB15k上可以达到相当不错的10点，但是平均等级差，特别是在WN上，甚至当我们使用大的潜在维度（Wordnet上的2000）时。

翻译术语的影响是巨大的。 当比较TransE和非结构化（即没有翻译的TransE）的性能时，非结构化的平均等级看起来相当好（WN上最好的亚军），但是命中@ 10是非常差的。 非结构化简单地将所有实体共同出现在一起，而不依赖于所涉及的关系，因此只能猜测哪些实体是相关的。 在FB1M上，TransE和Unstructured的平均排名几乎相似，但是TransE排名前十的预测值是前者的10倍。

[6]可在http://goo.gl/0PpKQe.

表4：关系类别的详细结果。 在我们的模型的过滤评估设置TransE和基线上，我们比较FB15k中的命中@ 10（以%计）。 （M.代表MANY）。

任务	预测头				预测尾巴			
REL。 类别	1-TO-1	1-TO-M。	M.-TO-1	M.-TO-M。	1-TO-1	1-TO-M。	M.-TO-1	M.-TO-M。
非结构化[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
中小企业（线性）[2]	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
中小型企业（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**

表5：使用TransE对FB15k测试集进行的示例预测。 粗体表示测试三联体的真实尾巴，斜体表示训练集中的其他真实尾巴。

输入（头和标签）	预测的尾巴
JK罗琳影响	*GK Chesterton，JRR托尔金，CS刘易斯，劳埃德●亚历山大，* Terry Pratchett，Roald Dahl，Jorge Luis Borges，Stephen King，Ian Fleming
安东尼LaPaglia执行	*马缨丹，山姆的夏天，快乐的脚，欢乐的房子，* 不忠实的，守护者传奇，裸体午餐，X战警，同名
卡姆登县毗邻	*伯灵顿县，大西洋省，格洛斯特县，联合县，* 埃塞克斯县，新泽西州，帕塞克县，海洋县，雄鹿县
40岁的维尔京提名	*最佳喜剧表演MTV电影奖，* *最佳喜剧"BFCA评论家选择奖"，最佳屏幕二重奏MTV电影奖，* 最佳突破性表演MTV电影奖， 最佳电影MTV电影奖，最佳吻MTV电影奖， DF Zanuck电影制片人，电影演员协会奖最佳男主角 - 电影
哥斯达黎加橄榄球队有位置	*向前，后卫，中场，守门员，* 投手，内野手，外野手，中锋，防守队员
Lil Wayne出生在	新奥尔良，亚特兰大，奥斯汀，圣路易斯， 多伦多，纽约市，惠灵顿，达拉斯，波多黎各
墙-E有这个流派	动画，电脑动画，喜剧电影， *冒险电影，科幻，幻想，定格，讽刺，戏剧*

详细的结果表4 根据几个关系类别和参数预测几种方法，对FB15k中的结果进行分类（命中@ 10）。 我们根据头尾参数的基数将这些关系分为四类：1对1，1对多，多对1，多对多。如果头部最多可以出现一条尾巴，那么给定的关系是1比1;如果头部可以出现多条尾巴，则比例为1比1;如果多个头部可以出现相同的尾巴，则比例为1比1;或者如果有多个头可以出现多条尾巴，则可以多对多。 我们通过计算每个关系f，计算FB15k数据集中出现的平均头数h（尊重尾数t），给出一对（f，t）（尊重a对（h， F））。 如果这个平均数低于1.5，那么这个参数被标记为1，否则就是很多。 例如，具有平均每尾1.2头和每头3.2尾的关系被归类为一对多。 我们得到FB15k的1对1关系占26.2%，1对多关系占22.7%，MANY-TO-1占28.3%，MANY-TO-MANY占22.8%。

