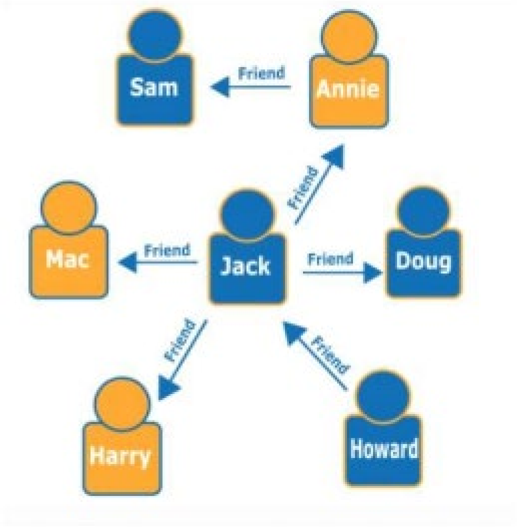
Translating Embeddings for Modeling Multi-relational Data (TransE)

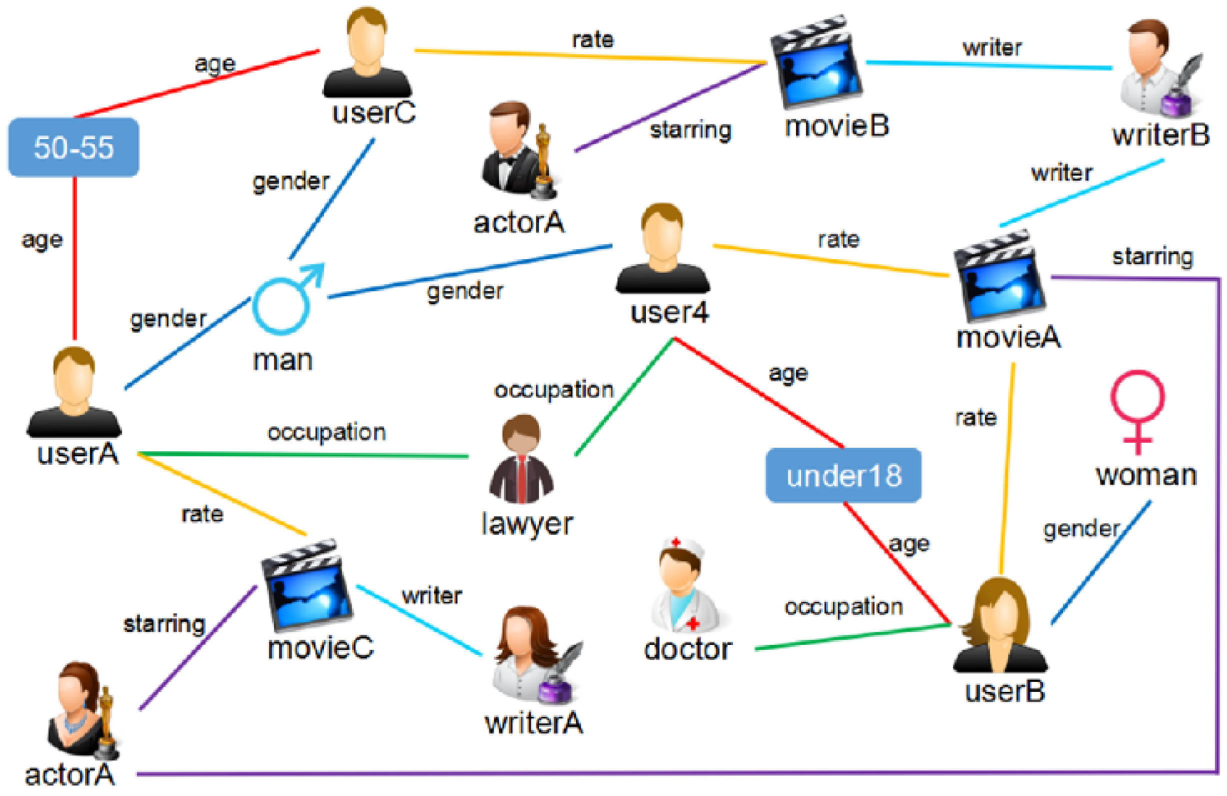
SoC 19, SeungAn Jung

Multi-Relational Data



Homogeneous relationship

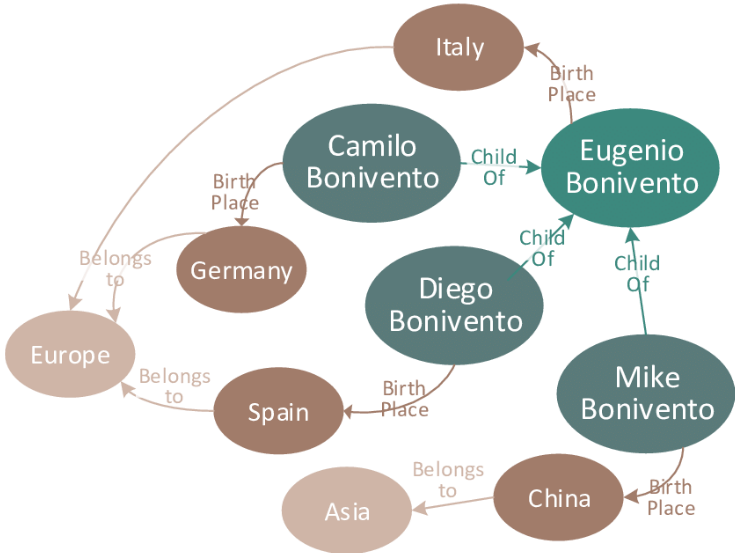
<cf.>



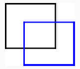
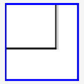
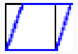
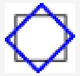
Heterogeneous relationships

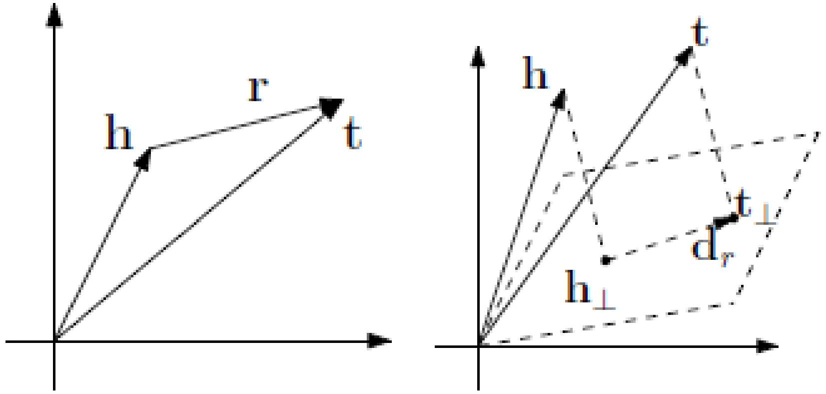
Knowledge Graph Embedding(1)

- A knowledge graph (KG) is a directed heterogeneous multigraph whose node and relation types have domain-specific semantics.
- Knowledge graph embedding is the task of completing the knowledge graphs by probabilistically inferring the missing arcs from the existing graph structure.



Translation

Affine Transform	Example	Transformation Matrix	
Translation		$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_x & t_y & 1 \end{bmatrix}$	t_x specifies the displacement along the x axis t_y specifies the displacement along the y axis.
Scale		$\begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$	s_x specifies the scale factor along the x axis s_y specifies the scale factor along the y axis.
Shear		$\begin{bmatrix} 1 & sh_y & 0 \\ sh_x & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	sh_x specifies the shear factor along the x axis sh_y specifies the shear factor along the y axis.
Rotation		$\begin{bmatrix} \cos(q) & \sin(q) & 0 \\ -\sin(q) & \cos(q) & 0 \\ 0 & 0 & 1 \end{bmatrix}$	q specifies the angle of rotation.



(a) TransE (b) TransH

Relationship Pattern

Definition 1. *A relation* $\mathtt{r}$ *is* ***symmetric (antisymmetric)*** *if* $\forall \mathtt{x}, \mathtt{y}$

$$\mathtt{r}(\mathtt{x}, \mathtt{y}) \Rightarrow \mathtt{r}(\mathtt{y}, \mathtt{x}) \;\; (\, \mathtt{r}(\mathtt{x}, \mathtt{y}) \Rightarrow \neg \mathtt{r}(\mathtt{y}, \mathtt{x}) \,)$$

