TranSHER: Translating Knowledge Graph Embedding with Hyper-Ellipsoidal Restriction

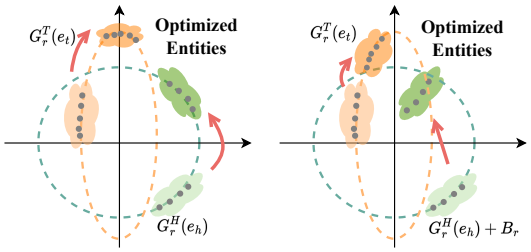
Yizhi Li[1]*, Wei Fan[2] , Chao Liu[3], Chenghua Lin[1]†, Jiang Qian[3]
[1] Department of Computer Science, The University of Sheffield, UK
[2] Department of Computer Science, University of Central Florida, USA
[3] Pingan Technology, China
{yizhi.li, c.lin}@sheffield.ac.uk, weifan@knights.ucf.edu,
lliuchao666@mail.ustc.edu.cn, jqian104@126.com

Abstract

Knowledge graph embedding methods are important for the knowledge graph completion (or link prediction) task. One existing efficient method, PairRE, leverages two separate vectors to model complex relations (i.e., 1-to-N, N-to-1, and N-to-N) in knowledge graphs. However, such a method strictly restricts entities on the hyper-ellipsoid surfaces which limits the optimization of entity distribution, leading to suboptimal performance of knowledge graph completion. To address this issue, we propose a novel score function *TranSHER*, which leverages relation-specific translations between head and tail entities to relax the constraint of hyper-ellipsoid restrictions. By introducing an intuitive and simple relation-specific translation, *TranSHER* can provide more direct guidance on optimization and capture more semantic characteristics of entities with complex relations. Experimental results show that *TranSHER* achieves significant performance improvements on link prediction and generalizes well to datasets in different domains and scales. Our codes are public available at https://github.com/yizhilll/TranSHER.

1 Introduction

Knowledge graph is proposed to structurally store human knowledge when the development of computer science has brought exponentially growing digitized information. Knowledge graphs have been widely adopted in important applications, such as semantic parsing (Berant et al., 2013), question generation and answering (Lin et al., 2015a; Hao et al., 2017; Peng et al., 2021), and information retrieval (Xiong et al., 2017). However, due to the expanding nature and difficulties of construction, knowledge graphs often suffer from incompleteness. Thus, knowledge graph completion (or link prediction) becomes a task of great importance from both a research and application perspective.



(a) PairRE (b) TranSHER

Figure 1: A Comparison Example of Distribution Optimization Between PairRE and TranSHER in 2-dimensional Space.

To construct a knowledge graph, entities and relations extracted from facts are treated as nodes and edges. A single fact in this graph is represented as a directed relation-specific link between two entities, which is denoted as a triplet like *(head, relation, tail)*. Since knowledge graphs usually contain a large number of entities and relations, knowledge graph embeddings are proposed to efficiently learn the representation of graphs and accomplish the link prediction task with *score functions*. Intuitions behind the design of score functions can be summarized as two important principles: 1) logical reasoning *relation patterns* such as symmetry (antisymmetry), inversion, and composition; 2) statistically categorized *relation types* like 1-to-1, 1-to-N, N-to-1, and N-to-N, where the latter three are called *complex relations*.

Inspired by word2vec (Mikolov et al., 2013), TransE (Bordes et al., 2013) represents knowledge graphs with relation-specific translations between head and tail entities to model the connections, where the score function works upon the translational distance $\|e_h + r - e_t\|$. Following this design, certain works (Wang et al., 2014; Lin et al., 2015b; Xiao et al., 2016) improve TransE by conducting relation-specific transformations on the entities be-

* Part of the work is done at Pingan Technology.
† Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8517 - 8528
December 7-11, 2022 ©2022 Association for Computational Linguistics

fore calculating distance. Certain works (Wang et al., 2014; Trouillon et al., 2016; Sun et al., 2019; Chao et al., 2021) further apply the principles of modeling relation patterns and complex relations to refine the design of score functions.

Among these methods, the recently proposed PairRE (Chao et al., 2021) uses separate relation vectors for head and tail entities to better model the complex relations and outperforms preceding methods. However, after our preliminary exploration, we notice that PairRE imposes a strong restriction that fixes the L_2 norm of the entity vector $\vec{e}$, and scales it with coefficients along all dimensions. Under such a restriction, the entities are essentially distributed on the surface of hyper-ellipsoids, where the foci are exactly laying on the axes. For the same reason, entity embeddings are strictly limited and can only be optimized around the hyper-ellipsoidal surface. In other words, entity distribution can only move along an arc path to match the true connections, as Fig. 1a shows, which may impose the close entity embeddings entangled and bring difficulties to the modeling.

To tackle the aforementioned challenges, we propose *TranSHER* which has a novel score function that leverages relation-specific translations between head and tail entities to relax the constraint of hyper-ellipsoid restrictions. We hypothesize that introducing translations for entities of PairRE can simplify the multi-relation modeling of entities on hyper-ellipsoids by providing an extra degree of freedom and thus improve the distribution learning of entities. In order to model complex relations, TranSHER first follows PairRE to conduct mappings on head and tail entities separately. Then, TranSHER additionally performs relation-specific translations on entities while holding the hyper-ellipsoidal restriction. Take Fig. 1b as an example: the mapped head entities (green) are moved closer to the tail entities (orange) without requiring obvious changes of the entities within the cluster and form a better distribution for relational modeling because the translational distance provides more direct guidance in the score function. The relation-specific translation in our score function allows more flexible optimization by relaxing the arc path constraint in PairRE brought by the hyper-ellipsoidal mapping, leading to a better distribution of entities and knowledge graph completion.

