
KBLRN: End-to-End Learning of Knowledge Base Representations with Latent, Relational, and Numerical Features

Alberto García-Durán
NEC Labs Europe
Heidelberg, Germany
alberto.duran@neclab.eu

Mathias Niepert
NEC Labs Europe
Heidelberg, Germany
mathias.niepert@neclab.eu

Abstract

We present KBLRN, a framework for end-to-end learning of knowledge base representations from latent, relational, and numerical features. KBLRN integrates feature types with a novel combination of neural representation learning and probabilistic product of experts models. To the best of our knowledge, KBLRN is the first approach that learns representations of knowledge bases by integrating latent, relational, and numerical features. We show that instances of KBLRN outperform existing methods on a range of knowledge base completion tasks. We contribute a novel data set enriching commonly used knowledge base completion benchmarks with numerical features. The data sets are available under a permissive BSD-3 license¹. We also investigate the impact numerical features have on the KB completion performance of KBLRN.

1 INTRODUCTION

The importance of knowledge bases (KBs) for AI systems has been demonstrated in numerous application domains. KBs provide ways to organize, manage, and retrieve structured data and allow AI system to perform reasoning. In recent years, KBs have been playing an increasingly crucial role in AI applications. Purely logical representations of knowledge bases have a long history in AI [27]. However, they suffer from being inefficient and brittle. Inefficient because the computational complexity of reasoning is exponential in the worst case and, therefore, the time required by a reasoner highly unpredictable. Brittle because a purely logical KB requires a large set of logical rules that are handcrafted and/or

¹<https://github.com/nle-ml/mmkb>

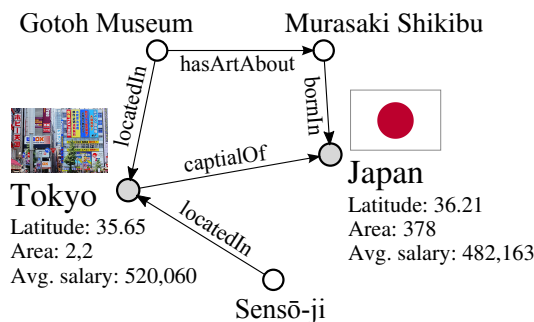


Figure 1: A small part of a knowledge base.

mined. These problems are even more pressing in applications whose environments are changing over time.

Motivated by these shortcomings, there has been a flurry of work on combining logical and statistical approaches to build systems capable of reasoning over and learning from incomplete structured data. Most notably, the statistical relational learning community has proposed numerous formalisms that combine logic and probability [24]. These formalisms are able to address the learning problem and make the resulting AI systems more robust to missing data and missing rules. Intuitively, logical formulas act as relational features and the probability of a possible world is determined by a sufficient statistic for the values of these features. These approaches, however, are in most cases even less efficient because logical inference is substituted with probabilistic inference.

More recently, the research community has focused on efficient machine learning models that perform well on restricted tasks such as link prediction in KBs. Examples are knowledge base factorization and embedding approaches [3, 21, 11, 20] and random-walk based ML models [15, 7]. The former learn latent features for the entities and relations in the knowledge base and use those to perform link prediction. The latter explore specific re-

lational features such as path types between two entities and train a machine learning model for link prediction.

With this work, we propose KBLRN, a novel approach to combining relational, latent (learned), and numerical features, that is, features that can take on a large or infinite number of real values. The combination of various features types is achieved by integrating embedding-based learning with probabilistic models in two ways. First, we show that modeling numerical features with radial basis functions is beneficial and can be integrated in an end-to-end differentiable learning system. Second, we propose a probabilistic product of experts (PoE) [13] approach to combine the feature types. Instead of training the PoE with contrastive divergence, we approximate the partition function with a negative sampling strategy. The PoE approach has the advantage of being able to train the model jointly and end-to-end.

The paper is organized as follows. First, we discuss relational, latent, and numerical features. Second, we describe KBLRN. Third, we present empirical evidence that instances of KBLRN outperform state of the art methods for KB completion. We also investigate in detail under what conditions numerical features are beneficial.

2 RELATIONAL, LATENT, AND NUMERICAL FEATURES

We assume that the facts of a knowledge base (KB) are given as a set of triples of the form (h, r, t) where h and t are the head and tail entities and r is a relation type. Figure 1 depicts a small fragment of a KB with relations and numerical features. KB completion is the problem of answering queries of the form $(?, r, t)$ or $(h, r, ?)$. While the proposed approach can be generalized to more complex queries, we focus on completion queries for the sake of simplicity. We now discuss the three feature types used in KBLRN and motivate their utility for knowledge base completion. How exactly we extract features from a given KB is described in the experimental section.

2.1 Relational Features

Each relational feature is given as a logical formula which is evaluated in the KB to determine the feature’s value. For instance, the formula $\exists x (A, \text{bornIn}, x) \wedge (x, \text{capitalOf}, B)$ corresponds to a binary feature which is 1 if there exists a path of that type from entity A to entity B, and 0 otherwise. These features are often used in relational models [30, 22] and random-walk based models such as PRA and SFE [15, 7]. In this work, we use relational paths of length one and two and use the rule mining approach AMIE+ [5]. We detail the generation of

the relational features in the experimental section. For a pair of entities (h, t) , we denote the feature vector computed based on a set of relational features by $r_{(h,t)}$.

