Probabilistic Box Embeddings for Uncertain Knowledge Graph Reasoning

Xuelu Chen[1*], Michael Boratko[2*], Muhao Chen[3,4]
Shib Sankar Dasgupta[2], Xiang Lorraine Li[2], Andrew McCallum[2]
[1]Department of Computer Science, UCLA
[2]College of Information and Computer Sciences, UMass Amherst
[3]Department of Computer Science, USC
[4]Information Sciences Institute, USC
shirleychen@cs.ucla.edu; mboratko@iesl.cs.umass.edu;
muhaoche@usc.edu; {ssdasgupta,xiangl,mccallum}@cs.umass.edu

Abstract

Knowledge bases often consist of facts which are harvested from a variety of sources, many of which are noisy and some of which conflict, resulting in a level of *uncertainty* for each triple. Knowledge bases are also often incomplete, prompting the use of embedding methods to generalize from known facts, however existing embedding methods only model triple-level uncertainty and reasoning results lack global consistency. To address these shortcomings, we propose BEUrRE, a novel uncertain knowledge graph embedding method with calibrated probabilistic semantics. BEUrRE models each entity as a *box* (i.e. axis-aligned hyperrectangle), and relations between two entities as affine transforms on the head and tail entity boxes. The geometry of the boxes allows for efficient calculation of intersections and volumes, endowing the model with calibrated probabilistic semantics and facilitating the incorporation of relational constraints. Extensive experiments on two benchmark datasets show that BEUrRE consistently outperforms baselines on confidence prediction and fact ranking due to it's probabilistic calibration and ability to capture high-order dependencies among facts.[1]

1 Introduction

Knowledge graphs (KGs) provide structured representations of facts about real-world entities and relations. In addition to deterministic KGs (Bollacker et al., 2008; Lehmann et al., 2015; Mahdisoltani et al., 2015), much recent attention has been paid to uncertain KGs (or UKGs). UKGs, such as ProBase (Wu et al., 2012), NELL (Mitchell et al., 2018), and ConceptNet (Speer et al., 2017), associate each fact (or triple) with a confidence score representing the likelihood of that fact to be true. Such uncertain knowledge representations critically capture the uncertain nature of reality, and provide more precise reasoning. For example, while both *(Honda, competeswith, Toyota)* and *(Honda, competeswith, Chrysler)* look somewhat correct, the former fact should have a higher confidence than the latter one, since Honda and Toyota are both Japanese car manufacturers and have highly overlapping customer bases. Similarly, while *(The Beatles, genre, Rock)* and *(The Beatles, genre, Pop)* are both true, the first one may receive a slightly higher confidence, since the Beatles is generally considered a rock band. Such confidence information is important when answering questions like *Who is the main competitor of Honda?*, or extracting confident knowledge for drug repurposing (Sosa et al., 2020).

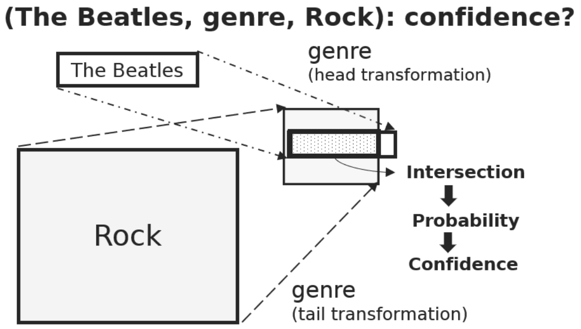

Figure 1: BEUrRE models entities as boxes and relations as two affine transforms.

To facilitate automated knowledge acquisition for UKGs, some UKG embedding models (Chen et al., 2019; Kertkeidkachorn et al., 2019) have recently been proposed. Inspired by the works about deterministic KG embeddings (Yang et al., 2015; Bordes et al., 2013), existing approaches model entities and relations as points in low-dimensional vector space, measure triple plausibility with vector similarity (eg. distance, dot-product), and map the plausibility to the confidence range of $[0, 1]$. For instance, the representative work UKGE (Chen et al., 2019) models the triple plausibility in the

[*] Indicating equal contribution.