A clause with such form is a ***symmetry (antisymmetry)*** *pattern.*

Definition 2. *Relation* $\mathtt{r}_1$ *is* ***inverse*** *to relation* $\mathtt{r}_2$ *if* $\forall \mathtt{x}, \mathtt{y}$

$$\mathtt{r}_2(\mathtt{x}, \mathtt{y}) \Rightarrow \mathtt{r}_1(\mathtt{y}, \mathtt{x})$$

A clause with such form is a ***inversion*** *pattern.*

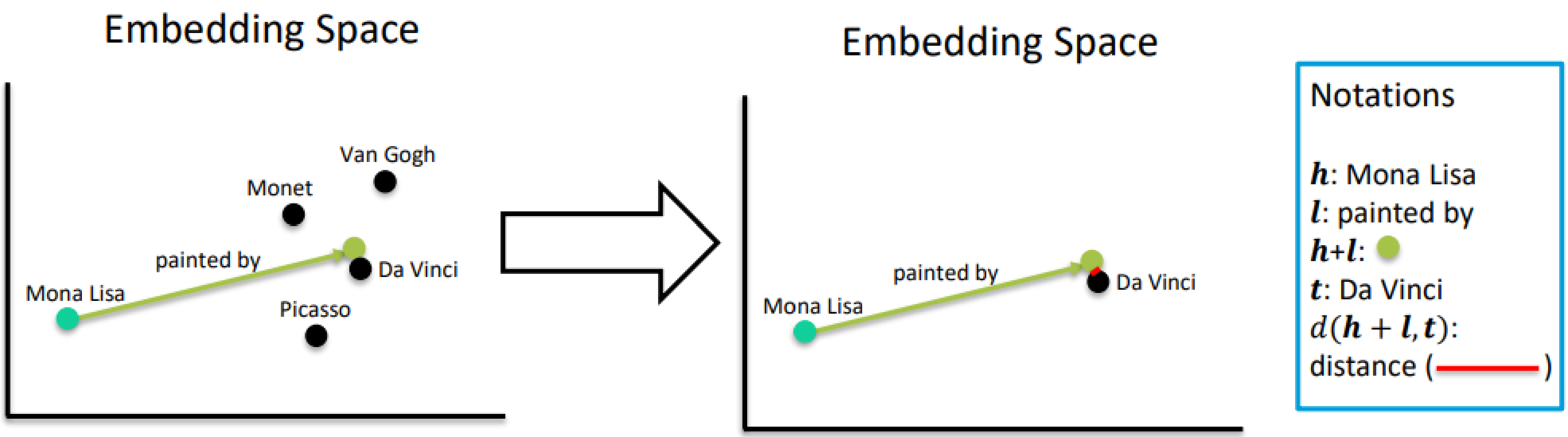
Definition 3. *Relation* $\mathtt{r}_1$ *is* ***composed*** *of relation* $\mathtt{r}_2$ *and relation* $\mathtt{r}_3$ *if* $\forall \mathtt{x}, \mathtt{y}, \mathtt{z}$

$$\mathtt{r}_2(\mathtt{x}, \mathtt{y}) \wedge \mathtt{r}_3(\mathtt{y}, \mathtt{z}) \Rightarrow \mathtt{r}_1(\mathtt{x}, \mathtt{z})$$

A clause with such form is a ***composition*** *pattern.*

TransE

$$< translation > \ h(entity) + r(relationship) \rightarrow t(entity)$$



TransE - Algorithm

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
    $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
    $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
loop
  $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
  $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b
  $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
  for $(h, \ell, t) \in S_{batch}$ **do**
    $(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet
    $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
  end for
  Update embeddings w.r.t. $\displaystyle\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$
end loop

$h + r \cong t \text{ when } (h, l, t) \text{ holds}$

d: Dissimilarity measure (L1 or L2 norm)

TransE - Algorithm

- Corrupted triplets: Replace either head or tail with random entity

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

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

- Loss: margin-based ranking criterion (triplet margin loss function)

- Lower dissimilarity for $d(h+l,t)$ (= valid triplet)
- Higher dissimilarity for $d(h'+l,t')$ (= corrupted triplet)