Moreover, TranSHER can better model the semantic characteristics of entities and utilize them to improve entity retrieval for complex relations. The semantic characteristics usually decide the categories of entities (e.g., people or companies); TranSHER can model the connected structure between entities of different categories more accurately in complex relations. For those complex relation triplets, TranSHER further enhances link prediction by retrieving multiple candidate entities in the correct category. For example, for a tail prediction query (film, produced_by, ?), TranSHER ranks the names of film producers at the front while other models may be confused by the related entertainment company or studio (cf. §5.6).

In addition, while achieving qualitative embeddings, TranSHER can further maintain the ability to model important relation patterns, namely, symmetry (antisymmetry), inversion, and composition under certain constraints (cf. §4). With such ability, TranSHER has the potential to be generalized well on datasets from different knowledge domains in different real-world settings.

We conduct comprehensive experiments across five datasets from different domains, which demonstrate the effectiveness of TranSHER. Impressively, TranSHER has achieved substantial improvement compared with the best baseline: the MRR increase has reached up to 4.6% on YAGO37 and 3.2% on ogbl-wikikg2. We also conduct many analytical experiments to study how translations of TranSHER behave and enhance knowledge graph completion. Some case studies further demonstrate the superiority of our approach.

2 Related Work

Knowledge graph embedding methods are proposed to model the intrinsic properties of facts and to conduct knowledge graph completion. To further complete knowledge graphs, these methods use designed score functions to model complex relations and various patterns. By measuring the relation-related distance between entities, KGE models predict the probabilities of given triplet queries $(e_h, r, ?)$ or $(?, r, e_t)$ to complete the graph. Such models are characterized by their corresponding score functions and the embeddings are usually optimized with gradient descent algorithms.

Distance-based score functions model the triplet facts by calculating distances between entity embeddings in the Euclidean space. Proposed by TransE (Bordes et al., 2013), one popular practice is to conduct a relation-specific translation on

the given entity before distance calculation, i.e., let $e_h + r \approx e_t$ in the case where the fact holds. Such a translational principle empowers the models to conduct knowledge graph completion on large-scale datasets while maintaining their performance. Many score functions, such as TransH, TransR, and ManifoldE (Wang et al., 2014; Lin et al., 2015b; Xiao et al., 2016), followed this principle of translation and achieved fair performances. However, RotatE (Sun et al., 2019) claims these extended works of TransE are weak in modeling certain relation patterns and thus proposes a solution of modeling in the complex space. Moreover, PairRE (Chao et al., 2021) argues that complex relations can be better modeled by separating relation vectors for heads and tails. Our TranSHER aims to provide a more effective model for complex relations by bridging the gap between translational distance-based models and the latest model PairRE.

Semantic matching score functions aim to predict the existence of facts by measuring the semantic similarity among entities and the given relation in the same representation space. RESCAL (Nickel et al., 2011) introduces a bilinear function $h^\top M_r t$ to represent the similarity score but suffers from modeling complexity. Some following work such as DistMult, HolE, and ComplEx (Yang et al., 2015; Nickel et al., 2016; Trouillon et al., 2016) intend to simplify the model while preserving critical features. A more recent work SEEK (Xu et al., 2020) generalizes the existing semantic matching score functions by segmenting the embedding to facilitate feature interactions, while keeping the same model size. In general, the semantic matching models struggled to distinguish similar entities and lack the ability to simultaneously model multiple relation patterns. Our TranSHER model tackles these challenges by proposing a novel score function which leverages relation-specific translations, yielding an effective improvement in modeling complex relations.

3 Problem Formulation

In this section, we describe the task definition and notations for better illustration. The set of known facts in a knowledge graph is represented by $\mathcal{T}$, which includes triplets (e_h, r, e_t). The notation $(e_h, r, e_t) \in \mathcal{T}$ will be used when the fact holds. The set of entity e and the set of relation r are denoted as $\mathcal{E}$ and $\mathcal{R}$.

Knowledge graph completion (also regarded as link prediction) aims to predict the missing links of knowledge graphs. Specifically, given a triplet $(e_h, r, e_{t'})$ or $(e_{h'}, r, e_t)$, a score function f_r is required to output an existence probability of the triplet. Since all the entities in $\mathcal{E}$ are provided in the training set, knowledge graph completion can also be regarded as a ranking task. For the true candidate entities, models are expected to assign higher rankings than false ones.

For a given entity-relation query $(e_h, r, ?)$ or $(?, r, e_t)$, there may exist multiple correct answers to complete the triplet, i.e., the quantities of those entities satisfy $\|\{e_{t'}|(h, r, e_{t'}) \in \mathcal{T}\}\| > 1$ or $\|\{e_{h'}|(e_{h'}, r, e_t) \in \mathcal{T}\}\| > 1$. According to the average *heads per tail* and *tails per head* counted through the dataset, relations in $\mathcal{R}$ can be categorized into 4 types: 1-to-1, 1-to-N, N-to-1, and N-to-N (Wang et al., 2014).

Relations can also be summarized by several significant patterns: symmetry/antisymmetry, inversion, and composition. The definitions are given as follows:

- A relation r is **symmetric** or **antisymmetric** if $(e_1, r, e_2) \in \mathcal{T}, \forall e_1, e_2 \in \mathcal{E} \Rightarrow (e_2, r, e_1) \in \mathcal{T}$ or $(e_2, r, e_1) \notin \mathcal{T}$.
- Relation r_1 and relation r_2 are **inverse** if $(e_1, r_1, e_2) \in \mathcal{T}, \forall e_1, e_2 \in \mathcal{E} \Rightarrow (e_2, r_2, e_1) \in \mathcal{T}$.
- Relation r_3 is **composed** by relation r_1 and r_2 if $(e_1, r_1, e_2) \in \mathcal{T} \wedge (e_2, r_2, e_3) \in \mathcal{T}, \forall e_1, e_2, e_3 \in \mathcal{E} \Rightarrow (e_1, r_3, e_3) \in \mathcal{T}$.