这些详细的结果在表中4 允许对方法的行为进行精确的评估和理解。 首先，正如人们所期望的那样，预测三联体的**"第一方"**（即，预测1-TO-MANY的头部和MANY-TO-1的尾部）的实体更容易，也就是说，实体指向它。 这些是适当的情况。 中小企业（双线性）在这种情况下被证明是非常准确的，因为它们是那些训练案例最多的。 非结构化在1对1关系上表现良好：这表明这种关系的参数必须共享隐藏类型，非结构化能够通过在嵌入空间中链接在一起的聚类实体来稍微发现。 但是这个策略在任何其他类型的关系中都失败了。 添加翻译术语（即，将非结构化升级到TransE）能够通过遵循关系在嵌入空间中从一个实体簇移动到另一个实体簇。 对于适当的情况来说，这一点尤为壮观。

插图表5 给出了FB15k测试集（预测尾部）上TransE链路预测结果的例子。 这说明了我们模型的功能。 给一个头和一个标签，顶部

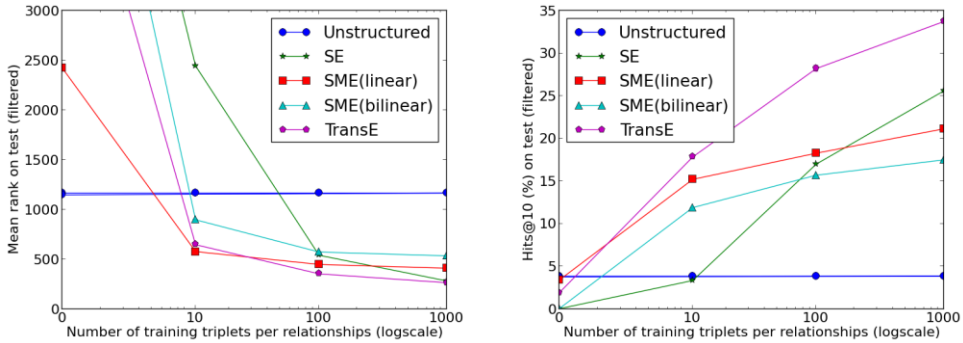

图1：通过几个例子学习新的关系。 对FB15k数据的比较实验以平均等级（左）和命中@ 10（右）评估。 文中更多细节。

预测的尾巴（和真实的尾巴）被描绘。 这些例子来自FB15k测试装置。 即使好的答案并不总是排在前列，预测反映的是常识。

4.4 学习用少数例子来预测新的关系

使用FB15k，我们想通过检查他们学习新关系的速度来测试方法如何能够推广到新的事实。为此，我们随机选择了40个关系，并将数据分为两组：一组（包含所有三元组，包含这40个关系的集合（名为FB15k-40rel），另一组包含其余集合（FB15k-rest）。 我们确保两个集合都包含所有的实体。 然后将FB15k-rest分成353,788三胞胎的训练集和53266和FB15k-40rel的训练集，成为40,000三胞胎训练集（每种关系1,000）和45,159的测试集。使用这些数据集，我们进行了以下实验：（1）使用FB15k-剩余训练和验证集合训练和选择模型，（2）随后在训练集合FB15k-40rel上训练它们，但仅仅学习与（3）在FB15k-40rel测试集的链路预测中（仅包含阶段（1）期间看不到的关系）评估它们。 我们在阶段（2）中使用0,10,100和1000个每个关系的例子重复这个过程。

图中给出了非结构化，SE，SME（线性），SME（双线性）和TransE的结果1. 当未提供未知关系的例子时，非结构化的性能是最好的，因为它不使用这个信息来预测。 但是，当然，这个性能并没有提高，同时提供了标记的例子。 TransE是学习速度最快的方法：只有10个新关系的例子，@ 10的命中率已经是18%，并且随着提供的样本数目单调增加。 我们相信TransE模型的简单性使其能够很好地概括，而不必修改任何已经过训练的嵌入。

