# Translating Embeddings for Knowledge Graph Completion with Relation Attention Mechanism

Wei Qian, Cong Fu, Yu Zhu, Deng Cai, Xiaofei He

Xiaofan Yan

# Outline

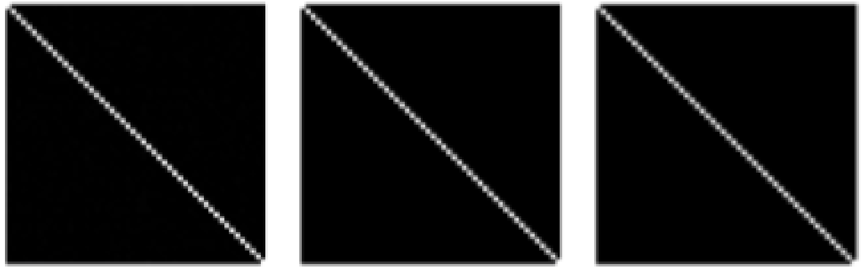
- Abstract
- Introduction
- Related work
- Method
- Expriments
- Conclusion and future work

# Abstract

- Translation based embedding method
  - Previous models
    - Failed creating attention mechanism
    - Because they ignore the hierarchical routine of human coghition
  - When people predict relations
    - first check the category of entities
    - then focus on fined-grained relation-related attributes to make the decision
  - Propose **TransAt**
    - Relation-related categories of entities
    - Relation-related attention

# Introduction

- Embedding based methods encode each object in knowledge graphs into a continuous vector space
- For a triple $(h, r, t)$, the basic idea of TransE model is that this triple induce a functional relation corresponds to a translation of the embeddings of entities in vector space $R^k$, that is, $h + r \approx t$.



295 353 610

- Actually, all transformation matrices are similar and approximate to the unit matrix. This means TransR fails to learn transformations only focusing on part of dimensions.
- Propose **a knowledge graph embedding method with attention mechanism** named TransAt (Translation with Attention)

# Related Work

- Translation-based models
  - TransE (Bordesetal.,2013)
    - Has problems when dealing with reflexive or many-to-one/one-to-many/many-to-many relations
  - TransH (Wanget al., 2014)
    - Distributed representations when the entity is involved in different relations
  - TransR/CTransR (Lin et al., 2015)
    - Models entities and relations in different embedding spaces
  - TransD (Ji et al., 2015)
    - considers the diversity of both relations and entities, can be applied to large scale graphs.
  - KG2E (He et al., 2015b)
    - Ignore that different entities and relations may contain different (un)certainties
  - TranSparse (Ji et al., 2016)
    - solving the heterogeneity and imbalance issues in knowledge graphs

# Related Work

- Other methods
  - StructuredEmbedding(SE) (Bordesetal.,2011)
  - Semantic Matching Energy (SME) (Bordes et al., 2012; 2014)
  - Latent Factor Model (LFM) (Jenatton et al., 2012).
  - SingleLayerModel(SLM) (Socheretal.,2013).
  - NeuralTensorNetwork(NTN) (Socheretal.,2013).
  - GAKE (Feng et al., 2016).
- The ORC structure
  - (Zhang, 2017) Some entity nodes form a circle by edges representing the same relation r in a knowledge graph. Then the corresponding relation vector r would approximate 0 (i.e. $r \approx 0$) when using traditional translation-based methods to model related triplets.