4 TranSHER

4.1 Score Function

We propose a simple yet effective translational distance-based score function *TranSHER*. The key intuition behind TranSHER is to provide more freedom with the relation-specific translation to ease the hyper-ellipsoidal restriction, while still keeping enough ability for complex relations modeling and training stability.

For this aim, TranSHER first maps the entity vectors to hyper-ellipsoids with underlying fixed-norm restriction for the actual entity embeddings and hence brings about general training stability; then conducts a relation-specific translation on the restricted entities for modeling the distances between mapped head and tail clusters. Specifically, we first define a mapping function $G(e)$ to restrict the entities on the hyper-ellipsoid surface. Since

the fact triplets are directional, we use two separate relation-specific mapping functions $G_r^H(e_h)$ and $G_r^T(e_t)$ to manage the cases when an entity is considered as head or tail, correspondingly:

$$G_r^H(e_h) = r^H \circ \frac{e_h}{|e_h|}, r^H, h \in \mathbb{R}^k \quad (1)$$

$$G_r^T(e_t) = r^T \circ \frac{e_t}{|e_t|}, r^T, t \in \mathbb{R}^k \quad (2)$$

where the $\circ$ here stands for the element-wise product. The mapping can be regarded as restricting the entities on unit hyper-spheres by fixing the L_2 norm $||\vec{e}||_2 = 1$ and conducting further linear scaling w.r.t. different relations. As the softmax loss in training intends to learn radially distributed entity representations (Wang et al., 2017), such a fixed-norm restriction of entity representations consequently benefits the stability of TranSHER optimization process (Xu and Durrett, 2018; Wang and Isola, 2020). With G_r^H and G_r^T, we are able to map the entity vectors to two hyper-ellipsoids according to the relations and whether they are heads or tails, as shown in Fig. 1a. Note that the entity embeddings will distribute on the unit hyper-sphere if $r^H = \mathbb{1}$ or $r^T = \mathbb{1}$.

Then, we introduce a relation-specific translation item $B_r \in \mathbb{R}^k$, which not only eases the hard hyper-ellipsoidal restriction but also encourages to identify the hard-to-distinguish entities close in space for complex relations. Accordingly, the final score function of TranSHER can be derived as:

$$f_r(e_h, e_t) = \gamma - ||G_r^H(e_h) + B_r - G_r^T(e_t)||_1 \quad (3)$$

where γ is an adjustable constant margin. Note that all the embeddings share the same dimension setting k, i.e., $r^H, r^T, B_r, e \in \mathbb{R}^k$. The additional translation for the entities with hyper-ellipsoidal restriction increases the degree of freedom of the score function and hence could provide extra optimization options for modeling the distances between entity clusters connected by complex relations.

4.2 Initialization

The initialization strategy plays an important role in neural network optimization (Erhan et al., 2009; Hayou et al., 2018), especially in the case of knowledge graph modeling that consists of enormous numbers of entity-relation-entity interactions. Classic knowledge graph embedding works pay less attention to initialization and adopt simple strategies. For example, both TransE and PairRE randomly initialize all the embedding weights with the same uniform distribution. After our practical implementation, we notice different initialization strategies place different assumptions on embeddings, which also largely influence the performances.

In this regard, all three main components in TranSHER need to be well-considered and initialized, i.e., the relation embeddings $\mathcal{R}$, entity embeddings $\mathcal{E}$, and translations B_r. To better model the distribution of knowledge graph embeddings, we propose to conduct initialization searching for the optimal initial distributions for each component in TranSHER. We select empirically validated strategies from two main categories, uniform, and normal distributions, for the component-independent initialization of TranSHER. For uniform distribution we follow the setting in Sun et al. (2019), where the weights are initialized with the gamma uniform $\mathcal{U}_\gamma(-\frac{\gamma+\epsilon}{k}, \frac{\gamma+\epsilon}{k})$. The ϵ here is a placeholder constant for the edge case that $\gamma = 0$. For the normal distribution, the Xavier normal proposed in Glorot and Bengio (2010) is adapted as $\mathcal{N}_X(0, g \cdot \sqrt{\frac{2}{k}})$ for TranSHER, where the gain g is defined as a scaling factor.

Without strict distribution assumptions for the parameters in the three components (i.e., $\mathcal{R}$, E, and B_r), TranSHER allows the components to be initialized independently with different distributions. The selection of the initial distribution for each component in TranSHER can vary through different knowledge graphs since they do not share the same data distribution. We will discuss in section 5.5 that, due to a more effective optimization process and potentially introducing appropriate inductive bias, such an initialization searching strategy allows TranSHER to produce better results.

4.3 Training and Optimization

Following the standard self-adversarial training framework (Sun et al., 2019), the general loss function for optimization is:

$$\mathcal{L} = -\log \sigma(f_r(e_h, e_t)) + \sum_i^N p(e_{h'_i}, r, e_{t'_i}) \log \sigma(-f_r(e_{h'_i}, e_{t'_i})) \quad (4)$$

where σ stands for the sigmoid activate function and $p(e_{h'_i}, r, e_{t'_i})$ is the self-adversarial weight calculated according to the scores.