[1] Resources and software are available at https://github.com/stasl0217/beurre

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 882–893
June 6–11, 2021. ©2021 Association for Computational Linguistics

form of embedding product (Yang et al., 2015), and trains the embedding model as a regressor to predict the confidence score. One interpretation of existing methods is that they model each triple using a binary random variable, where the latent dependency structure between different binary random variables is captured by vector similarities. Without an explicit dependency structure it is difficult to enforce logical reasoning rules to maintain global consistency.

In order to go beyond triple-level uncertainty modeling, we consider each entity as a binary random variable. However, representing such a probability distribution in an embedding space and reasoning over it is non-trivial. It is difficult to model marginal and joint probabilities for entities using simple geometric objects like vectors. In order to encode probability distributions in the embedding space, recent works (Lai and Hockenmaier, 2017; Vilnis et al., 2018; Li et al., 2019; Dasgupta et al., 2020) represent random variables as more complex geometric objects, such as cones and axis-aligned hyperrectangles (*boxes*), and use *volume* as the probability measure. Inspired by such advances of probability measures in embeddings, we present `BEUrRE` (**B**ox **E**mbedding for **U**nce**r**tain **RE**lational Data)[2]. `BEUrRE` represents entities as boxes. Relations are modeled as two separate affine transforms on the head and tail entity boxes. Confidence of a triple is modeled by the intersection between the two transformed boxes. Fig. 1 shows how a fact about the genre of the Beatles is represented under our framework.

Such representation is not only inline with the human perception that entities or concepts have different levels of granularity, but also allows more powerful domain knowledge representation. UKGE (Chen et al., 2019) has demonstrated that introducing domain knowledge about relation properties (e.g. transitivity) can effectively enhance reasoning on UKGs. While UKGE uses Probabilistic Soft Logic (PSL) (Bach et al., 2017) to reason for unseen facts and adds the extra training samples to training, such a method can lead to error propagation and has limited scope of application when UKG is sparse. In our work, we propose sufficient conditions for these relation properties to be preserved in the embedding space and directly model the relation properties by regularizing relation-specific transforms based on constraints. This technique is more robust to noise and has wide coverage that is not restricted by the scarcity of the existing triples. Extensive experiments on two benchmark datasets show that `BEUrRE` effectively captures the uncertainty, and consistently outperforms the baseline models on ranking and predicting confidence of unseen facts.

2 Related Work

We discuss two lines of related work.

UKG Embeddings. A UKG assigns a confidence score to each fact. The development of relation extraction and crowdsourcing in recent years enabled the construction of many large-scale uncertain knowledge bases. ConceptNet (Speer et al., 2017) is a multilingual KG of commonsense concepts, where triples are assigned with confidence measures reflecting crowdsourcing agreement. NELL (Mitchell et al., 2018) collects facts from web pages with an active-learnable information extraction model, and measures their confidence scores by semi-supervised learning with the Expectation-Maximum (EM) algorithm. Probase (Wu et al., 2012) is a general probabilistic taxonomy obtained from syntactic extraction. Aforementioned UKGs have supported numerous knowledge-driven applications, such as literature-based drug repurposing (Sosa et al., 2020).

Recently, a few UKG embedding methods have been proposed, which seek to facilitate automated knowledge acquisition for UKGs. UKGE (Chen et al., 2019) is the first work of this kind, which models triple plausibility as product of embedding vectors (Yang et al., 2015), and maps the plausibility to the confidence score range of $[0, 1]$. To further enhance the performance, UKGE incorporates PSL based constraints (Bach et al., 2017) to help enforce the global consistency of predicted knowledge. UOKGE (Boutouhami et al., 2020) jointly encodes the graph structure and the ontology structure to improve the confidence prediction performance, which however requires an additional ontology of entity types that is not always available to all KGs. In addition to the above UKG embeddings models, there is also a matrix-factorization-based approach URGE that seeks to embed uncertain graphs (Hu et al., 2017). However, URGE only considers the node proximity in the networks. URGE cannot handle multi-relational data and only generates node embeddings.

[2] "Beurre" is French for "butter".