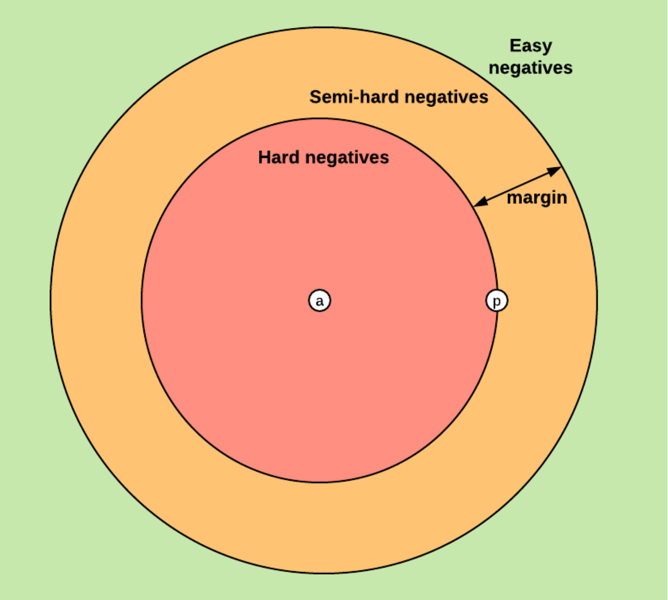
Algorithm 1 Learning TransE

input Training set $S = \{(h,\ell,t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
    $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
    $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$

loop
  $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
  $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b
  $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
  for $(h,\ell,t) \in S_{batch}$ **do**
    $(h',\ell,t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet
    $T_{batch} \leftarrow T_{batch} \cup \{((h,\ell,t),(h',\ell,t'))\}$
  end for
  Update embeddings w.r.t. $\sum_{((h,\ell,t),(h',\ell,t'))\in T_{batch}} \nabla\left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$

end loop



Related Work

- SE (Structured Embedding)
  - Embeds relationships into two matrices L1, L2
    - $d(\boldsymbol{L_1}h, \boldsymbol{L_2}t)$ is large for corrupted triplets
    - L1 reproduce translation, L2 = Identity Matrix => SE is same as TransE
  - Greater expressiveness than TransE
    - Synonymous to underfitting than to better performance

- Neural Tensor Model
  - TransE with squared Euclidean distance as dissimilarity
    - $l = l_1 = -l_2$ => same as TransE
    - TransE has much fewer parameter
      - Simplify training and prevent underfitting

$$s(h, \ell, t) = \boldsymbol{h}^T \boldsymbol{L} \boldsymbol{t} + \boldsymbol{\ell}_1^T \boldsymbol{h} + \boldsymbol{\ell}_2^T \boldsymbol{t}$$

$$d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) = \| \boldsymbol{h} \|_2^2 + \| \boldsymbol{\ell} \|_2^2 + \| \boldsymbol{t} \|_2^2 - 2\big(\boldsymbol{h}^T \boldsymbol{t} + \boldsymbol{\ell}^T (\boldsymbol{t} - \boldsymbol{h})\big) .$$

Experiment

- Data sets
  - Wordnet
    - Entity: meaning of word, Relationship: lexical relation
  - Freebase
    - KB of general facts
    - 15k, 1m (entity number in dataset)

Table 2: **Statistics of the data sets** used in this paper and extracted from the two knowledge bases, Wordnet and Freebase.

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

Experiment

- Evaluation Metric
    1. Mean Rank
    2. Hits@10 (%)

- Baselines
    1. Unstructured
       - consider data as single-relational, translation to 0
    2. RESCAL
       - collective matrix factorization model
    3. SE, SME(linear/bilinear)
    4. LFM
       - Energy-based models

METHOD	NB. OF PARAMETERS	ON FB15K
Unstructured [2]	$O(n_e k)$	0.75
RESCAL [11]	$O(n_e k + n_r k^2)$	87.80
SE [3]	$O(n_e k + 2n_r k^2)$	7.47
SME(LINEAR) [2]	$O(n_e k + n_r k + 4k^2)$	0.82
SME(BILINEAR) [2]	$O(n_e k + n_r k + 2k^3)$	1.06
LFM [6]	$O(n_e k + n_r k + 10k^2)$	0.84
TransE	$O(n_e k + n_r k)$	0.81