2.2 Latent Features

Numerous embedding methods for KBs have been proposed in recent years [21, 3, 11, 20]. Embedding methods provide fixed-size vector representations (embeddings) for all entities in the KB. In the simplest of cases, relations are modeled as translations in the entity embedding space [3]. We incorporate typical embedding learning objectives into KBLRN and write e_h and e_t to refer to an embedding of a head entity and a tail entity, respectively. The advantages of latent feature models are their computational efficiency and their ability to learn latent entity types suitable for downstream ML tasks without hand-crafted or mined logical rules.

2.3 Numerical Features

Numerical features are entity features whose values can take on a very large or infinite number of real values. To the best of our knowledge, there does not exist a principled approach that integrates numerical features into a relational ML model for KB completion. This is surprising, considering that numerical data is available in almost all existing large-scale KBs. The assumption that numerical data is helpful for KB completion tasks is reasonable. For several relations types the *differences* between the *head* and *tail* are characteristic of the relation itself. For example, while the mean difference of birth years is 0.4 for the Freebase relation `/people/marriage/spouse`, it is 32.4 for the relation `/person/children`. These observations motivate specifically the use of *differences* of numerical feature values. Taking the difference has the advantage that even if a numerical feature is not distributed according to a normal distribution (e.g., birth years in a KB), the difference is often normally distributed. This is important as we need to fit simple parametric distributions to the sparse numerical data. We detail the fully automated extraction and generation of the numerical features in the experimental section.

3 KBLRN: LEARNING END-TO-END JOINT REPRESENTATIONS FOR KNOWLEDGE BASES

With KBLRN we aim to provide a framework for end-to-end learning of KB representations. Since we want to combine different feature types (relational, latent or learned, and numerical) we need to find a suitable

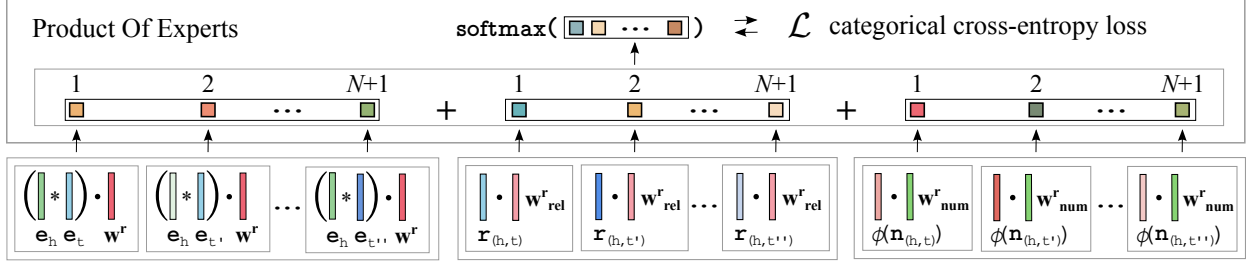


Figure 2: An illustration of an instance of KBLRN implemented with standard deep learning framework components. For every relation type, there is a separate expert for each of the different feature types. The entities τ' and τ'' are two of N randomly sampled entities. The scores of the various submodels are added and normalized with a softmax function. A categorical cross-entropy loss is applied to the normalized scores.

method for integrating the respective submodels, one per feature type. We propose a product of experts (PoE) approach where one expert is trained for each (relation type, feature type) pair. We extend the product of experts approach in two novel ways. First, we create dependencies between the experts by sharing the parameters of the entity embedding model across relation types. By doing this, we combine a probabilistic model with a model that learns vector representations from discrete and numerical data. Second, while product of experts are commonly trained with contrastive divergence [13], we train it with negative sampling and a cross-entropy loss.

In general, a PoE's probability distribution is

$$p(\mathbf{d} \mid \theta_1, \dots, \theta_n) = \frac{\prod_m f_m(\mathbf{d} \mid \theta_m)}{\sum_{\mathbf{c}} \prod_m f_m(\mathbf{c} \mid \theta_m)},$$

where \mathbf{d} is a data vector in a discrete space, θ_m are the parameters of individual model m , $f_m(\mathbf{d} \mid \theta_m)$ is the value of d under model m , and the \mathbf{c} 's index all possible vectors in the data space. The PoE model is now trained to assign high probability to observed data vectors.

In the KB context, the data vector \mathbf{d} is always a triple $\mathbf{d} = (\mathbf{h}, \mathbf{r}, \mathbf{t})$ and the objective is to learn a PoE that assigns high probability to true triples and low probabilities to triples assumed to be false. If $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ holds in the KB, the pair's vector representations are used as positive training examples. Let $\mathbf{d} = (\mathbf{h}, \mathbf{r}, \mathbf{t})$. We can now define one individual expert $f_{(\mathbf{r}, \mathbf{F})}(\mathbf{d} \mid \phi_{(\mathbf{r}, \mathbf{F})})$ for each (relation type \mathbf{r} , feature type \mathbf{F}) pair.