Geometric Embeddings. Developing embedding methods to represent elements using geometric objects with more complex structures than (Euclidean) vectors is an active area of study. *Poincaré embeddings* (Nickel and Kiela, 2017) represent entities in hyperbolic space, leveraging the inductive bias of negative curvature to fit hierarchies. *Order embeddings* (Vendrov et al., 2016) take a region-based approach, representing nodes of a graph using infinite cones, and using containment between cones to represent edges. *Hyperbolic entailment cones* (Ganea et al., 2018) combine order embeddings with hyperbolic geometry. While these methods show various degrees of promise when embedding hierarchies, they do not provide scores between entities that can be interpreted probabilistically, which is particularly useful in our setting.

Lai and Hockenmaier (2017) extend order embeddings with a probabilistic interpretation by integrating the volume of the infinite cones under the negative exponential measure, however the rigid structure imposed by the cone representation limits the representational capacity, and the resulting model cannot model negative correlation or disjointness. Introduced by Vilnis et al. (2018), *probabilistic box embeddings* represent elements using axis-aligned hyperrectangles (or *boxes*). Box embeddings not only demonstrate improved performance on modeling hierarchies, such embeddings also capture probabilistic semantics based on box volumes, and are capable of compactly representing conditional probability distributions. A few training improvement methods for box embeddings have been proposed (Li et al., 2019; Dasgupta et al., 2020), and we make use of the latter, which is termed *GumbelBox* after the distribution used to model endpoints of boxes.

While box embeddings have shown promise in representing hierarchies, our work is the first use of box embeddings to represent entities in multi-relational data. *Query2Box* (Ren et al., 2020) and *BoxE* (Abboud et al., 2020) make use of boxes in the loss function of their models, however entities themselves are represented as vectors, and thus these models do not benefit from the probabilistic semantics of box embeddings, which we rely on heavily for modeling UKGs. In (Patel et al., 2020), the authors demonstrate the capability of box embeddings to jointly model two hierarchical relations, which is improved upon using a learned transform in (Dasgupta et al., 2021). Similarly to Ren et al. (2020) and Dasgupta et al. (2021), we also make use of a learned transform for each relation, however we differ from Ren et al. (2020) in that entities themselves are boxes, and differ from both in the structure of the learned transform.

3 Background

Before we move on to the presented method in this work, we use this section to introduce the background of box embeddings and the addressed task.

3.1 Uncertain Knowledge Graphs

A UKG consists of a set of weighted triples $\mathcal{G} = \{(l, s_l)\}$. For each pair (l, s_l), $l = (h, r, t)$ is a triple representing a fact where $h, t \in \mathcal{E}$ (the set of entities) and $r \in \mathcal{R}$ (the set of relations), and $s_l \in [0, 1]$ represents the confidence score for this fact to be true. Some examples of weighted triples from NELL are *(Honda, competeswith, Toyota)*: 1.00 and *(Honda, competeswith, Chrysler)*: 0.94.

UKG Reasoning. Given a UKG $\mathcal{G}$, the *uncertain knowledge graph reasoning* task seeks to predict the confidence of an unseen fact (h, r, t).

3.2 Probabilistic Box Embeddings

In this section we give a formal definition of probabilistic box embeddings, as introduced by Vilnis et al. (2018). A *box* is an n-dimensional hyperrectangle, i.e. a product of intervals

$$\prod_{i=1}^{d} [x_i^{\mathrm{m}}, x_i^{\mathrm{M}}], \quad \text{where} \quad x_i^{\mathrm{m}} < x_i^{\mathrm{M}}.$$

Given a space $\Omega_{\text{Box}} \subseteq \mathbb{R}^n$, we define $\mathcal{B}(\Omega_{\text{Box}})$ to be the set of all boxes in Ω_{Box}. Note that $\mathcal{B}(\Omega_{\text{Box}})$ is closed under intersection, and the volume of a box is simply the product of side-lengths. Vilnis et al. (2018) note that this allows one to interpret box volumes as unnormalized probabilities. This can be formalized as follows.