4.4 Modeling Ability

Given the aforementioned design, our score function can be better optimized in training without losing the ability to learn different relation patterns. We prove that TranSHER can model symmetric/antisymmetric, inverse, and composed relations with the following constraints:

- **symmetry**: $(r_1^H \circ \frac{e_1}{|e_1|} + B_{r_1} - r_1^T \circ \frac{e_2}{|e_2|} = 0) \wedge (r_1^H \circ \frac{e_2}{|e_2|} + B_{r_1} - r_1^T \circ \frac{e_1}{|e_1|} = 0) \Rightarrow r_1^T = -r_1^H$
- **antisymmetry**: $(r_1^H \circ \frac{e_1}{|e_1|} + B_{r_1} - r_1^T \circ \frac{e_2}{|e_2|} = 0) \wedge (r_1^H \circ \frac{e_2}{|e_2|} + B_{r_1} - r_1^T \circ \frac{e_1}{|e_1|} \neq 0) \Rightarrow r_1^T \neq -r_1^H$
- **inversion**: $(r_1^H \circ \frac{e_1}{|e_1|} + B_{r_1} - r_1^T \circ \frac{e_2}{|e_2|} = 0) \wedge (r_2^H \circ \frac{e_2}{|e_2|} + B_{r_2} - r_2^T \circ \frac{e_1}{|e_1|} = 0) \Rightarrow (r_1^T \circ r_2^T = r_1^H \circ r_2^H) \wedge (B_{r_1} = -B_{r_2})$
- **composition**: $(r_1^H \circ \frac{e_1}{|e_1|} + B_{r_1} - r_1^T \circ \frac{e_2}{|e_2|} = 0) \wedge (r_2^H \circ \frac{e_2}{|e_2|} + B_{r_2} - r_2^T \circ \frac{e_3}{|e_3|} = 0) \wedge (r_3^H \circ \frac{e_1}{|e_1|} + B_{r_3} - r_3^T \circ \frac{e_3}{|e_3|} = 0) \Rightarrow (r_3^H = r_1^H \circ r_2^H) \wedge (r_3^T = r_1^T \circ r_2^T) \wedge (B_{r_3} = r_2^H \circ B_{r_1} + r_1^T \circ B_{r_2})$

This proof shows that TranSHER introduces additional relation-specific translations for easing the hyper-ellipsoidal restriction without losing the ability of modeling various relation patterns.

5 Experiments

5.1 Experimental Setup

Datasets and Evaluation. We conduct extensive experimentation on five publicly available datasets for two evaluation settings of the link prediction task. For the classic full ranking setting that requires an exhaustive search through the entity set $\mathcal{E}$, we select FB15k-237 (Toutanova and Chen, 2015), YAGO37 (Guo et al., 2018), and DB100K (Ding et al., 2018), which are extracted and constructed from knowledge databases (Suchanek et al., 2007; Bollacker et al., 2008; Bizer et al., 2009). Adopted from the Open Graph Benchmark (OGB) (Hu et al., 2020), the other setting of partial ranking only requires distinguishing the target entity from a randomly sampled entity subset. The ogbl-wikikg2 dataset with a massive number of triplets and the ogbl-biokg dataset consisting of biomedical facts are selected from the OGB link property prediction leaderboard. The statistics of the dataset are listed in Tab. 1. For both evaluation settings, the Mean Reciprocal Rank (MRR) is regarded as the main metric and Hits at N (HIT@N) as the auxiliary metric. More details of the datasets and evaluation protocol can be referred to in Appendix A.1 and A.2.

Dataset	Rel.	Ent.	Train	Valid	Test
FB15k-237	237	15k	272k	18k	20k
DB100K	470	100k	598k	50k	50k
YAGO37	37	123k	989k	50k	50k
ogbl-wikikg2	535	2,500k	16,109k	429k	598k
ogbl-biokg	51	94k	4,763k	163k	163k

Table 1: Statistics of Datasets.

Baselines. Our baselines include the two main categories of knowledge graph embedding methods for comparison. For the similarity-based *semantic matching* methods, DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), and SEEK (Xu et al., 2020) are selected. We choose the classic TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), and the recently proposed PairRE (Chao et al., 2021) as the baselines in the other category of *distance-based* methods. The model PairRE is the main baseline in our work. More details can be referred to in Appendix A.3.

Implementation Details. Following our design in §4.2, the three components of TranSHER are initialized with the gamma uniform (Sun et al., 2019) and the Xavier normal (Glorot and Bengio, 2010) distribution alternatively to achieve the best result. Our model is also fine-tuned with light parameter search on γ and regularization weights on translation embeddings according to datasets. All the experiments are set up in a GPU-accelerated hardware environment. We follow Wang et al. (2014) to counts *hpt* and *tph* to categorize relation types through the given dataset $\mathcal{T}$. Further implementation details could be found in Appendix A.4.

5.2 Overall Results

As revealed in Tab. 2 and Tab. 3, our model achieves significant performance improvement in the main metric MRR on all five different link prediction datasets compared to the strong baselines. On the datasets that require full ranking through all entities, TranSHER makes substantial improvement on the main metric MRR as shown in Tab. 2, surpassing PairRE and SEEK. As for results on datasets in Hu et al. (2020), TranSHER also has considerable performance gain as shown in Tab. 3. For the two datasets using the par-

Dataset	FB15k-237				DB100K				YAGO37			
Metric	MRR	HIT@1	HIT@3	HIT@10	MRR	HIT@1	HIT@3	HIT@10	MRR	HIT@1	HIT@3	HIT@10
TransE	.294	-	-	.465	.111	.016	.164	.270	.303	.218	.336	.475
DistMult	.241	.155	.263	.419	.233	.115	.301	.448	.365	.262	.411	.575
ComplEx	.247	.158	.275	.428	.242	.126	.312	.440	.417	.320	.471	.603
RotatE	.338	.241	.375	.533	-	-	-	-	-	-	-	-
SEEK	.338	**.268**	.370	.467	.338	.268	.370	.467	.454	.370	.498	.622
PairRE	.351	.256	.387	.544	.412	.309	.472	**.600**	-	-	-	-
TranSHER	**.360**	.264	**.397**	**.551**	**.431**	**.345**	**.476**	.589	**.490**	**.404**	**.538**	**.647**

* $p < 0.01$ is satisfied in the significance testing.

Table 2: Results on Full Ranking Settings Datasets. Results on FB15k-237 and DB100K are taken from (Chao et al., 2021), while results on YAGO37 are taken from (Xu et al., 2020). Dim is referred to the dimension parameter k of entity embeddings.

