On Multi-Relational Link Prediction with Bilinear Models

Yanjie Wang
University of Mannheim, Germany
ywang@uni-mannheim.de

Rainer Gemulla
University of Mannheim, Germany
rgemulla@uni-mannheim.de

Hui Li
The University of Hong Kong
hli2@cs.hku.hk

Abstract

We study bilinear embedding models for the task of multi-relational link prediction and knowledge graph completion. Bilinear models belong to the most basic models for this task, they are comparably efficient to train and use, and they can provide good prediction performance. The main goal of this paper is to explore the expressiveness of and the connections between various bilinear models proposed in the literature. In particular, a substantial number of models can be represented as bilinear models with certain additional constraints enforced on the embeddings. We explore whether or not these constraints lead to *universal models*, which can in principle represent every set of relations, and whether or not there are *subsumption relationships* between various models. We report results of an independent experimental study that evaluates recent bilinear models in a common experimental setup. Finally, we provide evidence that relation-level ensembles of multiple bilinear models can achieve state-of-the art prediction performance.

1 Introduction

Multi-relational link prediction is the task of predicting missing links in an edge-labeled graph. We focus and use the terminology of *knowledge base completion* throughout. Large-scale knowledge bases (KB) such as DBPedia (Lehmann et al. 2015) or YAGO (Rebele et al. 2016) contain millions of entities and facts, but they are nevertheless far from being complete (Nickel et al. 2016). Given a set of entities (vertices) and relations (edge labels) that hold between these entities, the goal of multi-relational link prediction (Bordes et al. 2013) is to determine whether or not some entity e_1 links to some entity e_2 via a relation R, i.e., whether the fact $R(e_1, e_2)$ is true.

Embedding models have recently received considerable attention for knowledge base completion tasks (Bordes et al. 2013; Nickel, Rosasco, and Poggio 2016; Trouillon et al. 2016a). Such models embed both entities and relations in a low-dimensional latent space such that the structure of the knowledge base is (largely) maintained. The embeddings are subsequently used to predict missing facts or to detect erroneous facts.

The perhaps most basic class of embedding models is given by bilinear models. Such models predict a “score” for each fact $R(e_1, e_2)$ by computing a weighted sum—where the weights depend on R—of the pairwise interactions of the entity embeddings of e_1 and e_2. The scores are used to rank (pairs of) entities according to their predicted truthfulness. Bilinear models are comparably efficient to train and use and they can provide good prediction performance (Trouillon and Nickel 2017).

A large number of bilinear models has been proposed in the literature, including RESCAL (Nickel, Tresp, and Kriegel 2011), TransE (Bordes et al. 2013), DISTMULT (Yang et al. 2014), HolE (Nickel, Rosasco, and Poggio 2016), and ComplEx (Trouillon et al. 2016a). There is, however, little work on the expressiveness of and the connections between various bilinear models. In this paper, we argue that all of the aforementioned models can be seen as bilinear models subject to certain constraints. We study whether and under which conditions each model is *universal* in that it can represent every possible set of relation instances (or, more precisely, entity rankings). We also explore the size of the embeddings needed for universality and derive upper bounds for the embedding size needed to obtain embeddings consistent with a given dataset. We establish a number of subsumption relationships between various models by giving explicit constructions on how to transform instances of one model to instances of another model (sometimes with a different embedding size). A summary of our results is given in Tab. 1.

We report on an independent experimental study that compared various bilinear models on standard datasets in a common experimental setup. We found that the relative performance among the models is highly relation-dependent. We thus propose a simple relation-level ensemble of multiple bilinear models, which—according to our experiments—significantly and consistently improved prediction performance over individual models. In fact, we found that the ensemble performed competitively to the state-of-the-art embedding approaches, whether or not they are bilinear.

2 Multi-Relational Link Prediction

Let $\mathcal{E}$ and $\mathcal{R}$ be a set of entities and relation names. A knowledge base $\mathcal{K} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a collection of triples (i, k, j) where i, j, and k refer to subject, object and relation, resp. We denote by $K = |\mathcal{R}| \geq 1$ and $N = |\mathcal{E}| \geq 2$ the number of entities and relations, resp. We represent knowledge base $\mathcal{K}$ via a binary tensor $\boldsymbol{\mathcal{X}} \in \{0,1\}^{N\times N\times K}$, where $x_{ijk} = 1$ if and only if $(i,k,j) \in \mathcal{K}$. By convention, vectors $\boldsymbol{a}_i$ refer to rows of matrix $\boldsymbol{A}$ (as a column vector) and scalars a_{ij} to individual entries. Given dimensionalities r and r', we denote by $\boldsymbol{e}_{i,r}$ the i-th standard basis vector, by $\mathbf{0}_r$ the zero vector, and by $\mathbf{0}_{r\times r'}$ the zero matrix of the respective shape. Finally, let $\operatorname{diag}(\cdot)$ refer to a block-diagonal matrix built from the arguments (a vector or a list of matrices).

2.1 Preliminaries

A *score-based ranking model* is a model m that associates a *score* $s_k^m(i,j) \in \mathbb{R}$ with each subject-relation-object triple. Denote by $\boldsymbol{S}_k^m \in \mathbb{R}^{N\times N}$ the corresponding *scoring matrix* for relation k, i.e., $[\boldsymbol{S}_k^m]_{ij} = s_k^m(i,j)$. Denote by $\boldsymbol{\mathcal{S}}^m \in \mathbb{R}^{N\times N\times K}$ the *scoring tensor* of m, i.e., the tensor with frontal slices $\boldsymbol{\mathcal{S}}_{(k)}^m = \boldsymbol{S}_k^m$.