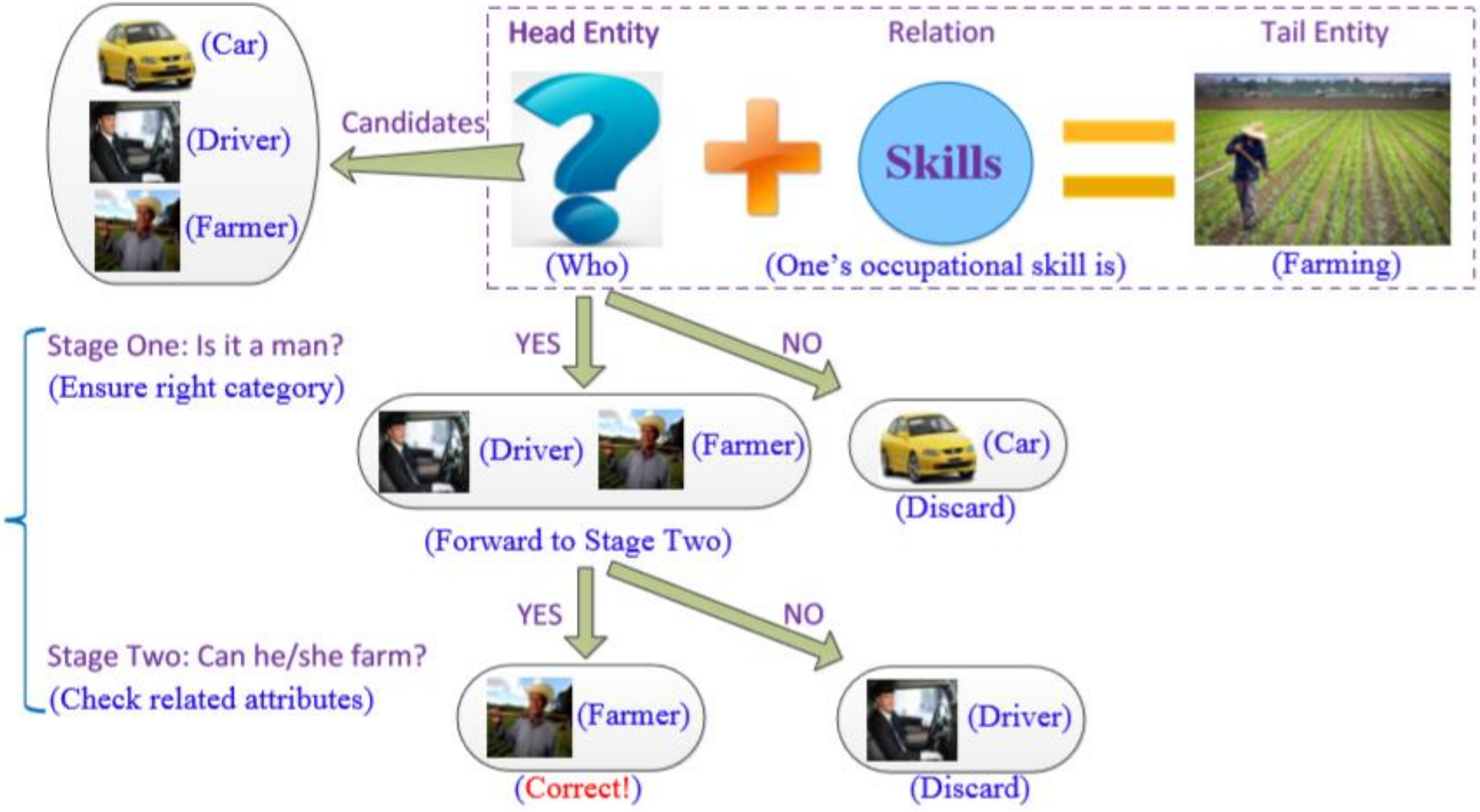
# Method

- Motivation
  - Human cognition for a relation follows a hierarchical routine
- TransAt
  - Learn the relation-related candidates and relation-related attention simultaneously

$$f_r(\mathbf{h}, \mathbf{t}) = \begin{cases} P_r(\mathbf{h}) + \mathbf{r} - P_r(\mathbf{t}) & h \in H_r, t \in T_r \\ \infty & otherwise \end{cases}$$

  - If the $h$ or the $t$ is not suitable for relation $r$, their distance for $r$ is infinite. If both the $h$ and the $t$ are suitable for $r$, their distance to $r$ is similar to TransE but only focuses on interested dimensions.