Dataset	ogbl-wikikg2			ogbl-biokg		
Metric	Dim	Test MRR	Valid MRR	Dim	Test MRR	Valid MRR
TransE	500	.4256	.4272	2000	.7452	.7456
DistMult	500	.3729	.3506	2000	.8043	.8055
ComplEx	250	.4027	.3759	1000	.8095	.8105
RotatE	250	.4332	.4353	1000	.7989	.7997
PairRE	200	.5208	.5423	2000	.8164	.8172
TranSHER	200	**.5536**	**.5662**	2000	**.8233**	**.8244**

Table 3: Results on Open Graph Benchmark Link Prediction Datasets. Results of baselines are taken from its official leaderboard (Hu et al., 2020). The *Dim* column is referred to the dimension parameter k of entity embeddings.

tial ranking setting, TranSHER also achieves incremental performances. On the very large-scale dataset ogbl-wikikg2, TranSHER improves about 3% MRR while keeping the dimension number of the entity and thus show its potential to extend on a knowledge graph with a large size. On the ogbl-biokg dataset, which distinguishes the challenge by isolating entities by types, TranSHER also shows its superiority at generalizing well across domains.

5.3 Complex Relations Modeling

The triplets with complex relations (1-to-N, N-to-1, and N-to-N) hold a large portion in many datasets. More importantly, they are more difficult to model than 1-to-1 relations. In this regard, we conduct experiments to analyze the performance of TranSHER on different types of triplets.

Results on FB15k-237 along relation types in Tab. 4 demonstrate that TranSHER makes stable gains on triplets with complex relations. For the N-to-N relation triplets (accounting for 87% of the whole dataset), TranSHER achieves better performances in MRR and HIT@10 on both head prediction and tail prediction tasks, even compared to the best baseline PairRE. This signifies that our proposed model can actually model complex relations better. A potential reason for this improvement of TranSHER is its most distinct part, relation-specific translation. We will give a more detailed analysis in the following sections.

To further learn the intrinsic influence of the relation-specific translation of TranSHER on different relation types, we visualize the translation embeddings on the FB15k-237 and DB100K by presenting the absolute value heat maps. As shown in Fig. 2, the translation embeddings of complex relations have obvious color differences from those 1-to-1 relation embeddings. The embeddings of 1-to-1 relations are closer to zero than complex relations; this signifies the translational item mainly contributes to complex relations 1-to-N, N-to-1, and N-to-N, whilst 1-to-1 relations are learned less. Specifically, translations of N-to-N relations are the most active, which suggests TranSHER puts more effort into these hard-to-learn relations. The behavior of translations on 1-to-1 relations implies that the easier relations require less optimization during training. The observation of this preference for complex relations from the relation-specific translations provides supportive evidence for the experimental results in Tab. 4.

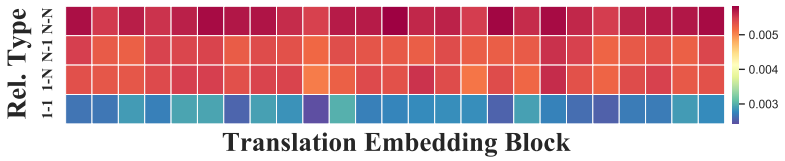
5.4 Translation Impacts

In order to explore how exactly the relation-specific translation affects modeling, we analyze the influence of the translations from the perspectives of training optimization and the score function itself.

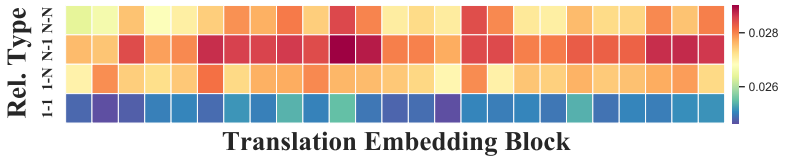
To study TranSHER from the view of the training process, the standard deviation of weight gradients for entities and relations are plotted in Fig. 3,

Rel. Type	1-to-1	1-to-N	N-to-1	N-to-N	1-to-1	1-to-N	N-to-1	N-to-N
Task	**Head Prediction (MRR)**				**Tail Prediction (MRR)**			
TransE*	.496	.457	.084	.251	.482	.074	.750	.365
DistMult*	.215	.441	.074	.231	.214	.052	.728	.346
ComplEx*	.357	.462	.092	.247	.371	.060	.741	.353
RotatE*	**.504**	.467	.090	.258	.488	.077	.758	.373
PairRE*	.494	.482	.112	.275	**.495**	.076	.766	.380
TranSHER	.501	**.487**	**.119**	**.285**	.494	**.079**	**.779**	**.389**
Task	**Head Prediction (HIT@10)**				**Tail Prediction (HIT@10)**			
TransE*	.609	.661	.166	.469	.594	.138	.879	.610
DistMult*	.453	.644	.140	.422	.448	.113	.844	.560
ComplEx*	.521	.657	.181	.453	.510	.126	.861	.588
RotatE*	.604	.671	.170	.471	.589	.146	.884	.611
PairRE*	.599	.667	**.211**	.488	.589	.147	.884	.619
TranSHER	**.615**	**.674**	**.211**	**.500**	**.604**	**.162**	**.891**	**.624**

Table 4: Evaluation Result on Different Relation Types on FB15k-237. Models with $*$ are reproduced with the self-adversarial framework (Sun et al., 2019).