We are ultimately interested in rankings, not in scores. In particular, score-based models are used to rank (pairs of) entities by their predicted truthfulness, given a query of form $R(i,?)$, $R(?,j)$, or $R(?,?)$. Generally, a result with a higher score is considered more likely to be correct. We say that an $N \times N$ matrix is a *ranking matrix* if all its entries are in $\{\, 1, 2, \ldots, N^2 \,\}$ and whenever there is any entry with value $s > 1$, there is at least one other entry with value $s - 1$. Denote by $\pi(\boldsymbol{S})$ the unique ranking matrix associated with scoring matrix $\boldsymbol{S}$, where $\pi_{ij}(\boldsymbol{S}) \overset{\text{def}}{=} [\pi(\boldsymbol{S})]_{ij}$ is the *dense rank* of s_{ij} in the multiset of the entries of $\boldsymbol{S}$. For every pair of tuples $(i,j) \in N \times N$ and $(i',j') \in N \times N$, we have

$$s_{ij} \leq s_{i'j'} \iff \pi_{ij}(\boldsymbol{S}) \geq \pi_{i'j'}(\boldsymbol{S}).$$

For example,

$$\boldsymbol{S} = \begin{pmatrix} 0.2 & 2.4 & 1 \\ -1 & 4 & 2 \\ -3 & 0.2 & 0 \end{pmatrix} \quad \Longrightarrow \quad \pi(\boldsymbol{S}) = \begin{pmatrix} 5 & 2 & 4 \\ 7 & 1 & 3 \\ 8 & 5 & 6 \end{pmatrix}$$

In a slight abuse of notation, we overload π to also apply to tensors, sets of matrices, and sets of tensors. In particular, the *ranking tensor* $\pi(\boldsymbol{\mathcal{S}})$ for a score tensor $\boldsymbol{\mathcal{S}}$ is the $N \times N \times K$ tensor produced from $\boldsymbol{\mathcal{S}}$ by replacing every frontal slice $\boldsymbol{\mathcal{S}}_{(k)}$ with $\pi(\boldsymbol{\mathcal{S}}_{(k)})$. Moreover, for any set X, set $\pi(X) = \{\, \pi(x) : x \in X \,\}$. Observe that $\pi(\mathbb{R}^{N\times N})$ corresponds to the set of all possible ranking matrices, $\pi(\mathbb{R}^{N\times N\times K})$ to all possible ranking tensors, and that $\pi(-\boldsymbol{P}) = \boldsymbol{P}$ for any ranking matrix (or ranking tensor) $\boldsymbol{P}$.

2.2 Bilinear Models

Bilinear models are models whose scoring function $s_k(i,j)$ has form $\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j$, where $\boldsymbol{a}_i, \boldsymbol{a}_j \in \mathbb{R}^r$ and $\boldsymbol{R}_k \in \mathbb{R}^{r\times r}$ are model parameters and are referred to as the *embeddings* of entities i and j as well as relation k, resp. We refer to $r \in \mathbb{N}$ as the *size* of the model.

In this paper, we consider bilinear models as well as models that can be represented as bilinear models with an at most linear increase in model size. Although some of the model considered here may not "look" bilinear at first glance, we show that they are closely related to bilinear models. We denote throughout the set of all models of type t (and of size r) and by M^t (M_r^t).

RESCAL (Nickel, Tresp, and Kriegel 2011). An unconstrained bilinear model. Each model $m \in M_r^{\text{RESCAL}}$ is parameterized by an entity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times r}$ and K relation matrices $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_K \in \mathbb{R}^{r\times r}$. We have

$$s_k^m(i,j) = \boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j.$$

RESCAL can be seen as an extension of the low-rank matrix factorization methods prominent in recommender systems to more then one relation.

DISTMULT (Yang et al. 2014). Each model $m \in M_r^{\text{DISTMULT}}$ is parameterized by an entity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times r}$ and a relation matrix $\boldsymbol{R} \in \mathbb{R}^{K\times r}$. We have

$$s_k^m(i,j) = \boldsymbol{a}_i^T \operatorname{diag}(\boldsymbol{r}_k)\, \boldsymbol{a}_j.$$

DISTMULT can be seen as a variant of RESCAL that puts a diagonality constraint on the relation matrices. Due to this constraint, it can only model symmetric relations. The model is equivalent to the INDSCAL tensor decomposition (Carroll and Chang 1970).

HolE (Nickel, Rosasco, and Poggio 2016). Each model $m \in M_r^{\text{HolE}}$ is parameterized by an entity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times r}$ and a relation matrix $\boldsymbol{R} \in \mathbb{R}^{K\times r}$. We have

$$s_k^m(i,j) = \boldsymbol{r}_k^T(\boldsymbol{a}_i \star \boldsymbol{a}_j),$$

where $\star$ refers to the *circular correlation* between $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$, i.e., $(\boldsymbol{a}_i \star \boldsymbol{a}_j)_k = \sum_{t=1}^{r} a_{it} a_{j((k+t-2 \mod r)+1)}$. The idea of using circular correlation relates to associative memory (Nickel, Rosasco, and Poggio 2016). Hayashi and Shimbo (2017) provide an alternative viewpoint in terms of ComplEx, discussed next.

ComplEx (Trouillon et al. 2016a). Each model $m \in M_r^{\text{ComplEx}}$ is parameterized by an entity matrix $\boldsymbol{A} \in \mathbb{C}^{N\times r}$ and a relation matrix $\boldsymbol{R} \in \mathbb{C}^{N\times r}$. We have

$$s_k^m(i,j) = \operatorname{Re}(\boldsymbol{a}_i^T \operatorname{diag}(\boldsymbol{r}_k)\, \boldsymbol{a}_j),$$

where $\operatorname{Re}(\cdot)$ extracts the real part of a complex number ($\boldsymbol{a}_i^T \operatorname{diag}(\boldsymbol{r}_k)\, \boldsymbol{a}_j$ is not necessarily real). ComplEx is superficially related to DISTMULT but uses complex-valued parameter matrices.

TransE (Bordes et al. 2013). Each model $m \in M_r^{\text{TransE}}$ is parameterized by an entity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times r}$ and an relation matrix $\boldsymbol{R} \in \mathbb{R}^{K\times r}$. We have[1]

$$s_k^m(i,j) = -\left\| \boldsymbol{a}_i + \boldsymbol{r}_k - \boldsymbol{a}_j \right\|_2^2.$$

In contrast to the models presented above, TransE is a translation-based model, not a factorization-based model. The use of translations—i.e., differences between entity embeddings—is inspired by Word2Vec's word analogy results (Mikolov et al. 2013). Note that TransE can also be used with L_1 norm instead of L_2; we focus on the L_2 variant given above throughout.

[1] This definition differs from the original definition of TransE in that we negate all scores in order to rank larger scores higher.

3 Subsumption and Expressiveness

For a given class M_r^t of models, denote by $\mathcal{M}_r^t = \{\, \boldsymbol{\mathcal{S}}^m : m \in M_r^t \,\}$ the set of scoring tensors that the model class can represent. Let $\mathcal{M}^t = \cup_{r \in \mathbb{N}^+} \mathcal{M}_r^t$. Note that $\pi(\mathcal{M}_r^t)$ and $\pi(\mathcal{M}^t)$ denote the set of ranking tensors that can be represented by M_r^t and M^t, respectively.