Definition 3.1. Let $(\Omega_{\text{Box}}, \mathcal{E}, P_{\text{Box}})$ be a probability space, where $\Omega_{\text{Box}} \subseteq \mathbb{R}^n$ and $\mathcal{B}(\Omega_{\text{Box}}) \subseteq \mathcal{E}$. Let $\mathcal{Y}$ be the set of binary random variables Y on Ω_{Box} such that $Y^{-1}(1) \in \mathcal{B}(\Omega_{\text{Box}})$. A *probabilistic box embedding* of a set S is a function $: S \to \mathcal{Y}$. We typically denote $f(s) =: Y_s$ and $Y_s^{-1}(1) =: \text{Box}(s)$.

Essentially, to each element of S we associate a box which, when taken as the support set of a binary random variable, allows us to interpret each element of S as a binary random variable. Using boxes for the support sets allows one to easily calculate marginal and conditional probabilities, for example if we embed the elements $\{\textsc{cat}, \textsc{mammal}\}$ as boxes in $\Omega_{\text{Box}} = [0,1]^d$ with P_{Box} as Lebesgue measure, then

$$P(\textsc{mammal} \mid \textsc{cat}) = P_{\text{Box}}(X_{\textsc{mammal}} | X_{\textsc{cat}}) \\ = \frac{\text{Vol}(\text{Box}(\textsc{mammal}) \cap \text{Box}(\textsc{cat}))}{\text{Vol}(\text{Box}(\textsc{cat}))}.$$

3.3 Gumbel Boxes

We further give a brief description of the *Gumbel-Box* method, which we rely on for training our box embeddings (Dasgupta et al., 2020).

As described thus far, probabilistic box embeddings would struggle to train via gradient descent, as there are many settings of parameters and objectives which have no gradient signal. (For example, if boxes are disjoint but should overlap.) To mitigate this, Dasgupta et al. (2020) propose a latent noise model, where the min and max coordinates of boxes in each dimension are modeled via Gumbel distributions, that is

$$\begin{aligned} \text{Box}(X) &= \prod_{i=1}^{d} [x_i^{\text{m}}, x_i^{\text{M}}] \quad \text{where} \\ x_i^{\text{m}} &\sim \text{GumbelMax}(\mu_i^{\text{m}}, \beta), \\ x_i^{\text{M}} &\sim \text{GumbelMin}(\mu_i^{\text{M}}, \beta). \end{aligned}$$

μ_i^{m} thereof is the *location* parameter, and β is the (global) variance. The Gumbel distribution was chosen due to its min/max stability, which means that the set of all "Gumbel boxes" are closed under intersection. Dasgupta et al. (2020) go on to provide an approximation of the expected volume of a Gumbel box,

$$\mathbb{E}\left[\text{Vol}(\text{Box}(X))\right] \approx \\ \prod_{i=1}^{d} \beta \log \left(1 + \exp\left(\tfrac{\mu_i^{\text{M}} - \mu_i^{\text{m}}}{\beta} - 2\gamma\right)\right).$$

A first-order Taylor series approximation yields

$$\mathbb{E}[P_{\text{Box}}(X_{\text{A}} \mid X_{\text{B}})] \approx \frac{\mathbb{E}[\text{Vol}(\text{Box}(A) \cap \text{Box}(B))]}{\mathbb{E}[\text{Vol}(\text{Box}(B))]},$$

and Dasgupta et al. (2020) empirically demonstrate that this approach leads to improved learning when targeting a given conditional probability distribution as the latent noise essentially ensembles over a large collection of boxes which allows the model to escape plateaus in the loss function. We therefore use this method when training box embeddings.

Remark 3.1. While we use Gumbel boxes for training, intuition is often gained by interpreting these boxes as standard hyperrectangles, which is valid as the Gumbel boxes can be seen as a distribution over such rectangles, with the Gumbel variance parameter β acting as a global measure of uncertainty. We thus make statements such as $\text{Box}(X) \subseteq \text{Box}(Y)$, which, strictly speaking, are not well-defined for Gumbel boxes. However we can interpret this probabilistically as $P(Y \mid X) = 1$ which coincides with the conventional interpretation when $\beta = 0$.

4 Method

In this section, we present our UKG embedding model `BEUrRE`. The proposed model encodes entities as probabilistic boxes and relations as affine transforms. We also discuss the method to incorporate logical constraints into learning.