The specific experts we utilize here are based on simple linear models and the DistMult embedding method.

$$\begin{aligned} f_{(\mathbf{r}, \mathbf{L})}(\mathbf{d} \mid \theta_{(\mathbf{r}, \mathbf{L})}) &= \exp((\mathbf{e}_h * \mathbf{e}_t) \cdot \mathbf{w}^{\mathbf{r}}) \\ f_{(\mathbf{r}, \mathbf{R})}(\mathbf{d} \mid \theta_{(\mathbf{r}, \mathbf{R})}) &= \exp(\mathbf{r}_{(\mathbf{h}, \mathbf{t})} \cdot \mathbf{w}_{\text{rel}}^{\mathbf{r}}) \\ f_{(\mathbf{r}, \mathbf{N})}(\mathbf{d} \mid \theta_{(\mathbf{r}, \mathbf{N})}) &= \exp(\phi(\mathbf{n}_{(\mathbf{h}, \mathbf{t})}) \cdot \mathbf{w}_{\text{num}}^{\mathbf{r}}) \text{ and} \\ f_{(\mathbf{r}', \mathbf{F})}(\mathbf{d} \mid \theta_{(\mathbf{r}', \mathbf{F})}) &= 1 \text{ for all } \mathbf{r}' \neq \mathbf{r} \text{ and } \mathbf{F} \in \{\mathbf{L}, \mathbf{R}, \mathbf{N}\}, \end{aligned}$$

where $*$ is the element-wise product, \cdot is the dot product, $\mathbf{w}^{\mathbf{r}}$, $\mathbf{w}_{\text{rel}}^{\mathbf{r}}$, $\mathbf{w}_{\text{num}}^{\mathbf{r}}$ are the parameter vectors for the latent, relational, and numerical features corresponding to the relation \mathbf{r} , and ϕ is the radial basis function (RBF) applied element-wise to $\mathbf{n}_{(\mathbf{h}, \mathbf{t})}$. $f_{(\mathbf{r}, \mathbf{L})}(\mathbf{d} \mid \theta_{(\mathbf{r}, \mathbf{L})})$ is equivalent to the exponential of the DISTMULT [35] scoring function but with KBLRN we can use any of the existing KB embedding scoring functions.

The probability for triple $\mathbf{d} = (\mathbf{h}, \mathbf{r}, \mathbf{t})$ of the PoE model is now

$$p(\mathbf{d} \mid \theta_1, \dots, \theta_n) = \frac{\prod_{\mathbf{F} \in \{\mathbf{R}, \mathbf{L}, \mathbf{N}\}} f_{(\mathbf{r}, \mathbf{F})}(\mathbf{d} \mid \theta_{(\mathbf{r}, \mathbf{F})})}{\sum_{\mathbf{c}} \prod_{\mathbf{F} \in \{\mathbf{R}, \mathbf{L}, \mathbf{N}\}} f_{(\mathbf{r}, \mathbf{F})}(\mathbf{c} \mid \theta_{(\mathbf{r}, \mathbf{F})})},$$

where \mathbf{c} indexes all possible triples.

For numerical features, an activation function should fire when the difference of values is in a specific *range*. For example, we want the activation to be high when the difference of the birth years between a parent and its child is close to 32.4 years. Commonly used activation functions such as sigmoid or tanh are not suitable here, since they saturate whenever they exceed a certain threshold. For each relation \mathbf{r} and the d_n corresponding relevant numerical features, therefore, we apply a radial basis function over the differences of values $\phi(\mathbf{n}_{(\mathbf{h}, \mathbf{t})}) = [\phi(\mathbf{n}_{(\mathbf{h}, \mathbf{t})}^{(1)}), \dots, \phi(\mathbf{n}_{(\mathbf{h}, \mathbf{t})}^{(d_n)})]$, where

$$\phi(\mathbf{n}_{(\mathbf{h}, \mathbf{t})}^{(i)}) = \exp\left(\frac{-\|\mathbf{n}_{(\mathbf{h}, \mathbf{t})}^{(i)} - c_i\|_2^2}{\sigma_i^2}\right).$$

This results in the RBF kernel being activated whenever the difference of values is close to the expected value c_i . We discuss and evaluate several alternative strategies for incorporating numerical features in the experimental section.

3.1 Learning

Product of experts are usually trained with contrastive divergence (CD) [13] which relies on an approximation of the gradient of the log-likelihood using a short Markov chain started at the current seen example. The advantage of CD is that the partition function, that is, the denominator of the probability p , which is intractable to compute, does not need to be approximated. Due to the parameterization of the PoE we have defined here, however, it is not possible to perform CD since there is no way to sample a hidden state given a triple \mathbf{d} . Hence, instead of using CD, we approximate the partition function by performing negative sampling.