3.1 Subsumption

We first explore subsumption relationships between different model classes as well as the the size of the entity representations needed for a subsumption to hold. We assume throughout that the number $N \geq 2$ of entities and the number $K \geq 1$ of relations are arbitrary but fixed.

We say that class M^{t_2} *subsumes* class M^{t_1} whenever $\pi(\mathcal{M}^{t_1}) \subseteq \pi(\mathcal{M}^{t_2})$. In other words, M^{t_2} is at least as expressive in terms of rankings as M^{t_1}. If $\pi(\mathcal{M}^{t_1}) \subset \pi(\mathcal{M}^{t_2})$, we say that M^{t_2} *strictly subsumes* M^{t_1}, indicating that M^{t_2} is strictly more expressive than M^{t_1}. Note that it is good when M^{t_2} is more expressive than M^{t_1} because M^{t_2} can in principle express more rankings. It can also be problematic, however, because efficient training and the avoidance of overfitting become more challenging.

We first show subsumption by specifying an explicit model transformation, then strictness via a counterexample.

Theorem 1. *For all $r \in \mathbb{N}^+$, M_{2r+1}^{RESCAL} subsumes M_r^{TransE}.*

Proof. Fix some $r \in \mathbb{N}^+$. Pick any TransE model $m_T \in M_r^{\text{TransE}}$, denote by $\boldsymbol{A} \in \mathbb{R}^{N \times r}$ and $\boldsymbol{R} \in \mathbb{R}^{K \times r}$ the corresponding parameter matrices, and by $\boldsymbol{\mathcal{S}}^{m_T}$ the scoring tensor. We show that $\pi(\boldsymbol{\mathcal{S}}^{m_T}) \in \pi(\mathcal{M}_{2r+1}^{\text{RESCAL}})$. We do this by explicitly constructing a corresponding RESCAL model $m_R \in M_{2r+1}^{\text{RESCAL}}$ by specifying its parameters $\boldsymbol{A}' \in \mathbb{R}^{N \times (2r+1)}$ and $\boldsymbol{R}'_k \in \mathbb{R}^{(2r+1) \times (2r+1)}$. Setting

$$\begin{aligned} \boldsymbol{a}'_i &= \begin{pmatrix} \mathbf{1}_r^T & \boldsymbol{a}_i^T & \boldsymbol{a}_i^T \boldsymbol{a}_i \end{pmatrix}^T, \\ \boldsymbol{R}'_k &= - \begin{pmatrix} \mathbf{0}_{r \times r} & -2\,\mathrm{diag}\,(\boldsymbol{r}_k) & \boldsymbol{e}_{1,r} \\ 2\,\mathrm{diag}\,(\boldsymbol{r}_k) & -2\boldsymbol{I}_{r \times r} & 0_{r \times 1} \\ \boldsymbol{e}_{1,r}^T & \mathbf{0}_{1 \times r} & 0 \end{pmatrix}, \end{aligned} \tag{1}$$

we can now verify[2] that

$$s_k^{m_R}(i,j) \leq s_k^{m_R}(i',j') \iff s_k^{m_T}(i,j) \leq s_k^{m_T}(i',j'),$$

which implies that m_T and m_R agree on the ranking for each relation, i.e., $\pi(\boldsymbol{\mathcal{S}}^{m_T}) = \pi(\boldsymbol{\mathcal{S}}^{m_R})$. Since $m_R \in M_{2r+1}^{\text{RESCAL}}$, we obtain $\pi(\boldsymbol{\mathcal{S}}^{m_T}) \in \pi(\mathcal{M}_{2r+1}^{\text{RESCAL}})$ as claimed. □

The proof above shows that TransE can be viewed as a bilinear model with the constraints specified in Eq. (1).

Theorem 2. *M^{TransE} does not subsume M_r^{RESCAL} for any $r \geq 2$.*

Note that the theorem implies that there are RESCAL models with $r = 2$ that cannot be expressed with any TransE model, no matter how large its size.

Proof. Fix some $r \geq 2$ and consider the RESCAL model $m_R \in M_r^{\text{RESCAL}}$ specified by parameters

$$\boldsymbol{a}'_i = \begin{cases} \boldsymbol{e}_{i,r} & \text{for } i \in \{\, 1, 2 \,\} \\ \mathbf{0}_r & \text{otherwise} \end{cases},$$

$$\boldsymbol{R}'_k = \begin{cases} \begin{pmatrix} 1 & 1 & \mathbf{0}_{r-2} \\ 1 & 0 & \mathbf{0}_{r-2} \\ \mathbf{0}_{(r-2)\times 1} & \mathbf{0}_{(r-2)\times 1} & \mathbf{0}_{(r-2)\times(r-2)} \end{pmatrix} & k = 1 \\ \mathbf{0}_{r \times r} & \text{otherw.} \end{cases}$$

We have $s_1^{m_R}(1,1) = 1$, $s_1^{m_R}(2,2) = 0$. Thus $s_1^{m_R}(1,1) \neq s_2^{m_R}(2,2)$ and consequently $\pi_{11}(\boldsymbol{S}_{(1)}^{m_R}) \neq \pi_{22}(\boldsymbol{S}_{(1)}^{m_R})$. Now pick any TransE model $m_T \in \mathcal{M}^{\text{TransE}}$, denote by $\boldsymbol{A}$ and $\boldsymbol{R}$ its parameters, and observe that $s_k^{m_T}(1,1) = s_k^{m_T}(2,2) = -\left\|\boldsymbol{r}_k\right\|_2^2$. Thus $\pi_{11}(\boldsymbol{S}_{(1)}^{m_T}) = \pi_{22}(\boldsymbol{S}_{(1)}^{m_T})$. Since this holds for any TransE model, we conclude that $\pi(\boldsymbol{\mathcal{S}}^{m_R}) \notin \pi(\mathcal{M}^{\text{TransE}})$. □

Nickel, Rosasco, and Poggio (2016) argued that HolE can be viewed as a compressed version of RESCAL and implicitly established the subsumption relationship to RESCAL. We present their argument formally below.