4.1 Modeling UKGs with Box Embeddings

`BEUrRE` represents entities as Gumbel boxes, and a relation r acting on these boxes by translation and scaling. Specifically, we parametrize a Gumbel box $\text{Box}(X)$ using a center $\text{cen}(\text{Box}(X)) \in \mathbb{R}^d$ and offset $\text{off}(\text{Box}(X)) \in \mathbb{R}_+^d$, where the location parameters are given by

$$\begin{aligned} \mu_i^{\text{m}} &= \text{cen}(\text{Box}(X)) - \text{off}(\text{Box}(X)), \\ \mu_i^{\text{M}} &= \text{cen}(\text{Box}(X)) + \text{off}(\text{Box}(X)). \end{aligned}$$

We consider transformations on Gumbel boxes parametrized by a translation vector $\tau \in \mathbb{R}^d$ and a scaling vector $\Delta \in \mathbb{R}_+^d$ such that

$$\begin{aligned} \text{cen}(f(\text{Box}(X); \tau, \Delta)) &= \text{cen}(\text{Box}(X)) + \tau, \\ \text{off}(f(\text{Box}(X); \tau, \Delta)) &= \text{off}(\text{Box}(X)) \circ \Delta, \end{aligned}$$

where $\circ$ is the Hadamard product. We use separate actions for the head and tail entities of a relation, which we denote f_r and g_r, and omit the explicit dependence on the learned parameters τ and Δ.

Remark 4.1. Note that these relations are not an affine transformations of the *space*, Ω_{Box}, rather they perform a transformation of a *box*. These functions form an Abelian group under composition, and furthermore define a transitive, faithful group action on the set of (Gumbel) boxes.

Given a triple (h, r, t), `BEUrRE` models the confidence score using the (approximate) conditional probability given by

$$\phi(h, r, t) = \frac{\mathbb{E}[\text{Vol}(f_r(\text{Box}(h)) \cap g_r(\text{Box}(t)))]}{\mathbb{E}[\text{Vol}(g_r(\text{Box}(t)))]}.$$

We can think of the box $f_r(\mathrm{Box}(h))$ as the support set of a binary random variable representing the concept h in the context of the head position of relation r, for example $\mathrm{Box}(\text{THEBEATLES})$ is a latent representation of the concept of The Beatles, and $f_{\text{GENRE}}(\mathrm{Box}(\text{THEBEATLES}))$ represents The Beatles in the context of genre classification as the object to be classified.

4.2 Logical Constraints

The sparsity of real-world UKGs makes learning high quality representations difficult. To address this problem, previous work (Chen et al., 2019) introduces domain knowledge about the properties of relations (e.g., transitivity) and uses PSL over first-order logical rules to reason for unseen facts and create extra training samples. While this technique successfully enhances the performance by incorporating constraints based on relational properties, the coverage of such reasoning is still limited by the density of the graph. In UKGE, the confidence score of a triple can be inferred and benefit training only if all triples in the rule premise are already present in the KG. This leads to a limited scope of application, particularly when the graph is sparse.

In our work, we propose sufficient conditions for these relation properties to be preserved in the embedding space and directly incorporating the relational constraints by regularizing relation-specific transforms. Compared to previous work, our approach is more robust to noise since it does not hardcode inferred confidence for unseen triples, and it has wide coverage that is not restricted by the scarcity of the existing triples.

In the following, we discuss the incorporation of two logical constraints — transitivity and composition — in the learning process. We use capital letters A, B, C to represent universally quantified entities from UKG and use Φ to denote a set of boxes sampled from $\mathcal{B}(\Omega_{\text{Box}})$.

Transitivity Constraint. A relation r is *transitive* if $(A, r, B) \wedge (B, r, C) \implies (A, r, C)$. An example of a transitive relation is *hypernymy*.