The logarithmic loss for the given training triples \mathbf{T} is defined as

$$\mathcal{L} = - \sum_{\mathbf{t} \in \mathbf{T}} \log p(\mathbf{t} | \theta_1, \dots, \theta_n).$$

To fit the PoE to the training triples, we follow the derivative of the log likelihood of each observed triple $\mathbf{d} \in \mathbf{T}$ under the PoE

$$\frac{\partial \log p(\mathbf{d} | \theta_1, \dots, \theta_n)}{\partial \theta_m} = \frac{\partial \log f_m(\mathbf{d} | \theta_m)}{\partial \theta_m} - \frac{\partial \log \sum_c \prod_m f_m(\mathbf{c} | \theta_m)}{\partial \theta_m}$$

Now, to approximate the intractable second term of the right hand side of the above equation, we generate for each triple $\mathbf{d} = (\mathbf{h}, \mathbf{r}, \mathbf{t})$ a set \mathbf{E} consisting of N triples $(\mathbf{h}, \mathbf{r}, \mathbf{t}')$ by sampling exactly N entities \mathbf{t}' uniformly at random from the set of all entities. The term

$$\frac{\partial \log \sum_c \prod_m f_m(\mathbf{c} | \theta_m)}{\partial \theta_m}$$

is then approximated by the term

$$\frac{\partial \log \sum_{\mathbf{c} \in \mathbf{E}} \prod_m f_m(\mathbf{c} | \theta_m)}{\partial \theta_m}.$$

Analogously for the head of the triple. This is often referred to as negative sampling. Figure 2 illustrates the KBLRN framework.

4 RELATED WORK

A combination of latent and relational features has been explored by Toutanova et al. [30, 31]. There, a weighted combination of two independently learned models, a latent feature model [35] and a model fed with a binary vector reflecting the presence of paths of length one between the *head* and *tail*, is proposed. These simple relational features aim at capturing association

strengths between pairs of relationships (e.g. `contains` and `contained_by`). Riedel et al. [25] proposed a method that learns implicit associations between pairs of relations in addition to a latent feature model in the context of relation extraction. Gardner et al. [8] modifies the path ranking algorithm (PRA) [15] to incorporate latent representations into models based on random walks in KBs. Gardner et al. [7] extracted relational features other than paths to better capture entity type information. There are a number of recent approaches that combine relational and latent representations by incorporating known logical rules into the embedding learning formulation [26, 10, 18]. Despite its simplicity, KBLRN’s combination of relational and latent representations significantly outperforms all these approaches.

There exists additional work on combining various types of KB features. Nickel et al. [19] proposed a modification of the well-known tensor factorization method RESCAL [21], called ARE, which adds a learnable matrix that weighs a set of metrics (e.g. Common Neighbors) between pairs of entities; Garcia-Duran et al. [6] proposed a combination of latent features, aiming to take advantage of the different interaction patterns between the elements of a triple. The integration of different feature types into relational machine learning models has been previously addressed [1] [17], but not in the context of link prediction in multi-relational graphs.

KBLRN is different to these approaches in that (i) we incorporate numerical features for KB completion, (ii) we propose a unifying end-to-end learning framework that integrates arbitrary relational, latent, and numerical features.

More recent work [23] combines numerical, visual, and textual features by learning feature type specific encoders and using the vector representations in an off-the-shelf scoring function such as DISTMULT. In contrast to this approach, KBLRN combines experts that are specialized to a specific feature type with a product of expert approach. Moreover, by taking the *difference* between numerical features and explicitly modeling relational features that hold *between* head and tail entities, KBLRN incorporates dependencies between modalities of the head and tail entities. These dependencies cannot be captured with a model that only includes modalities of either the head or the tail entity but not both at the same time.

5 EXPERIMENTS

We conducted experiments on six different knowledge base completion data sets. Primarily, we wanted to understand for what type of relations numerical features are helpful and what input representation of numerical fea-

Data set	FB15k	FB15k-num	FB15k-237	FB15k-237-num	WN18	FB122
Entities	14,951	14,951	14,541	14,541	40,943	9,738
Relation types	1,345	1,345	237	237	18	122
Training triples	483,142	483,142	272,115	272,115	141,442	91,638
Validation triples	50,000	5,156	17,535	1,058	5,000	9,595
Test triples	59,071	6,012	20,466	1,215	5,000	11,243
Relational features	90,318	90,318	7,834	7,834	14	47

Table 1: Statistics of the data sets.

tures achieves the best results. An additional objective was the comparison to state of the art methods.

5.1 Datasets

We conducted experiments on six different data sets: FB15k, FB15k-237, FB15k-num, FB15k-237-num, WN18, and FB122. FB15k [3] and Wordnet (WN) [2] are knowledge base completion data sets commonly used in the literature. The FB15k data set is a representative subset of the Freebase knowledge base. WN18 represents lexical relations between word senses. The two data sets are being increasingly criticized for the frequent occurrence of reverse relations causing simple relational baselines to outperform most embedding-based methods [30]. For these reasons, we also conducted experiments with FB15k-237 a variant of FB15k where reverse relations have been removed [30]. FB122 is a subset of FB15k focusing on relations pertaining to the topics of “people”, “location”, and “sports.” In previous work, a set of 47 logical rules was created for FB122 and subsequently used in experiments for methods that take logical rules into account [10, 18].