Theorem 3. *M_r^{RESCAL} subsumes M_r^{HolE}.*

Proof. From the definition of HolE, we rewrite

$$\begin{aligned} \boldsymbol{r}_k^T(\boldsymbol{a}_i \star \boldsymbol{a}_j) &= \sum_{t=1}^{d} r_{kt} \sum_{u=1}^{d} a_{iu} a_{j((t+u-2 \bmod r)+1)} \\ &= \boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j, \end{aligned}$$

where $\boldsymbol{R}_k = \begin{pmatrix} r_{k1} & r_{k2} & \cdots & r_{kr} \\ r_{kr} & r_{k1} & \cdots & r_{k(r-1)} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k2} & r_{k3} & \cdots & r_{k1} \end{pmatrix}$. □

Recently, Hayashi and Shimbo (2017) proved that $\mathcal{M}_{2r+1}^{\text{HolE}} \supseteq \mathcal{M}_r^{\text{ComplEx}}$ and $\mathcal{M}_r^{\text{HolE}} \subseteq \mathcal{M}_r^{\text{ComplEx}}$. Putting this together with Th. 3, we obtain:

Corollary 1. M_{2r+1}^{RESCAL} subsumes M_r^{ComplEx}.

Finally, since DISTMULT differs from RESCAL only in that DISTMULT adds a diagonality constraint, we directly obtain:

Theorem 4. *M_r^{RESCAL} subsumes M_r^{DISTMULT}.*

3.2 Universality

We say that class M^t is *universal* if $\pi(\mathcal{M}^t) = \pi(\mathbb{R}^{N \times N \times K})$, i.e., any ranking tensor can be expressed. As with subsumption, universality does by no means imply that a model class is suitable for use in practice. If a model class is not universal, however, care must be taken because certain relations cannot be modeled.

A direct consequence of Th. 2 is:

Corollary 2. M^{TransE} is not universal.

We establish the universality of RESCAL, HolE, and ComplEx next.

[2] A more detailed derivation can be found in the online appendix.

Table 1: Summary of our main results. Each row corresponds to a model of size r. All conditions are sufficient conditions. ? means that no bound other than the universal bound is known.

Model	# Parameters	Universal when $r \geq$	Consistent with $\boldsymbol{\mathcal{B}}$ when $r \geq$	Subsumption of model of size r' when $r \geq$				
				RESCAL	HolE	ComplEx	DISTMULT	TransE
RESCAL	$Nr + Kr^2$	N	$\min\{N, 2\sum_k \operatorname{rrank}(\boldsymbol{B}_k)\}$	r'	r'	$2r' + 1$	r'	$2r' + 1$
HolE	$Nr + Kr$	$2KN + 1$	$2\min\{KN, 2\sum_k \operatorname{rrank}(\boldsymbol{B}_k)\} + 1$	?	r'	$2r' + 1$	$2r' + 1$	?
ComplEx	$2Nr + 2Kr$	KN	$\min\{KN, 2\sum_k \operatorname{rrank}(\boldsymbol{B}_k)\}$	?	r'	r'	r'	?
DISTMULT	$Nr + Kr$	No	No	No	No	No	r'	No
TransE	$Nr + Kr$	No	No	No	No	No	No	r'

Theorem 5. *M_N^{RESCAL} is universal.*

Proof. Pick any ranking tensor $\boldsymbol{\mathcal{P}} \in \pi(\mathbb{R}^{N \times N \times K})$. Consider the model $m \in M_N^{\text{RESCAL}}$ with parameterization $\boldsymbol{A} = \boldsymbol{I}_N$ and $\boldsymbol{R}_k = -\boldsymbol{\mathcal{P}}_{(k)}$. Then $\boldsymbol{S}_k^m = \boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^T = -\boldsymbol{\mathcal{P}}_{(k)}$ and thus $\boldsymbol{\mathcal{S}}^m = -\boldsymbol{\mathcal{P}}$. Using the fact that $\pi(-\boldsymbol{\mathcal{P}}) = \boldsymbol{\mathcal{P}}$, we conclude that $\boldsymbol{\mathcal{P}} \in \pi(\mathcal{M}_N^{\text{RESCAL}})$. □

Note that models in M_N^{RESCAL} have very large embeddings. It is more involved to establish whether or not M_r^{RESCAL} is universal for some $r < N$. We approach this question below and show that r needs to be linear in N to obtain universality.

Theorem 6. *$M_{\lfloor N/32-1 \rfloor}^{\text{RESCAL}}$ is not universal.*

The proof (given below) makes use of the notion of rounding rank (Neumann, Gemulla, and Miettinen 2016). Given a *rounding threshold* $\tau \in \mathbb{R}$, denote by $\operatorname{round}_\tau(x) = I(x \geq \tau)$ the rounding function (1 if $x \geq \tau$, else 0). We apply $\operatorname{round}_\tau$ to matrices and tensors by rounding elementwise. In particular, when $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is any real-valued matrix, then $\operatorname{round}_\tau(\boldsymbol{A})$ is the $m \times n$ binary matrix with $[\operatorname{round}_\tau(\boldsymbol{A})]_{ij} = \operatorname{round}_\tau(a_{ij})$. We assume $\tau = 1/2$ unless explicitly stated otherwise and write round for $\operatorname{round}_{1/2}$.

Definition 1. For $\tau \in \mathbb{R}$, the *rounding rank w.r.t.* τ of a binary matrix $\boldsymbol{B} \in \{0, 1\}^{m \times n}$ is given by $\operatorname{rrank}_\tau(\boldsymbol{B}) = \min\{\operatorname{rank}(\boldsymbol{A}) : \boldsymbol{A} \in \mathbb{R}^{m \times n}, \operatorname{round}_\tau(\boldsymbol{A}) = \boldsymbol{B}\}$.

Given a Boolean matrix $\boldsymbol{B}$, say that a scoring matrix $\boldsymbol{S}$ is *consistent* with $\boldsymbol{B}$ if

$$b_{ij} = 1 \text{ and } b_{i'j'} = 0 \implies \pi_{ij}(\boldsymbol{S}) < \pi_{i'j'}(\boldsymbol{S}). \qquad (2)$$

The rounding rank can be interpreted as the minimum rank of a scoring matrix that is consistent with $\boldsymbol{B}$. Neumann, Gemulla, and Miettinen (2016) proved that the rounding rank differs by at most 1 for different choices of τ and that it is connected to the *sign rank* (Alon, Moran, and Yehudayoff 2016). The rounding rank can be much smaller than the matrix rank in practice, which partially explains the the success of bilinear models.