# Method



(Car)
(Driver)
(Farmer)
Candidates
Head Entity
Relation
Tail Entity
Skills
(Who)
(One's occupational skill is)
(Farming)
Stage One: Is it a man?
(Ensure right category)
YES
NO
(Driver)
(Farmer)
(Forward to Stage Two)
(Car)
(Discard)
YES
NO
Stage Two: Can he/she farm?
(Check related attributes)
(Farmer)
(Correct!)
(Driver)
(Discard)

# Method

| | |
|---|---|
| Kmeans | Aggregate all entities to some clusters and merge clusters |
| PCA | Obtain variances on dimension to determine to retain the dimension or not |
| An asymmetric operation | make h and t are different to solve ORC problem |

- Hope
  - If h and t in the suitable candidate set  --  massive score
  - Otherwise, entities not in the same are candidate set farfrom each other

**Initialize** → **Fair comparison** → **Sample batches** → **Optimize**

| Initialize | Fair comparison | Sample batches | Optimize |
|---|---|---|---|
| Xavier initialization | Sampling method | “bern” method | SGD |

# Experiments

- Datasets
  - WN11, WN18, FB15k, FB13
- Link prediction
  - Mean Rank: Take triples -> rank -> adjust entities in candidate set ahead of non-candidate
  - Hits@10: Hit count of $h$ -> hit count of $t$ -> average
  - Results:
    - have the best Mean Rank performance on both two datasets
    - has bigger γ and k in the optimal configuration
    - Asymmetric version of method outperforms the symmetrical one

# Experiments

- Link prediction

| Datasets | #WN18 | | | | #FB15k | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | Mean Rank | | Hits@10 | | Mean Rank | | Hits@10 | |
| | Raw | Filt | Raw | Filt | Raw | Filt | Raw | Filt |
| SE | 1,011 | 985 | 68.5 | 80.5 | 273 | 162 | 28.8 | 39.8 |
| SME (linear/bilinear) | 545/526 | 533/509 | 65.1/54.7 | 74.1/61.3 | 274/284 | 154/158 | 30.7/31.3 | 40.8/41.3 |
| LFM | 469 | 456 | 71.4 | 81.6 | 283 | 164 | 26.0 | 33.1 |
| GAKE | - | - | - | - | 228 | 119 | 44.5 | 64.8 |
| TransE | 263 | 251 | 75.4 | 89.2 | 243 | 125 | 34.9 | 47.1 |
| TransH (unif/bern) | 318/401 | 303/388 | 75.4/73.0 | 86.7/82.3 | 211/212 | 84/87 | 42.5/45.7 | 58.5/64.4 |
| TransR (unif/bern) | 232/238 | 219/225 | 78.3/79.8 | 91.7/92.0 | 226/198 | 78/77 | 43.8/48.2 | 65.5/68.7 |
| CTransR (unif/bern) | 243/231 | 230/218 | 78.9/79.4 | 92.3/92.3 | 233/199 | 82/75 | 44.0/48.4 | 66.3/70.2 |
| TransD (unif/bern) | 242/224 | 229/212 | 79.2/79.6 | 92.5/92.2 | 211/194 | 67/91 | 49.4/53.4 | 74.2/77.3 |
| TranSparse (share, S, unif/bern) | 248/237 | 236/224 | 79.7/80.4 | 93.5/93.6 | 226/194 | 95/88 | 48.8/53.4 | 73.4/77.7 |
| TranSparse (share, US, unif/bern) | 242/233 | 229/221 | 79.8/80.5 | 93.7/93.9 | 231/191 | 101/86 | 48.9/53.5 | 73.5/78.3 |
| TranSparse (separate, S, unif/bern) | 235/224 | 223/221 | 79.0/79.8 | 92.3/92.8 | 211/187 | **63**/82 | 50.1/53.3 | 77.9/79.5 |
| TranSparse (separate, US, unif/bern) | 233/223 | 221/211 | 79.6/80.1 | 93.4/93.2 | 216/190 | 66/82 | 50.3/**53.7** | 78.4/**79.9** |
| TransAt (bern) | 214 | 202 | **81.4** | **95.1** | 187 | 83 | 52.6 | 78.1 |
| TransAt (asy,bern) | **169** | **157** | **81.4** | 95.0 | **185** | 82 | 52.9 | 78.2 |

Table 2: Experimental results of link prediction.