The main objective of this paper is to investigate the impact of incorporating numerical features. Hence, we created two additional data set by removing those triples from FB15k’s and FB15k-237’s validation and test sets where numerical features are never used for the triples’ relation type. Hence, the remaining test and validation triples lead to completion queries where the numerical features under consideration are potentially used. We refer to these new data sets as FB15k-num and FB15k-237-num. A similar methodology can be followed to evaluate the performance on a different set of numerical features.

We extracted numerical data from the 1.9 billion triple Freebase RDF dump by mining triples that associate entities to literals of some numerical type. For example, the relation `/location/geocode/latitude` maps entities to their latitude. We performed these extractions for all entities in FB15k but only kept a numerical feature if at least 5 entities had values for it. This resulted in 116 different numerical features and 12,826 entities for which at least one of the numerical features had a value. On average each entity had 2.3 numerical features with a

$n_{(h,t)}$	MR	MRR	Hits@1	Hits@10
sign	231	29.7	20.1	50.1
RBF	121	31.4	21.2	52.3

Table 2: KBLRN for two possible input representations of numerical features for FB15k-237-num.

value. Since numerical data is not available for Wordnet, we do not perform experiments with numerical features for variants of this KB.

Each data set contains a set of triples that are known to be true (usually referred to as positive triples). Statistics of the data sets are provided in Table 1. Since the identifiers for entities and relations have been changed in FB13 [29], we could not extract numerical features for the data set and excluded it from the experiments.

5.2 General Set-up

We evaluated the different methods by their ability to answer completion queries of the form $(h, r, ?)$ and $(?, r, t)$. For queries of the form $(h, r, ?)$, we replaced the tail by each of the KB’s entities in turn, sorted the triples based on the scores or probabilities returned by the different methods, and computed the rank of the correct entity. We repeated the same process for the queries of type $(?, r, t)$. We follow the filtered setting described in [3] which removes correct triples that are different to the target triple from the ranked list. The mean of all computed ranks is the Mean Rank (lower is better) and the fraction of correct entities ranked in the top n is called hits@ n (higher is better). We also compute the Mean Reciprocal Rank (higher is better) which is an evaluation measure for rankings that is less susceptible to outliers.

We conduct experiments with the scoring function of DISTMULT [35] which is an application of parallel factor analysis to multi-relational graphs. For a review on parallel factor analysis we refer the reader to [12]. We validated the embedding size of KBLRN from the values $\{100, 200\}$ for all experiments. These values are used in most of the literature on KB embedding methods. For all other embedding methods, we report the original results from the literature or run the authors’ original implementation. For FB15k and FB15k-237, the results for Dist-

	FB15k				FB15k-237			
	MR	MRR	Hits@1	Hits@10	MR	MRR	Hits@1	Hits@10
TRANSE	51	44.3	25.1	76.3	214	25.1	14.5	46.3
DISTMULT	-	65.4	54.6	82.4	-	19.1	10.6	37.6
COMPLEX	-	69.2	59.9	84.0	-	20.1	11.2	38.8
NODE+LINKFEAT	-	82.2	-	87.0	-	23.7	-	36.0
R-GCN+	-	69.6	60.1	84.2	-	24.8	15.3	41.7
CONVE	64	74.5	67.0	87.3	330	30.1	22.0	45.8
without numerical features								
KBL	69	77.4	71.2	87.6	231	30.1	21.4	47.5
KBR	628	78.7	75.6	84.3	2518	18.0	12.8	28.5
KBLR	45	79.0	74.2	87.3	231	30.6	22.0	48.2
with numerical features								
KBLN	66	78.3	72.6	87.8	229	30.4	22.0	47.0
KBRN	598	78.7	75.6	84.2	3303	18.2	13.0	28.7
KBLRN	44	79.4	74.8	87.5	209	30.9	21.9	49.3

Table 3: Results (filtered setting) for KBLRN and state of the art approaches.

	FB15k-num				FB15k-237-num			
	MR	MRR	Hits@1	Hits@10	MR	MRR	Hits@1	Hits@10
TRANSE	25	34.7	5.5	79.9	158	21.8	10.41	45.6
DISTMULT	39	72.6	62.1	89.7	195	26.4	16.4	47.3
without numerical features								
KBL	39	72.6	62.1	89.7	195	26.4	16.4	47.3
KBR	399	84.7	81.6	90.1	3595	23.6	17.8	36.1
KBLR	28	85.3	80.3	92.4	232	29.3	19.7	49.2
with numerical features								
KBLN	32	73.6	63.0	90.7	122	28.6	17.9	51.6
KBRN	68	84.0	80.6	90.0	600	26.1	19.3	39.7
KBLRN	25	85.9	81.0	92.9	121	31.4	21.2	52.3

Table 4: Results (filtered) on the data sets where the test and validation sets are comprised of those triples whose type could potentially benefit from numerical features.