Proof (of Th. 6). Alon, Frankl, and Rödl (1985) showed that there exist Boolean matrices in $\{0, 1\}^{N \times N}$ with rounding rank at least $N/32$ for every N. Pick any such matrix $\boldsymbol{B}$. The proof is by contradiction. Consider $K = 1$ and suppose there exists a scoring matrix $\boldsymbol{S} \in M_{\lfloor N/32-1 \rfloor}^{\text{RESCAL}}$ that satisfies Eq. (2), i.e., $s_{ij} > s_{i'j'}$ whenever $b_{ij} = 1$ and $b_{i'j'} = 0$. Observe that $\boldsymbol{S}$ has rank at most $\lfloor N/32 - 1 \rfloor$ because it is defined by a product involving a $\lfloor N/32 - 1 \rfloor \times \lfloor N/32 - 1 \rfloor$ matrix. But this implies that $\operatorname{rrank}(\boldsymbol{B}) \leq N/32 - 1$, a contradiction. □

Note that the proof implies that there exists ranking tensors with just two distinct ranks that cannot be expressed by $M_{\lfloor N/32-1 \rfloor}^{\text{RESCAL}}$. Since RESCAL is an unconstrained bilinear model, we can generalize to other model classes.

Corollary 3. No model class that only contains bilinear models of size less than $\frac{N}{32}$ is universal.

Theorem 7. *M_{KN}^{ComplEx} and M_{2KN+1}^{HolE} are universal.*

Proof. Pick any scoring tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{N \times N \times K}$. Trouillon et al. (2016b) showed that for every $N \times N$ real matrix and thus every scoring matrix $\boldsymbol{S}_k$, there exists $\boldsymbol{A}_k, \boldsymbol{D}_k \in \mathbb{C}^{N \times N}$, where $\boldsymbol{A}_k$ is unitary, $\boldsymbol{D}_k$ diagonal, and $\boldsymbol{S}_k = \operatorname{Re}(\boldsymbol{A}_k \boldsymbol{D}_k \boldsymbol{A}_k^*)$. Now consider the ComplEx model with

$$\begin{aligned} \boldsymbol{A} &= \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_K \end{pmatrix} \\ \boldsymbol{R}_k &= \operatorname{diag}\left(\boldsymbol{0}_{N \times N}, \ldots, \boldsymbol{D}_k, \ldots, \boldsymbol{0}_{N \times N}\right) \end{aligned}$$

We can verify $\boldsymbol{S}_k = \operatorname{Re}(\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^*)$ for each k. Thus $\boldsymbol{\mathcal{S}} \in \mathcal{M}_{KN}^{\text{ComplEx}}$ and it follows that M_{KN}^{ComplEx} is universal. The universiality of M_{2KN+1}^{HolE} follows from the fact that $\mathcal{M}_{2r+1}^{\text{HolE}} \supseteq \mathcal{M}_r^{\text{ComplEx}}$ for every r (Hayashi and Shimbo 2017). □

Finally, since DISTMULT's relation matrix is diagonal and thus symmetric, DISTMULT cannot model asymmetric relations.

Theorem 8. *$\mathcal{M}^{\text{DISTMULT}}$ is not universal.*

3.3 Consistent Ranking

Suppose we are given an $N \times N \times K$ Boolean tensor $\boldsymbol{\mathcal{B}}$ and we look for a ranking tensor $\boldsymbol{\mathcal{P}}$ that is consistent with $\boldsymbol{\mathcal{B}}$ in each frontal slice, i.e., $p_{ijk} < p_{i'j'k}$ whenever $b_{ijk} = 1$ and $b_{i'j'k} = 0$. In this section, we establish upper bounds on the size[3] that various bilinear models need to express a ranking that is consistent with $\boldsymbol{\mathcal{B}}$, i.e., which ranks 1s above 0s. Here we think of $\boldsymbol{\mathcal{B}}$ as the correct completed KB; there is no hope for a model class not consistent with $\boldsymbol{\mathcal{B}}$ to recover the correct KB.

Note that even if a model class is not universal, it may still contain consistent models for all Boolean tensors. This is not the case for DISTMULT and TransE, however. In particular, since DISTMULT produces symmetric scoring matrices, DISTMULT does not contain models consistent with any Boolean tensor that has an asymmetric frontal slice.

[3] The expressive power of models considered here is nondecreasing as their size grows.

For TransE, the proof of Th. 2 implies that TransE does not contain models for Boolean tensors with both 0s and 1s on the main diagonal of any of its frontal slices.

Theorem 9. *There exists Boolean tensors $\mathcal{B}$ such that no ranking tensor in $\pi(\mathcal{M}^{\textit{DISTMULT}})$ is consistent with $\mathcal{B}$.*

Theorem 10. *There exists Boolean tensors $\mathcal{B}$ such that no ranking tensor in $\pi(\mathcal{M}^{\textit{TransE}})$ is consistent with $\mathcal{B}$.*

For RESCAL, which is universal, we can make use of the rounding-rank decomposition to obtain a tighter bound than the one implied by its universality.

Theorem 11. *For any boolean tensor $\mathcal{B}$, $\pi(\mathcal{M}_r^{\textit{RESCAL}})$ contains a ranking tensor consistent with $\mathcal{B}$ if*

$$r \geq \min\left\{ N,\ 2\sum_{k=1}^{K} \operatorname{rrank}(\boldsymbol{B}_k) \right\}.$$

Proof. The case $r \geq N$ follows from Th. 6. Denote by r_k the rounding rank of slice $\boldsymbol{B}_k$ of $\mathcal{B}$; we explicitly construct a consistent RESCAL model with $r = 2\sum_k r_k$ (as asserted). To do so, pick any $\boldsymbol{L}_k, \boldsymbol{Q}_k \in \mathbb{R}^{N \times r_k}$ that form a *rounding-rank decomposition* of $\boldsymbol{B}_k$, i.e., for which $\boldsymbol{B}_k = \operatorname{round}(\boldsymbol{L}_k \boldsymbol{Q}_k^T)$. (By the definition of rounding rank, such matrices always exist.) Now set

$$\begin{aligned}
\boldsymbol{a}_i^T &= (\ [\boldsymbol{L}_1]_{i:} \quad [\boldsymbol{Q}_1]_{i:} \quad \cdots \quad [\boldsymbol{L}_K]_{i:} \quad [\boldsymbol{Q}_K]_{i:})^T \\
\boldsymbol{M}_k &= \begin{pmatrix} \mathbf{0}_{r_k \times r_k} & \boldsymbol{I}_{r_k \times r_k} \\ \mathbf{0}_{r_k \times r_k} & \mathbf{0}_{r_k \times r_k} \end{pmatrix} \\
(\boldsymbol{R}_k)_{ij} &= \operatorname{diag}\left(\mathbf{0}_{2r_1 \times 2r_1}, \ldots, \boldsymbol{M}_k, \ldots, \mathbf{0}_{2r_K \times 2r_K}\right)
\end{aligned}$$