(a) FB15K-237



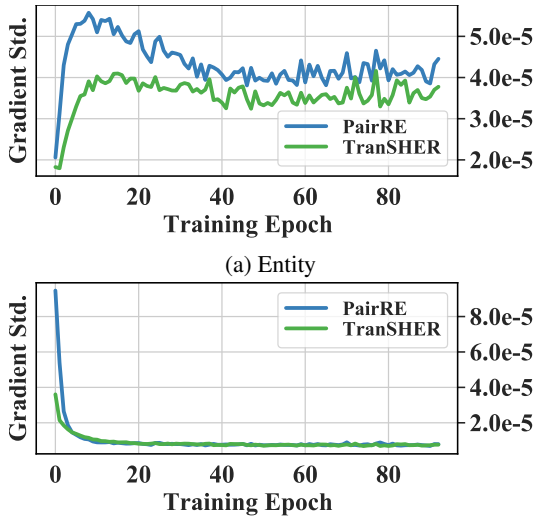
(b) DB100K

Figure 2: Heat Map Of Translation Embedding Values. The translation embeddings are first grouped by their relation types and their average absolute values are calculated along each dimension. We then implement a mean pooling operation on relation-type-wide average bias embeddings. Specifically, the pooling block size is set to 60 for the FB15k-237 model (1500-dim) and 20 for the DB100K model (500-dim).

Initialization			FB15k-237				DB100K			
$\mathcal{E}$	G_r	B_r	MRR	HIT@1	HIT@3	HIT@10	MRR	HIT@1	HIT@3	HIT@10
$\mathcal{N}_\mathcal{X}$	$\mathcal{N}_\mathcal{X}$	$\mathcal{N}_\mathcal{X}$	.357	.262	.394	.548	.426	.336	.475	**.592**
$\mathcal{N}_\mathcal{X}$	$\mathcal{N}_\mathcal{X}$	$\mathcal{U}_\gamma$	.353	.260	.389	.542	.429	.345	.474	.585
$\mathcal{N}_\mathcal{X}$	$\mathcal{U}_\gamma$	$\mathcal{N}_\mathcal{X}$	.353	.258	.389	.545	**.431**	.345	**.476**	.589
$\mathcal{U}_\gamma$	$\mathcal{N}_\mathcal{X}$	$\mathcal{N}_\mathcal{X}$	**.360**	**.264**	**.397**	**.551**	.423	.335	.470	.585
$\mathcal{N}_\mathcal{X}$	$\mathcal{U}_\gamma$	$\mathcal{U}_\gamma$	.347	.255	.381	.533	.430	**.347**	.473	. 583
$\mathcal{U}_\gamma$	$\mathcal{N}_\mathcal{X}$	$\mathcal{U}_\gamma$	.355	.261	.389	.543	.425	.343	.470	.580
$\mathcal{U}_\gamma$	$\mathcal{U}_\gamma$	$\mathcal{N}_\mathcal{X}$	.357	.262	.394	.548	.424	.338	.469	.583
$\mathcal{U}_\gamma$	$\mathcal{U}_\gamma$	$\mathcal{U}_\gamma$	.348	.255	.384	.535	.424	.341	.468	.577

Table 5: A Study of Different Initialization Strategies for TranSHER. The experiments are conducted on FB15k-237 and DB100K datasets with full ranking settings. The results are grouped by the number of gamma uniform or Xavier normal distributions used in the combinations.

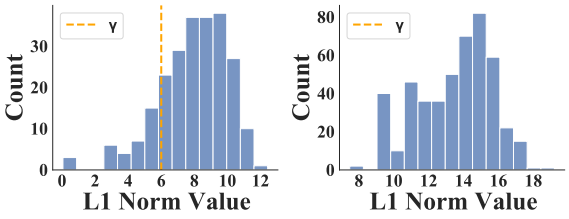
following similar settings in Glorot and Bengio (2010). We found that TranSHER largely reduces the gradient standard deviation of relation embeddings only at the beginning of training and keeps a similar trend to the baseline for the rest epochs. What most distinguishes the optimization process of TranSHER from PairRE is that TranSHER maintains a relatively low standard deviation of entity embedding gradients along with the whole training. Such a low standard deviation implies the stable optimization progress of TranSHER. The adjustment of entity embeddings usually requires more effort since the number of entities is about thirty times larger than the relations in FB15k-237 (and



(a) Entity

(b) Relation

Figure 3: Statistics of Model Gradients in FB15k-237 Training. We plot the standard deviation of weight gradients for entity and relation embeddings at the beginning of each training epoch. Gradient standard deviation of translations is not calculated and plotted since the baseline model does not contain such a module.



(a) FB15K-237

(b) DB100K

Figure 4: Histogram of the L1 Norm of TranSHER Translations on two dataset.

similar cases for other datasets, see Tab. 1). We suspect that the generally superior results of TranSHER are mainly brought about by a more stable optimization of entity embeddings.

From the perspective of the score function, the relation-specific translation item can directly affect the distance calculation. We plot the distribution of L1 Norm values of the k-dimensional translation embeddings to learn of such an impact (Fig. 4). Most L1 Norms of translation embeddings are larger than the margin γ in both experiments on FB15k-237 and DB100K, which suggests a prominent impact from the translation in modeling.

5.5 Initialization Strategy

During the implementation, we found that the initial strategy of TranSHER components is crucial to the performance, and thus we conduct a further study on the initialization strategies combinations on the FB15k-237 and DB100K datasets.

Query	(Cinderella, /film/film/produced_by, **?**)	
Answer	Walt Disney	
Model	TranSHER	PairRE
Rank 1	• **Walt Disney**	Walt Disney Animation Studios
Rank 2	Ivan Reitman	The Walt Disney Company
Rank 3	Hayao Miyazaki	Jerry Bruckheimer
Rank 4	Jerry Bruckheimer	Alan Menken
Rank 5	Walt Disney Pictures	•**Walt Disney**
Rank 6	Gary Goetzman	Hayao Miyazaki
Rank 7	Lawrence Golden	Walt Disney Picture
Rank 8	Howard Ashman	John Lasseter

* The • refers to the correct answer.

Table 6: A Case Study of Tail Prediction on FB15k-237.