	WN18+rules[10]					FB122-all[10]				
	MR	MRR	Hits@3	Hits@5	Hits@10	MR	MRR	Hits@3	Hits@5	Hits@10
TRANSE	-	45.3	79.1	89.1	93.6	-	48.0	58.9	64.2	70.2
TRANSH	-	56.0	80.0	86.1	90.0	-	46.0	53.7	59.1	66.0
TRANSR	-	51.4	69.7	77.5	84.3	-	40.1	46.4	52.4	59.3
KALE-PRE	-	53.2	86.4	91.9	94.4	-	52.3	61.7	66.2	71.8
KALE-JOINT	-	66.2	85.5	90.1	93.0	-	52.3	61.2	66.4	72.8
COMPLEX	-	94.2	94.7	95.0	95.1	-	64.1	67.3	69.5	71.9
ASR-COMPLEX	-	94.2	94.7	95.0	95.1	-	69.8	71.7	73.6	75.7
KBL	537	80.8	92.5	93.7	94.7	117	69.5	74.6	77.2	80.0
KBR	7113	72.0	72.1	72.1	72.1	2018	54.7	54.7	54.7	54.7
KBLR	588	93.6	94.5	94.8	95.1	113	70.2	74.0	77.0	79.7

Table 5: Results (filtered setting) for KB completion benchmarks where logical rules are provided.

Mult, Complex, and R-GCN+ are taken from [28]; results for the relational baseline Node+LinkFeat are taken from [30]; results for ConvE are taken from [4] and results for TransE were obtained by running the authors’ implementation. For WN18-rules and FB122-all, the results for TransE, TransH, TransR, and KALE are taken from [10], and results for ComplEx and ASR-ComplEx are taken from [18]. All methods were tuned for each of the respective data sets.

For KBLRN we used ADAM [14] for parameter learning in a mini-batch setting with a learning rate of 0.001, the categorical cross-entropy as loss function and the number of epochs was set to 100. We validated every 5 epochs and stopped learning whenever the MRR (Mean Reciprocal Rank) values on the validation set decreased. The batch size was set to 512 and the number N of negative samples to 500 for all experiments. We use the abbreviations KB_{suffix} to refer to the different instances of

KBLRN. *suffix* is a combination of the letters L (Latent), R (Relational) and N (Numerical) to indicate the inclusion of each of the three feature types.

5.3 Automated Generation of Relational and Numerical Features

For the data sets FB15k, FB15k-237, and their numerical versions, we used all relational paths of length one and two found in the training data as relational features. These correspond to the formula types (h, r, t) (1-hop) and $\exists x (h, r_1, x) \wedge (x, r_2, t)$ (2-hops). We computed these relational paths with AMIE+ [5], a highly efficient system for mining logical rules from knowledge bases. We used the standard settings of AMIE+ with the exception that the minimal head support was set to 1. With these settings, AMIE+ returns horn rules of the form $body \Rightarrow (x, r, y)$ that are present for at least 1% of the triples of the form (x, r, y) . For each relation r , we used the body of those rules where r occurs in the head as r 's relational path features. For instance, given a rule such as $(x, r_1, z), (z, r_2, y) \Rightarrow (x, r, y)$, we introduce the relational feature $\exists x (h, r_1, x) \wedge (x, r_2, t)$ for the relation r . Table 7 lists a sample of relational features that contributed positively to the performance of KBLRN for specific relation types. For the data sets WN18 and FB122, we used the set of logical formulas previously used in the literature [10]. Using the same set of relational features allows us to compare KBLRN with existing approaches that incorporate logical rules into the embedding learning objective [10, 18].

For each relation r we only included a numerical feature if, in at least $\tau = 90\%$ of training triples, both the head and the tail had a value for it. This increases the likelihood that the feature is usable during test time. For $\tau = 90\%$ there were 105 relations in FB15k for which at least one numerical feature was included during learning, and 33 relations in FB15k-237. With the exception of the RBF parameters, all network weights are initialized following [9]. The parameters of KBLRN's RBF kernels are initialized and fixed to $c_i = \frac{1}{|\mathbf{T}|} \sum_{(h,r,t) \in \mathbf{T}} \mathbf{n}_{(h,t)}^{(i)}$, where \mathbf{T} is the set of training triples (h, r, t) for the relation r for which both $\mathbf{n}_h^{(i)}$ and $\mathbf{n}_t^{(i)}$ have a value; and $\sigma_i = \sqrt{\frac{1}{|\mathbf{T}|} \sum_{(h,r,t) \in \mathbf{T}} (\mathbf{n}_{(h,t)}^{(i)} - c_i)^2}$.