We can now verify that $\operatorname{round}(\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^T) = \boldsymbol{B}_k$, which implies consistency. □

Theorem 12. *For any Boolean tensor $\mathcal{B}$, $\pi(\mathcal{M}_r^{\textit{ComplEx}})$ contains a ranking tensor consistent with $\mathcal{B}$ if*

$$r \geq \min\left\{ KN, 2\sum_{k=1}^{K} \operatorname{rrank}(\boldsymbol{B}_k) \right\}.$$

Proof. The case $r \geq KN$ follows directly from Th. 7. To obtain $r \geq 2\sum_{k=1}^{K} \operatorname{rrank}(\boldsymbol{B}_k)$, define r_k, $\boldsymbol{L}_k$, and $\boldsymbol{Q}_k$ as in the proof of Th. 11, and set $\boldsymbol{S}_k = \boldsymbol{L}_k \boldsymbol{Q}_k^T$. Then there exist matrices $\boldsymbol{A}_k \in \mathbb{C}^{N \times 2r_k}$ and $\boldsymbol{D}_k \in \mathbb{C}^{2r_k \times 2r_k}$, with $\boldsymbol{D}_k$ being diagonal, such that $\boldsymbol{S}_k = \operatorname{Re}(\boldsymbol{A}_k \boldsymbol{D}_k \boldsymbol{A}_k^*)$ (Trouillon et al. 2016b). Now define

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{A}_1 \quad \boldsymbol{A}_2 \quad \cdots \quad \boldsymbol{A}_K) \\
\boldsymbol{R}_k &= \operatorname{diag}\left(\mathbf{0}_{2r_1 \times 2r_1}, \ldots, \boldsymbol{D}_k, \ldots, \mathbf{0}_{2r_K \times 2r_K}\right)
\end{aligned}$$

and observe that $\boldsymbol{S}_k = \operatorname{Re}(\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^*)$. □

As a corollary of the above theorem, we have:

Corollary 4. For any Boolean tensor $\mathcal{B}$, $\pi(\mathcal{M}_r^{\text{HolE}})$ contains a ranking tensor consistent with $\mathcal{B}$ if

$$r \geq \min\left\{ 2KN + 1, 4\sum_{k=1}^{K} \operatorname{rrank}(\boldsymbol{B}_k) + 1 \right\}.$$

4 Training and Relation-Level Ensemble

We have seen that various prior models can be interpreted as a bilinear models subject to certain constraints. In other words, they are diverse with respect to their expressivity. So far, we did not touch on how to select a suitable model for a given dataset and from a given model class. In this section, we briefly discuss model training in a margin-based framework. We then propose a simple relation-level ensemble that combines multiple individual models. The rationale behind using an ensemble is that whether a model class can represent well or be trained well on a relation depends on properties of that relation. The ensemble thus aims to pick the best model (or a combination of models) for each relation.

4.1 Margin-Based Training

We assume throughout that we are given a set of positive triples $\mathcal{T}^+ \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, but no negative evidence. This is a common scenario in practice. To deal with the absence of negative evidence, ranking-based frameworks aim to produce a model that ranks triples in $\mathcal{T}^+$ higher than other triples. A common approach (Bordes et al. 2013; Nickel, Rosasco, and Poggio 2016) is to define a set of "negative" triples for each positive triple $(i,k,j) \in \mathcal{T}^+$ by perturbing subject or object:

$$\begin{aligned}
\mathcal{T}^-_{(i,k,j)} = &\left\{ (i',k,j) \mid i' \in \mathcal{E},\ (i',k,j) \notin \mathcal{T}^+ \right\} \cup \\
&\left\{ (i,k,j') \mid j' \in \mathcal{E},\ (i,k,j') \notin \mathcal{T}^+ \right\}.
\end{aligned}$$

This approach corresponds to a local closed-world assumption (Dong et al. 2014). We now briefly summarize a common margin-based framework for training (Bordes et al. 2013). There are a number of alternatives, including logistic loss (Riedel et al. 2013) and negative log-likelihood (Trouillon et al. 2016a). Margin-based frameworks often lead to faster training times in practice because they focus on "informative" pairs of positive and negative triples, i.e., they ignore parts of the data that are already more or less well-represented by the model. In particular, we minimize

$$\sum_{\substack{(i^+,k,j^+) \in \mathcal{T}^+, \\ (i^-,k,j^-) \in \mathcal{T}^-_{(i^+,k,j^+)}}} \frac{[f(i^-,k,j^-) + \gamma - f(i^+,k,j^+)]_+}{|\mathcal{T}^-_{(i^+,k,j^+)}|},$$

where $0 \leq \gamma \in \mathbb{R}^+$ is a *margin hyperparameter*, $[x]_+ = \max(0,x)$, and f depends on the model being trained. For all models but HolE, we set $f(i,k,j) = s_k^m(i,j)$. For HolE, we set $f(i,k,j) = \sigma(s_k^m(i,j))$, where σ denotes the logistic function, as suggested by the authors. In our experimental study, we also consider an additional L_2 regularization term over the model parameters. The models can be fit using stochastic gradient descent (SGD) as in (Bordes et al. 2013; Lin et al. 2015b). The computational cost per SGD step of RESCAL is $O(r^2)$, of HolE $O(r \log n)$, and of all other models $O(r)$.

4.2 Relation-Level Ensemble

The simplest way of combining multiple models is to construct an ensemble at the model level (Krompaß and Tresp 2015). Our experimental study suggests that the relative performance of different models is relation-dependent, however.

Table 2: Dataset statistics

Dataset	# Ent.	# Rel.	# Train.	# Valid.	# Test
WN18	40,943	18	141,442	5,000	5,000
FB15K	14,951	1,345	483,142	50,000	59,071

A more promising approach is therefore to combine models at the relation level. To the best of our knowledge, this simple approach has not been explored previously.

Our ensemble is based on stacking. A meta learner is used to combine the ranking matrices produced by the individual models such that some accuracy measure is maximized. Here we use logistic regression. To do so, we construct for each relation a dataset that contains all of its positive triples as well as an equal amount of negative triples obtained by randomly perturbing each positive triple following the same strategy as in training individual models. For logistic regression, we use rescaled scores of the individual models as features and the positive/negative class label as response variable. Rescaling accounts for the variety in range of scores of different models; we rescale each feature linearly into range $[0, 1]$ (Han, Kamber, and Pei 2011, Sec. 3.5.2).