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**

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**

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

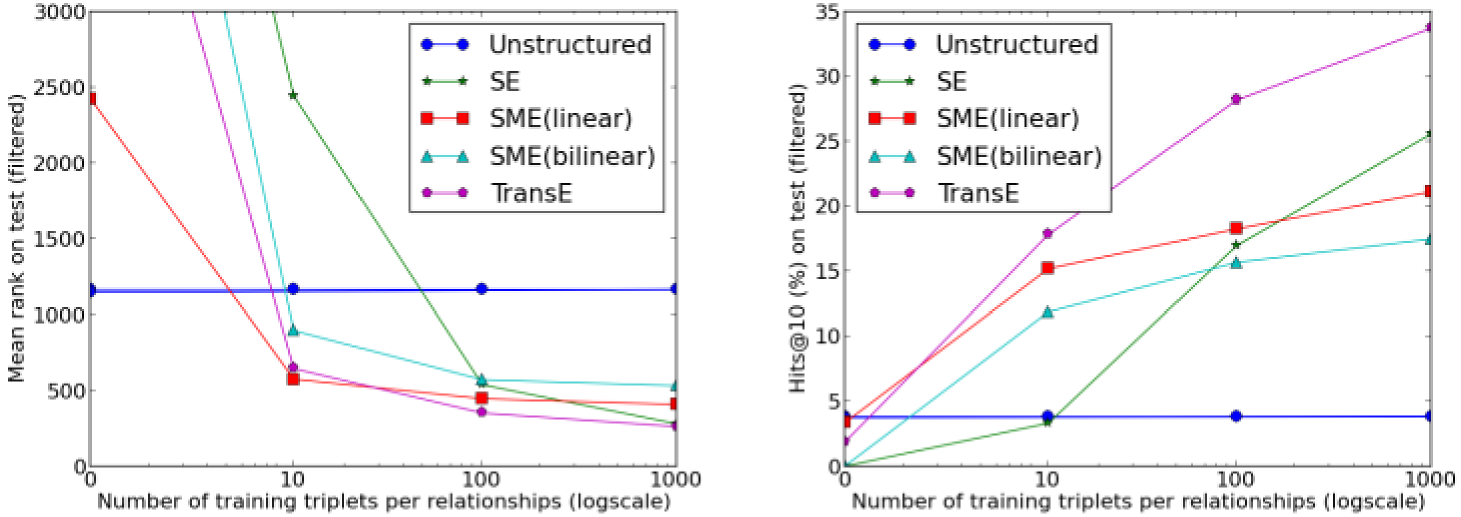


Figure 1: **Learning new relationships with few examples**. Comparative experiments on FB15k data evaluated in mean rank (left) and hits@10 (right). More details in the text.

TransE is the fastest method to learn

-> Simplicity of TransE makes it able to generalize well, without having to modify of the already trained embeddings.

Implementation

```
@torch.no_grad()
def test(
    self,
```

```
Epoch: 324, Loss: 0.0819
Epoch: 325, Loss: 0.0812
100%|████████████████████████████████████████| 17535/17535 [01:29<00:00, 194.92it/s]
(257.8840637207031, 0.3698317650413459)
Epoch: 326, Loss: 0.0809
Epoch: 327, Loss: 0.0822
Epoch: 328, Loss: 0.0821
Epoch: 329, Loss: 0.0818
Epoch: 330, Loss: 0.0825
Epoch: 331, Loss: 0.0822
Epoch: 332, Loss: 0.0820
Epoch: 333, Loss: 0.0816
Epoch: 334, Loss: 0.0821
Epoch: 335, Loss: 0.0815
Epoch: 336, Loss: 0.0810
Epoch: 337, Loss: 0.0822
Epoch: 338, Loss: 0.0817
Epoch: 339, Loss: 0.0828
Epoch: 340, Loss: 0.0814
Epoch: 341, Loss: 0.0810
Epoch: 342, Loss: 0.0808
Epoch: 343, Loss: 0.0813
Epoch: 344, Loss: 0.0805
Epoch: 345, Loss: 0.0809
Epoch: 346, Loss: 0.0800
Epoch: 347, Loss: 0.0817
Epoch: 348, Loss: 0.0808
Epoch: 349, Loss: 0.0820
Epoch: 350, Loss: 0.0814
100%|████████████████████████████████████████| 17535/17535 [01:31<00:00, 191.05it/s]
(258.5716552734375, 0.3679498146564015)
Epoch: 351, Loss: 0.0818
```

```
    return mean_rank, mrr, hits_at_k
```

torch_geometric.nn.kge.transe

torch_geometric.nn.kge.base

Conclusion

- Highly scalable
- Minimal parameterization for representing hierarchical relationships
- Simple!

- TransE performs well compared to other models across all setting
    - However, it is still unclear that TransE can applied to all relation types (1-Many, etc.)

    -> Limitation for symmetric, n-ary relation