We compare the results of several initialization methods to learn the effect of different distributions, as revealed in Tab. 5. Specifically, the gamma uniform distribution $\mathcal{U}_\gamma$ and the Xavier normal distribution $\mathcal{N}_\mathcal{X}$ are alternatively adopted for three components in TranSHER, i.e. the entities $\mathcal{E}$, the relational mapping G_r, and the relation-specific translations B_r. Such a strategy of initialization combination produces a set of eight experiments on each dataset while keeping the same size of model parameters.

Among all the initialization combinations, the strategy of '$\mathcal{U}_\gamma\mathcal{N}_\mathcal{X}\mathcal{N}_\mathcal{X}$' leads to the strongest performance on FB15K-237, while the '$\mathcal{N}_\mathcal{X}\mathcal{U}_\gamma\mathcal{N}_\mathcal{X}$' variant gets the best MRR result on DB100K. The observation of the best initialization strategy varying through datasets suggests that the initialization strategy of TranSHER can accommodate discrepancies through different knowledge graphs, which brings about performance gains on link prediction.

5.6 Case Study

We further provide a case study to illustrate the effectiveness of TranSHER in handling challenging link predictions. In Tab. 6, the query asks for the producer of a 1950 animated musical fantasy film *Cinderella* (NB: the produced_by relation defined in FB15k-237 *only* refers to producers). While PairRE is capable of retrieving relevant entities like the production studio/company, and even a composer that used to work on another *Walt Disney* film, it still struggles with learning the semantic meaning of the relation and mixing the neighbor entities in representation space. Meanwhile, the high-ranking entities found by TranSHER are mostly producers/directors and the exact subsidiary of the production studio. This implies that TranSHER does not only have the ability to cluster the rele-

vant entities restricted on hyper-ellipsoids but can also accurately model the semantic meaning of the particular relation "who is the producer of the film" with the relation-specific translation item. More cases can be found in Appendix B.

6 Conclusion

We propose a novel knowledge graph embedding model TranSHER for the link prediction task. TranSHER leverages relation-specific translation on entities with hyper-ellipsoidal restriction, which is explicitly encoded into the score function. By introducing the translation, TranSHER can improve the optimization of entities distributed on hyper-ellipsoids and shows ingenuity in understanding semantic characteristics. Moreover, we prove that TranSHER preserves the ability to represent logical reasoning relation patterns. Comprehensive experiments on different datasets show that TranSHER has robust performance and improves complex relation modeling.

Limitations

The proposed model TranSHER mainly provides insight into how the translation item can improve the knowledge graph embedding methods for the link prediction task with restricted entities. However, due to the enormous number of entities in knowledge graphs, this work does not directly show the learned entity representation distribution, which could potentially provide further information beneficial to the task.

Acknowledgement

Yizhi Li is fully funded by an industrial PhD studentship (Grant number: 171362) from the University of Sheffield, UK.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of Web Semantics*, (3):154–165.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*.

Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. PairRE: Knowledge Graph Embeddings via Paired Relation Vectors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4360–4369, Online.

Boyang Ding, Quan Wang, Bin Wang, and Li Guo. 2018. Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 110–121, Melbourne, Australia.

Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. 2009. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pages 153–160, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. 2018. Knowledge graph embedding with iterative guidance from soft rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. 2018. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, pages 22118–22133.

Chenghua Lin, Dong Liu, Wei Pang, and Zhe Wang. 2015a. Sherlock: A semi-automatic framework for quiz generation using a hybrid semantic similarity measure. *Cognitive computation*, (6):667–679.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 3111–3119, Red Hook, NY, USA.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 809–816, Madison, WI, USA.

Keqin Peng, Chuantao Yin, Wenge Rong, Chenghua Lin, Deyu Zhou, and Zhang Xiong. 2021. Named entity aware transfer learning for biomedical factoid question answering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, (10):78–85.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1041–1049, New York, NY, USA.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. From one point to a manifold: Knowledge graph embedding for precise link prediction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, page 1315–1321.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Jiacheng Xu and Greg Durrett. 2018. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Wentao Xu, Shun Zheng, Liang He, Bin Shao, Jian Yin, and Tie-Yan Liu. 2020. SEEK: Segmented embedding of knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3897, Online.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, 2015, Conference Track Proceedings*, San Diego, CA, USA.

A Experimental Details

A.1 Datasets

Extensive experiments are conducted on five publicly available datasets. The FB15k-237 dataset is filtered from another dataset FB15k (Toutanova and Chen, 2015; Bordes et al., 2013), which is built from a knowledge fact database, Freebase (Bollacker et al., 2008). Compared to FB15k-237, two larger-scale knowledge graphs DB100K (Ding et al., 2018) and YAGO37 (Guo et al., 2018) with about ten times of entities are also selected. The YAGO37 is a subset selected from the YAGO3 core facts (Guo et al., 2018; Suchanek et al., 2007). and the DB100K is constructed from the mapping-based objects of core DBpedia (Ding et al., 2018; Bizer et al., 2009). We additionally test our model on two distinguished link prediction datasets (Hu et al., 2020). ogbl-wikikg2 is a very large-scale dataset derived from the Wikidata knowledge base (Vrandečić and Krötzsch, 2014). ogbl-biokg uses data from biomedical data repositories and divides the entities into 5 types according to domain knowledge.

To prove the generalization ability of TranSHER, we select datasets of various entity quantities from 15k (FB15k-237) to 2,500k (ogbl-wikikg2), which spans three orders of magnitude.

A.2 Evaluation Protocol

Following standard implementation, we use Mean Reciprocal Rank (MRR) and Hits at N (HIT@N) as our metrics on all the datasets, whilst MRR is referred to as the main evaluation measure due to its relatively comprehensive perspective. The general evaluation settings for the link prediction task can be recognized as the full ranking setting and the partial ranking setting according to the selection methods of negative samples for testing.