5 Experiments

We conducted an experimental study on two real-world datasets, which are commonly used in prior work on KB completion. The primary goal of our study was to provide independent evidence for the performance of various bilinear models under the margin-based ranking framework. We also evaluated relation-level ensembles of such models and compared the results to prior results reported in the literature (for bilinear and other models).

5.1 Experimental Setup

All datasets, experimental results, and source code will be made publicly available.[4]

Data. We used the WN18 (Bordes et al. 2014) and FB15K (Bordes et al. 2013) datasets, which were extracted from WordNet (Miller 1995) and Freebase (Bollacker et al. 2008), respectively. WordNet contains words and their relationships. Freebase contains various facts across a large number of relations. The two datasets are presplit into a training set, a validation set, and a test set. Table 2 summarizes the key statistics.

Methods and training. We considered RESCAL (R), HolE (H), and TransE (T) in our experimental study. We reimplemented each method in C++, partly using the Intel Math Kernel Library. We trained each model in the margin-based ranking framework using Adagrad (Duchi, Hazan, and Singer 2011). In each step, we sampled a positive triple at random and obtained a negative triple by randomly perturbing subject or object. Sampling was done without replacement and we did not use mini-batches. When using the same hyperparameters, our implementation provided similar or better fits than the original implementations provided by the authors. Note that our study was limited in that we considered only one particular training method; no conclusions can be drawn about other training methods. We focused on margin-based ranking because it led to much faster training times, making this study more feasible. We used LIBLINEAR for logistic regression.

Table 3: Hyperparameters settings used in our study

Dataset	Model	r	γ	η	λ_e	λ_r
WN18	RESCAL	200	1.0	0.10	0.10	0.01
	HolE	200	0.2	0.10	0.01	0.00
	TransE	200	0.5	0.01	-	-
FK15k	RESCAL	200	4.0	0.10	0.10	0.01
	HolE	200	0.2	0.10	0.01	0.01
	TransE	200	0.2	0.01	-	-

Evaluation. We evaluated model performance for the tasks of entity ranking and triple classification on the test data. In *entity ranking*, we rank entities for queries of the form $R(?, e)$ or $R(e, ?)$. Our evaluation closely follows Bordes et al. (2013), and we report *mean reciprocal rank (MRR)*, *HITS@10*, and *mean rank (MR)* in the *filtered* setting, i.e., predictions that correspond to tuples in the training or validation datasets were discarded. In *triple classification*, we are given a triple (i, k, j) and are asked to classify it as positive or negative; we proceed as Socher et al. (2013) to produce the set of tuples to classify. To perform classification, we determined a score threshold σ_k for each relation and model; scores larger than σ_k were classified positive, else negative. We used optimal thresholds with respect to the validation set.

Model selection. Each of the models has a number of hyperparameters. For all models, we trained the models solely on the training data and used the validation data solely to tune hyperparameters. Test data was not touched for model selection. We considered the following hyperparameter settings: $r \in \{100, 200\}$, learning rate $\eta \in \{0.01, 0.1, 1\}$, weight of L2-regularization $\lambda_e, \lambda_r \in \{0, 0.1, 0.01\}$ for entity and relation parameters, resp., margin hyperparameter $\gamma \in \{1, 2, 4, 8\}$ for RESCAL, $\gamma \in \{0.2, 0.5, 0.7\}$ for HolE, and $\gamma \in \{0.2, 0.5, 0.7, 1.0, 1.5\}$ for TransE.[5]

We performed exhaustive grid search, using 50 (2000 for TransE) epochs (passes over the dataset) per hyperparameter setting and model. We then retrained the best-performing setting (w.r.t. HITS@10 on validation data) for each model on the training data for up to 2,000 epochs. Tab. 3 reports the hyperparameters ultimately selected.

5.2 Results

Entity ranking. Our results are summarized in Tab. 4. Detailed results can be found in Tab. 6, where we measured HITS@10 per relation category and per argument to be predicted as in (Bordes et al. 2013).

For the individual models, our results indicate that model

[4] `http://dws.informatik.uni-mannheim.de/en/resources/software/tf/`

[5] We used smaller margins than the ones suggested for TransE with L_1 distance (Lin et al. 2015b; Wang et al. 2014; Lin et al. 2015a). By doing this, we obtained comparable prediction performance as TransE-L_1.

Table 4: Entity ranking results of our experimental study. Best-performing entries marked bold.

Dataset	WN18			FB15K		
Model	HITS@10 (%)	MRR (%)	MR	HITS@10 (%)	MRR (%)	MR
HolE (Nickel, Rosasco, and Poggio 2016)	94.1	93.8	819	72.6	50.2	331
TransE (Bordes et al. 2013)	94.5	43.9	**474**	79.5	34.4	76
RESCAL (Nickel, Tresp, and Kriegel 2011)	87.8	79.9	905	59.6	38.1	247
RESCAL + TransE	94.8	87.3	510	79.7	51.1	61
RESCAL + HolE	94.4	**94.0**	743	79.1	57.5	165
HolE + TransE	94.9	93.8	507	84.6	61.0	67
RESCAL + HolE + TransE	**95.0**	**94.0**	507	**85.1**	**62.8**	**52**

Table 5: Entity ranking results as reported in the literature (not reproduced here, partly with different training methods, partly non-bilinear models). Entries marked "-" were not reported. Entries better than any result in our study are marked bold.

Dataset	WN18			FB15K		
Model	HITS@10 (%)	MRR (%)	MR	HITS@10 (%)	MRR (%)	MR
Gaifman (Niepert 2016)	93.9	-	**352**	84.2	-	75
ComplEx (Trouillon et al. 2016a), r=150/200	94.7	**94.1**	-	84.0	**69.2**	-
DISTMULT (Trouillon et al. 2016a), r=150/200	93.6	82.2	902	82.4	**65.4**	97
R-GCN+DISTMULT (Schlichtkrull et al. 2017), r=200	**96.4**	81.9	-	84.2	**69.6**	-
ANALOGY (Liu, Wu, and Yang 2017), r=200	94.7	**94.2**	-	**85.4**	**72.5**	-

Table 6: Detailed entity ranking results (FB15k, HITS@10)

Task	Predict subject				Predict object			
Relations	1:1	1:N	N:1	N:N	1:1	1:N	N:1	N:N
TransE	75.8	91.9	41.4	82.2	75.5	51.1	91.9	84.7
HolE	80.4	69.5	44.7	77.4	79.0	57.8	59.1	79.0
RESCAL	43.1	75.7	17.7	62.0	42.4	21.3	79.2	65.8
R+H+T	**87.5**	**94.3**	**55.2**	**86.7**	**87.0**	**65.0**	**93.3**	**89.4**

Table 7: Triple classification results (FB15K)

Model	T	H	R	R+T	R+H	H+T	R+H+T
Accuracy	96.2	93.7	94.6	96.7	95.8	96.5	**96.9**

performance depends on the relation category. No single model always performed best across all categories. HolE and TransE generally performed better than RESCAL; here constraints help. The relation-level ensembles generally improved performance w.r.t. HITS@10 and MRR. Performance of MR was not improved, however, mainly because this metric is sensitive to low-ranked triples (which existed in HolE and RESCAL predictions). Note that adding RESCAL to the ensemble was helpful. Finally, the ensemble of RESCAL, TransE, and HolE performed best w.r.t. HITS@10 on all relation categories and for both datasets.