5.4 Representations of Numerical Features

We experimented with different strategies for incorporating raw numerical features. For the difference of feature values the simplest method is the application of the sign function. For a numerical attribute i , the activation is either 1 or -1 depending on whether the difference $\mathbf{n}_h^{(i)} - \mathbf{n}_t^{(i)}$ is positive or negative. For a more

Relation	KBLR		KBLRN	
	MRR	H@10	MRR	H@10
capital_of	5.7	13.6	14.6	18.2
spouse_of	4.4	0.0	7.9	0.0
influenced_by	7.3	20.9	9.9	26.8

Table 6: MRR and hits@10 results (filtered) for KBLRN with and without numerical features in FB15k-237. Results improve for relations where the difference of the relevant features is approximately normal (see Figure 3).

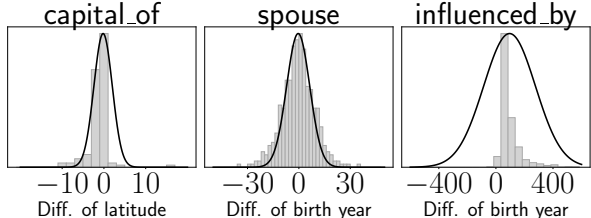


Figure 3: Histograms and fitted RBFs for three representative relations and numerical features.

nanced representation of differences of numerical features, a layer of RBF kernels is a suitable choice since the activation is here highest in a particular range of input values. The RBF kernel might not be appropriate, however, in cases where the underlying distribution is not normal.

To evaluate different input representations, we conducted experiments with KBLRN on the FB15k-237-num data set. Table 2 depicts the KB completion performance of two representation strategies for the difference of head and tail values. Each row corresponds to one evaluated strategy. “sign” stands for applying the sign function to the difference of numerical feature values. RBF stands for using an RBF kernel layer for the differences of numerical feature values. All results are for the FB15k-237-num test triples.

The RBF kernels outperform the sign functions significantly. This indicates that the difference of feature values is often distributed normally and that having a region of activation is beneficial. Given these results, we use the RBF input layer for $\mathbf{n}_{(h,t)}$ for the remainder of the experiments.

5.5 Comparison to State of the Art

For the standard benchmark data sets FB15k and FB15k-237, we compare KBLRN with TRANSE, DISTMULT, COMPLEX [32], R-GCN+ [28], and ConvE [4].

Table 3 lists the KB completion results. KBLRN is competitive with state of the art knowledge base comple-

Relational Feature	Triple (h, r, t)
$\exists x (h, \text{containedby}, x) \wedge (t, \text{locations_in_this_time_zone}, x)$	(h, time_zone, t)
$\exists x (t, \text{prequel}, x) \wedge (x, \text{character}, h)$	(h, character_in_film, t)
$\exists x (x, \text{cause_of_death}, h) \wedge (x, \text{cause_of_death}, t)$	(h, includes_causes_of_death, t)

Table 7: Relational features found by AMIE+ that positively contributed to the performance of KBLRN for the particular relation type holding between a head entity h and tail entity t.

	MR		MRR	
	ONE	MANY	ONE	MANY
KBLR	42	201	74.9	21.0
KBLRN	60	135	74.2	22.8

Table 8: MR and MRR results (filtered) on FB15k-237-num based on the cardinality of the test queries.

tion methods on FB15k and significantly outperforms all other methods on the more challenging FB15k-237 data set. Since the fraction of triples that can potentially benefit from numerical features is very small for FB15k, the inclusion of numerical features is only slightly beneficial. For FB15k-237, however, the numerical features significantly improve the results.

For the numerical versions of FB15k and FB15k-237, we compared KBLRN to TransE and DistMult. Table 4 lists the results for the KB completion task on these data sets. KBLRN significantly outperforms the KB embedding approaches. The positive impact of including the numerical features is significant.

For the data sets WN18-rules and FB122-all we compared KBLRN to KB embedding methods TransE, TransR [16], TransH [34], and ComplEx [32] as well as state of the art approaches for incorporating logical rules into the learning process. The experimental set-up is consistent with that of previous work. Table 5 lists the results for the KB completion task on these data sets. KBLRN combining relational and latent representations significantly outperforms existing approaches on FB122 with exactly the same set of rules. This provides evidence that KBLRN’s strategy to combine latent and relational features is effective despite its simplicity relative to existing approaches. For WN18+rules, KBLRN is competitive with ComplEx, the best performing method on this data set. In addition, KBLRN’s performance improves significantly when relational and latent representations are combined. In contrast, ASR-COMPLEX is not able to improve the results of ComplEx, its underlying latent representation.

5.6 The Impact of Numerical Features

The integration of numerical features improves KBLRN’s performance significantly. We performed

	AUC-PR	MR	MRR
TRANSE	0.837	231	26.5
KBLR	0.913	94	66.8
KBLRN	0.958	43	70.8

Table 9: AUC-PR, MR, and MRR results for the completion query (USA, /location/contains, ?).

several additional experiments so as to gain a deeper understanding of the impact numerical features have.

Table 6 compares KBLRN’s performance with and without integrating numerical features on three relations. The performance of the model with numerical features is clearly superior for all three relationships (capital_of, spouse and influenced_by). Figure 3 shows the normalized histograms for the values $n_{(h,t)}$ for these relations. We observe the differences of feature values are approximately normal.