5 结论和未来的工作

我们提出了一种学习知识库嵌入的新方法，着重于模型的最小参数化，主要表示层次关系。我们发现它与两种不同的知识库上的竞争方法相比效果很好，同时也是一个高度可扩展的模型，我们将它应用于一个非常大规模的Freebase数据块。 尽管我们仍不清楚所有关系类型是否可以通过我们的方法进行充分建模，但将评估分为几类（1对1，1对多，...），看起来效果不错其他方法跨所有设置。

未来的工作可以进一步分析这个模型，并集中在更多的任务，特别是应用，如学习单词表示的启发[8]. 将KB与文本组合在一起[2] 是我们的方法证明有用的另一个重要方向。 因此，我们最近巧妙地将TransE插入到从文本中提取关系的框架中 [16].

致谢

这项工作是在Labex MS2T（ANR-11-IDEX-0004-02）框架内进行的，由法国国家研究机构（EVEREST-12-JS02-005-01）资助。 我们感谢X.Glorot提供代码基础设施，T.Strohmann和K.Murphy进行有益的讨论。

参考

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge和J. Taylor。 Freebase：一个用于构建人类知识的协作创建的图形数据库。 在2008年ACM SIGMOD数据管理国际会议论文集中。

[2] A. Bordes, X. Glorot, J. Weston和Y. Bengio。 用于多关系数据学习的语义匹配能量函数。 机器学习，2013。

[3] A. Bordes, J. Weston, R. Collobert和Y. Bengio。 学习知识库的结构化嵌入。 2011年第25届人工智能年会（AAAI）会议录。

[4] X. Glorot和Y. Bengio。 了解训练深度前馈神经网络的难度。 在国际人工智能和统计学会议（AISTATS），2010年。

[5] RA Harshman和ME Lundy。 Parafac：平行因子分析。 计算统计和数据分析，18（1）：39-72，1994年8月。

[6] R. Jenatton, N. Le Roux, A. Bordes, G. Obozinski等人 高度多关系数据的潜在因子模型。 神经信息处理系统进展（NIPS 25），2012。

[7] C. Kemp, JB Tenenbaum, TL Griffiths, T. Yamada和N. Ueda。 用无限关系模型学习概念系统。 在第二十一届人工智能年会（AAAI）会议录中，2006。

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado和J. Dean。 单词和短语的分布式表示及其组合性。 神经信息处理系统进展（NIPS 26），2013。

[9] G. Miller。 WordNet：英语词汇数据库。 ACM通讯，38（11）：39-41,1995。

[10] K. Miller, T. Griffiths和M. Jordan。 用于链接预测的非参数潜在特征模型。 神经信息处理系统进展（NIPS 22），2009。

[11] M. Nickel, V. Tresp和H.-P. 克里格尔。 多关系数据集体学习的三维模型。 在第二十八届国际机器学习会议（ICML），2011年。

[12] M. Nickel, V. Tresp和H.-P. 克里格尔。 分解YAGO：可链接数据的可伸缩机器学习。 在第21届国际万维网（WWW）国际会议论文集，2012。

[13] AP Singh和GJ Gordon。 通过集体矩阵分解进行关系学习。 在第十四届ACM SIGKDD知识发现和数据挖掘会议（KDD）会议上，2008。

[14] R. Socher, D. Chen, CD Manning和AY Ng。 利用神经张量网络和语义词向量从知识库中学习新的事实。 神经信息处理系统进展（NIPS 26），2013。

[15] I. Sutskever, R. Salakhutdinov和J. Tenenbaum。 使用贝叶斯聚类张量分解对关系数据建模。 神经信息处理系统进展（NIPS 22），2009。

[16] J. Weston, A. Bordes, O. Yakhnenko和N. Usunier。 将语言和知识库与关系抽取的嵌入模型相结合。 在2013年自然语言处理经验方法会议（EMNLP）会议记录。

[17] J. Zhu。 用于链路预测的最大余量非参数潜在特征模型。 在2012年第29届国际机器学习会议（ICML）会议论文集中。