Except for datasets selected from Hu et al. (2020), the ranking candidates are all entities that appeared in the knowledge graph, i.e. a full ranking task setting. Following the standard protocol, the scores of the extra correct entities in each query are filtered during full ranking testing. Note that for ogbl-wikikg2 and ogbl-biokg, the ranking candidates are the positive entity and 500 randomly sampled negative entities (with a separate 500 for prediction head and tail tasks). Specifically, the sampled negative entities belong to the same type of positive ones in the ogbl-biokg.

A.3 Baselines

Two main categories of knowledge graph embedding methods are chosen in our work to compare and validate the performance of TranSHER:

- Semantic matching score functions
- Distance-based score functions

The semantic matching score functions intend to study interactions in knowledge graphs with similarity-based score functions. We select DistMult (Yang et al., 2015) and ComplEx (Trouillon et al., 2016) as the representatives of such methods that leverage inner-product largely. Moreover, a recently proposed framework SEEK (Xu et al., 2020) is selected as a strong baseline for the dataset YAGO37. We also categorize SEEK as a semantic matching due to its usage of inner-product calculation even if it conducts hard segmentation on the embeddings.

Methods in the other category, *distanced-based*, design score functions to model the distance between connected entities in low-dimensional space. Additional to the foundational distance-based function TransE (Bordes et al., 2013), we also take two more effective models RotatE (Sun et al., 2019) and the PairRE (Chao et al., 2021) as our baselines. PairRE is the main baseline since it has competitive performance and adopts the same restriction on the entities as TranSHER.

A.4 Implementation Details

As described in §4.2, the entity embeddings in $\mathcal{E}$, mapping weights of G_r, and relation-specific translations B_r can be initialized either with gamma uniform implemented in (Sun et al., 2019; Chao et al., 2021) or the Xavier normal distribution (Glorot and Bengio, 2010). The gamma uniform is set as $\mathcal{U}_\gamma(-\frac{\gamma+\epsilon}{k}, \frac{\gamma+\epsilon}{k}), \epsilon = 2.0$ the scaling gain for Xavier normal $\mathcal{N}_X(0, g \cdot \sqrt{\frac{2}{k}})$ is set to $g = 1.0$. Parameter searches on γ and regularization weights on translation embeddings are adopted, while the embedding dimension k remains the same as PairRE. The extra effort of embedding dimension tuning is spent on the YAGO37 dataset since the original work PairRE does not provide the implementation on it.

All the experiments occupy a capacity under 16GB RAM on an RTX 3090 GPU. To get the relation type information following Wang et al. (2014), we count *hpt* and *tph* through all the triplets from the given dataset including the test split.

Query	(**?**, /medicine/symptom/symptom_of, jaundice)		(**?**, /music/group_membership/group, USA for Africa)	
Answer	vomiting		Stevie Wonder	
Model	TranSHER	PairRE	TranSHER	PairRE
Rank 1	• **vomiting**	pancreatic cancer	Janet Jackson	USA for Africa
Rank 2	dyspnea	liver tumor	Norah Jones	Michael Sembello
Rank 3	fever	liver cirrhosis	• **Stevie Wonder**	Elton John
Rank 4	liver cirrhosis	fever	Quincy Jones	Bobby Darin
Rank 5	diarrhea	hepatitis B	Prince	Norah Jones
Rank 6	anorexia	diarrhea	Michael McDonald	Irene Cara
Rank 7	headache	• **vomiting**	USA for Africa	• **Stevie Wonder**
Rank 8	pancreatic cancer	headache	Barbra Streisand	Quincy Jones

* The • refers to the correct answer.

Table 7: Additional Cases for Head Prediction Task on FB15k-237.

Query	(Dena Higley, /people/person/profession, **?**)		(Almost Famous, /film/film/genre, **?**)	
Answer	writer		coming of age	
Model	TransHER	PairRE	TransHER	PairRE
Rank 1	• **writer**	television producer	romance film	romance film
Rank 2	television producer	actor	• **coming of age**	LGBT
Rank 3	actor	television director	independent film	historical period drama
Rank 4	journalist	film director	romantic comedy	independent film
Rank 5	film director	film producer	black comedy	romantic comedy
Rank 6	television director	television presenter	LGBT	biography
Rank 7	author	• **writer**	historical period drama	biographical film
Rank 8	film producer	journalist	biographical film	• **coming of age**

* The • refers to the correct answer.

Table 8: Additional Cases for Tail Prediction Task on FB15k-237.

B Supplementary Case Study

We provide extra cases for further analysis on the FB15k-237 dataset, where the head prediction task cases are shown in Tab. 7 and tail prediction tasks in Tab. 8, respectively.

In general, we can observe that both TranSHER and PairRE can retrieve relevant entities in the same or similar topic according to the given relation-entity queries, which distribute from the entertainment domain to medical knowledge facts. For instance, in the *jaundice* case at the left of Tab. 7, the entities recalled by the models are all terminologies from the practice of medicine. This implies that the score functions with hyper-ellipsoidal restriction can model the entities in the same neighborhood and assign close positions in the latent space with regard to their graphical interactions such as the n-hop distances.

However, similar to the case described in §5.6, TranSHER shows additional precision for predicting the correct entities by learning the inherent semantic categorization of entities with the extra relation-specific translations. Regarding the aforementioned *jaundice* example, the query specifically asks for the symptom of the jaundice disease, which could be highly related to liver diseases.[1] Although the PairRE has learned the close relationships between jaundice and liver diseases and distributed these entities close with the hyper-ellipsoidal mapping, it fails to distinguish the concept disease from the close concept symptom and assigns the liver-related with high ranks. In contrast with PairRE, our proposed model TranSHER can learn the nuanced differences in these two concepts defined in the relations. As a result, we suspect that the semantic characteristic modeling of score functions can be improved by the additional relation-specific translation item in TranSHER.

[1] More information can be found at the Jaundice wikipage.