# Experiments

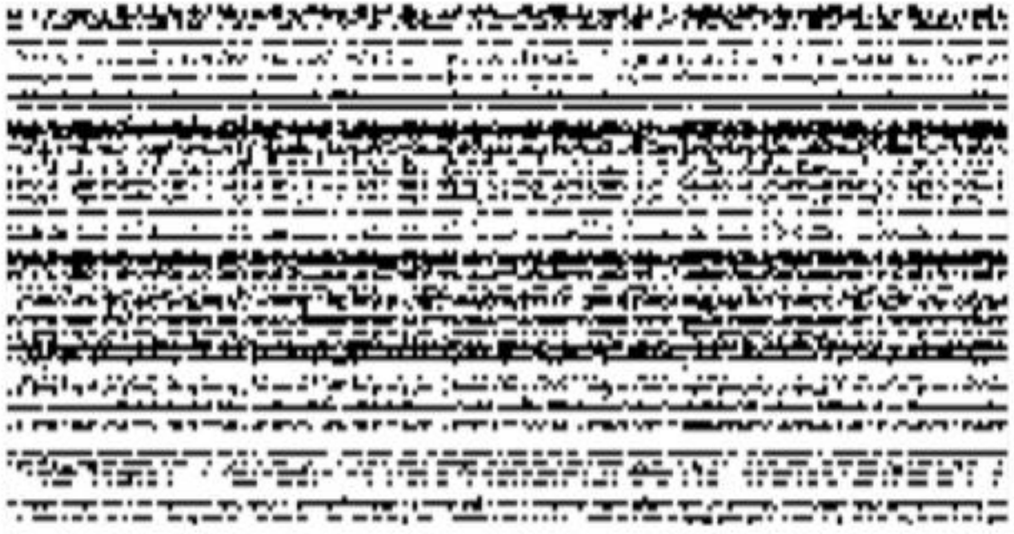
- Triple classification
  - Make a judgement on the correctness of a triple (binary classification task)

| Datasets | #WN11 | #FB13 |
|---|---|---|
| SE | 53.0 | 75.2 |
| SME(bilinear) | 70.0 | 63.7 |
| SLM | 69.9 | 85.3 |
| LFM | 73.8 | 84.3 |
| NTN | 70.4 | 87.1 |
| TransE(unif/bern) | 75.9/75.9 | 70.9/81.5 |
| TransH(unif/bern) | 77.7/78.8 | 76.5/83.3 |
| TransR(unif/bern) | 85.5/85.9 | 74.7/82.5 |
| CTransR(bern) | 85.7 | - |
| TransD(unif/bern) | 85.6/86.4 | 85.9/89.1 |
| TranSparse(share, S, unif/bern) | 86.2/86.3 | 85.5/87.8 |
| TranSparse(share, US, unif/bern) | 86.3/86.3 | 85.3/87.7 |
| TranSparse(separate, S, unif/bern) | 86.2/86.4 | 86.7/88.2 |
| TranSparse(separate, US, unif/bern) | 86.8/86.8 | 86.5/87.5 |
| TranAt(bern) | **88.2** | **89.1** |

Following the previous works, we learn the $\theta_r$ by maximizing the classification accuracy on the valid set. For a given triple $t\ (h, r, t)$, if its score is lower than $\theta_r$ and both the head entity and tail entity are in respective candidate sets, it will be classified as positive, otherwise, negative.

# Experiments

- Attention Mechanism
  - Pay attention to a part of the dimensions of relations
  - Some attention vectors $a_r$ are sparse



This figure demonstrates that TransAt can truly **pay attention to a part of the dimensions** of some relations.

# Conclusion and Future Work

- Propose a translation-based method
- Previous --  fail to learn attention
- Introduce a two stage discriminative method to achieve attention