Following previous work [3], we have classified each test query of FB15k-237-num as either ONE or MANY, depending on the whether one or many entities can complete that query. For the queries labeled ONE the model without numerical features shows a slightly worse performance with respect to the model that makes use of them, whereas for the queries labeled MANY, KBLRN significantly outperforms KBLR in both MR and MRR.

A well-known finding is the lack of completeness of FB15k and FB15k-237. This results in numerous cases where the correct entity for a completion query is not contained in the ground truth (neither in the training, nor test, nor validation data set). This is especially problematic for queries where a large number of entities are correct completions. To investigate the actual benefits of the numerical features we carried out the following experiment: We manually determined all correct completions for the query (USA, /location/contains, ?). We ended up with 1619 entities that correctly complete the query. FB15k-237 contains only 954 of these correct completions. With a complete ground truth, we can now use the precision-recall area under the curve (PR-AUC) metric to evaluate KB completion methods [21, 19, 6]. A high PR-AUC represents both high recall and high precision. Table 9 lists the results for the different methods. KBLRN with numerical features consistently and significantly outperformed all other approaches.

6 CONCLUSIONS

We introduced KBLRN, a class of machine learning models that, to the best of our knowledge, is the first proposal aiming at integrating relational, latent, and continuous features of head and tail entities in KBs into a single end-to-end differentiable framework. KBLRN outperforms state of the art KB completion methods on a range of data sets. We show that the inclusion of numerical features is beneficial for KB completion tasks.

Future work will primarily study instances of experts that can combine different numerical features such as life expectancy and latitude. Furthermore, combinations of multiple different embedding methods, which was shown to be beneficial in recent work [33], is also possible with our PoE approach.

References

- [1] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1171, 2011.
- [2] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, 2014.
- [3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795, 2013.
- [4] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. *arXiv preprint arXiv:1707.01476*, 2017.
- [5] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. Fast rule mining in ontological knowledge bases with amie+. *The VLDB Journal*, 24(6):707–730, 2015.
- [6] A. Garcia-Duran, A. Bordes, N. Usunier, and Y. Grandvalet. Combining two and three-way embeddings models for link prediction in knowledge bases. *arXiv preprint arXiv:1506.00999*, 2015.
- [7] M. Gardner and T. M. Mitchell. Efficient and expressive knowledge base completion using sub-graph feature extraction. In *EMNLP*, pages 1488–1498, 2015.
- [8] M. Gardner, P. P. Talukdar, J. Krishnamurthy, and T. Mitchell. Incorporating vector space similarity in random walk inference over knowledge bases. In *EMNLP*, 2014.
- [9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9, pages 249–256, 2010.
- [10] S. Guo, Q. Wang, L. Wang, B. Wang, and L. Guo. Jointly embedding knowledge graphs and logical rules. In *EMNLP*, pages 192–202, 2016.
- [11] K. Guu, J. Miller, and P. Liang. Traversing knowledge graphs in vector space. In *EMNLP*, 2015.
- [12] R. A. Harshman and M. E. Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.
- [13] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.
- [14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [15] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, pages 529–539, 2011.
- [16] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187, 2015.
- [17] X. Liu and J. S. Baras. Trust-aware crowdsourcing with domain knowledge. In *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, pages 2913–2918. IEEE, 2015.
- [18] P. Minervini, T. Demeester, T. Rocktäschel, and S. Riedel. Adversarial sets for regularised neural link predictors. In *UAI*, 2017.
- [19] M. Nickel, X. Jiang, and V. Tresp. Reducing the rank in relational factorization models by including observable patterns. In *NIPS*, pages 1179–1187, 2014.
- [20] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.
- [21] M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, pages 809–816, 2011.

- [22] M. Niepert. Discriminative gaifman models. In *NIPS*, pages 3405–3413, 2016.
- [23] P. Pezeshkpour, L. Chen, and S. Singh. Embedding multimodal relational data. In *Workshop on Automated Knowledge Base Construction (AKBC)*, 2017.
- [24] L. D. Raedt, K. Kersting, S. Natarajan, and D. Poole. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(2):1–189, 2016.
- [25] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. Relation extraction with matrix factorization and universal schemas. 2013.
- [26] T. Rocktäschel, S. Singh, and S. Riedel. Injecting logical background knowledge into embeddings for relation extraction. In *HLT-NAACL*, pages 1119–1129, 2015.
- [27] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, 3rd edition, 2009.
- [28] M. Schlichtkrull, T. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.
- [29] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, pages 926–934, 2013.
- [30] K. Toutanova and D. Chen. 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, 2015.
- [31] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, pages 1499–1509, 2015.
- [32] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080, 2016.
- [33] Y. Wang, R. Gemulla, and H. Li. On multi-relational link prediction with bilinear models. *arXiv preprint arXiv:1709.04808*, 2017.
- [34] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119, 2014.
- [35] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Learning multi-relational semantics using neural-embedding models. *arXiv preprint arXiv:1411.4072*, 2014.