In Tab. 5, we compare to some recent results reported in the literature. Note that training methods were different than the one used in our study for some of these models, and that some models are not bilinear. Nevertheless, a direct comparison indicates that a relation-level ensemble of multiple bilinear models is competitive to the state-of-the-art.

Triple classification. Tab. 7 summarizes the HITS@10 performance of each individual model and various relation-level ensembles for triple classification on FB15k. The results are generally in line with the results for entity ranking. A notable exception is that RESCAL outperforms HolE here; we conjecture that this is due to HolE's high MR on this dataset.

6 Related Work

We focus on recent embedding models that solely use the KB as input. There are a number of methods that modify TransE in one way or another: TransH (Wang et al. 2014) and TransR (Lin et al. 2015b) improve support symmetric and many-to-one relations, TransG (Xiao et al. 2015) adds refines relation embeddings by semantic components, and PTransE (Lin et al. 2015a) adds multiple-step relation paths. Gaifman (Niepert 2016) exploits structural features in the form of Horn clauses to construct embeddings. Socher et al. (2013) combined neural networks with tensors. Schlichtkrull et al. (2017) models relational data with graph convolutional networks. ANALOGY (Liu, Wu, and Yang 2017) is a recent bilinear model that constrains relation embeddings be real normal matrices. Finally, Nickel, Jiang, and Tresp (2014) provided a rank bound for exact recovery of a Boolean tensor with RESCAL. Our results differ in that we consider consistency, not exact recovery.

7 Conclusion

We studied the expressive power of and subsumption relationships between recent bilinear embedding models for knowledge graphs. We introduced the concepts of universality and consistency, which capture different aspects of model expressiveness, and provided bounds on model sizes needed for universality or consistency with a given dataset. We argued that using a relation-level ensembles are beneficial for multi-relational learning. Finally, we conducted an independent experimental study that compared various bilinear models in a common setup.

Future work includes tightening the bounds provided here, studying which relation types can be reportedresented by which models, and exploring the relationship between additional models. We also expect an in-depth study of model performance with various alternative datasets and training methods to be insightful.

References

Alon, N.; Frankl, P.; and Rödl, V. 1985. Geometrical realization of set systems and probabilistic communication complexity. In *FOCS*, 277–280.

Alon, N.; Moran, S.; and Yehudayoff, A. 2016. Sign rank versus vc dimension. In *COLT*, 47–80.

Bollacker, K. D.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 1247–1250.

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.

Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2014. A semantic matching energy function for learning with multi-relational data - application to word-sense disambiguation. *Machine Learning* 94(2):233–259.

Carroll, J. D., and Chang, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika* 35(3):283–319.

Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *KDD*, 601–610.

Duchi, J. C.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159.

Han, J.; Kamber, M.; and Pei, J. 2011. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann.

Hayashi, K., and Shimbo, M. 2017. On the equivalence of holographic and complex embeddings for link prediction. In *ACL (2)*, 554–559.

Krompaß, D., and Tresp, V. 2015. Ensemble solutions for link-prediction in knowledge graphs. In *LD4KD*.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; and Bizer, C. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6(2):167–195.

Lin, Y.; Liu, Z.; Luan, H.; Sun, M.; Rao, S.; and Liu, S. 2015a. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, 705–714.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2181–2187.

Liu, H.; Wu, Y.; and Yang, Y. 2017. Analogical inference for multi-relational embeddings. In *ICML*, volume 70, 2168–2178.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Neumann, S.; Gemulla, R.; and Miettinen, P. 2016. What you will gain by rounding: Theory and algorithms for rounding rank. In *ICDM*, 380–389.

Nickel, M.; Murphy, K.; Tresp, V.; and Gabrilovich, E. 2016. A review of relational machine learning for knowledge graphs. *IEEE Computer Society* 104(1):11–33.

Nickel, M.; Jiang, X.; and Tresp, V. 2014. Reducing the rank of relational factorization models by including observable patterns. In *NIPS*, 1179–1187.

Nickel, M.; Rosasco, L.; and Poggio, T. A. 2016. Holographic embeddings of knowledge graphs. In *AAAI*, 1955–1961.

Nickel, M.; Tresp, V.; and Kriegel, H. 2011. A three-way model for collective learning on multi-relational data. In *ICML*, 809–816.

Niepert, M. 2016. Discriminative gaifman models. In *NIPS*, 3405–3413.

Rebele, T.; Suchanek, F. M.; Hoffart, J.; Biega, J.; Kuzey, E.; and Weikum, G. 2016. YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In *ISWC*, LNCS, 9982:177–185.

Riedel, S.; Yao, L.; McCallum, A.; and Marlin, B. M. 2013. Relation extraction with matrix factorization and universal schemas. In *HLT-NAACL*, 74–84.

Schlichtkrull, M. S.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2017. Modeling relational data with graph convolutional networks. *CoRR* abs/1703.06103.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. Y. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 926–934.

Trouillon, T., and Nickel, M. 2017. Complex and holographic embeddings of knowledge graphs: A comparison. *CoRR* abs/1707.01475.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016a. Complex embeddings for simple link prediction. In *ICML*, 2071–2080.

Trouillon, T.; Dance, C. R.; Gaussier, E.; and Bouchard, G. 2016b. Decomposing real square matrices via unitary diagonalization. *CoRR*.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 1112–1119.

Xiao, H.; Huang, M.; Hao, Y.; and Zhu, X. 2015. Transg : A generative mixture model for knowledge graph embedding. *CoRR* abs/1509.05488.

Yang, B.; Yih, W.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *CoRR* abs/1412.6575.