The objective of imposing a transitivity constraint in learning is to preserve this property of the relation in the embedding space, i.e. to ensure that (A, r, C) will be predicted true if (A, r, B) and (B, r, C) are true. This objective is fulfilled if $g_r(\mathrm{Box}(B))$ contains $f_r(\mathrm{Box}(B))$. An illustration of the box containment relationships is given in Fig 2. Thus, we constrain f_r and g_r so that $g_r(u)$ contains $f_r(u)$ for any $u \in \Omega_{\text{Box}}$. We impose the constraint with the following regularization term:

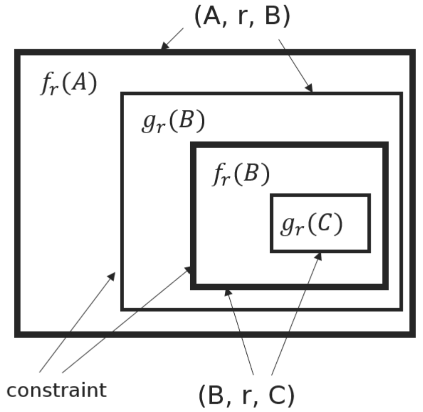

Figure 2: Illustration of how the constraint that $g_r(u)$ contains $f_r(u)$ preserves transitivity of relation r in the embedding space. A triple (h, r, t) is true if and only if $f_r(\mathrm{Box}(h))$ contains $g_r(\mathrm{Box}(t))$. By adding this constraint, $f_r(\mathrm{Box}(A))$ is guaranteed to contain $g_r(\mathrm{Box}(C))$ if (A, r, B) and (B, r, C) are true.

$$L_{\text{tr}}(r) = \frac{1}{|\Phi|} \sum_{u \in \Phi} \|P_{\text{Box}}(g_r(u) \mid f_r(u)) - 1\|^2 .$$

Composition Constraint. A relation r_3 is *composed of* relation r_1 and relation r_2 if $(A, r_1, B) \wedge (B, r_2, C) \implies (A, r_3, C)$. For example, the relation *atheletePlaysSports* can be composed of relations *atheletePlaysForTeam* and *teamPlaysSports*.

To preserve the relation composition in the embedding space, we constrain that the relation-specific mappings f_{r_3} and g_{r_3} are the *composite mappings* of f_{r_1}, f_{r_2} and g_{r_1}, g_{r_2} respectively:

$$f_{r_3} = f_{r_2} \cdot f_{r_1};\; g_{r_3} = g_{r_2} \cdot g_{r_1}.$$

where $\cdot$ is the mapping composition operator. Thus, for any $u \in \Omega_{\text{Box}}$, we expect that $f_{r_3}(u)$ is the same as $f_{r_2}(f_{r_1}(u))$ and $g_{r_3}(u)$ is the same as $g_{r_2}(g_{r_1}(u))$. We accordingly add the following regularization term

$$\begin{aligned} L_{\text{c}}(r_1, r_2, r_3) =& \frac{1}{|\Phi|} \sum_{u \in \Phi} f_{r_3}(u) \oplus f_{r_2}(f_{r_1}(u)) \\ &+ g_{r_3}(u) \oplus g_{r_2}(g_{r_1}(u)) \end{aligned}$$

where $\oplus$ is defined as

$$\begin{aligned} \mathrm{Box}_1 \oplus \mathrm{Box}_2 =& \|1 - P_{\text{Box}}(\mathrm{Box}_1 \mid \mathrm{Box}_2)\|^2 \\ &+ \|1 - P_{\text{Box}}(\mathrm{Box}_2 \mid \mathrm{Box}_1)\|^2 . \end{aligned}$$

4.3 Learning Objective

The learning process of BEUrRE optimizes two objectives. The main objective optimizes the loss for a regression task and, simultaneously, a constrained regularization loss enforces the aforementioned constraints.

Let $\mathcal{L}^+$ be the set of observed relation facts in training data. The goal is to minimize the mean squared error (MSE) between the ground truth confidence score s_l and the prediction $\phi(l)$ for each relation $l \in \mathcal{L}^+$. Following UKGE (Chen et al., 2019), we also penalize the predicted confidence scores of facts that are not observed in UKG. The main learning objective is as follows:

$$\mathcal{J}_1 = \sum_{l \in \mathcal{L}^+} |\phi(l) - s_l|^2 + \alpha \sum_{l \in \mathcal{L}^-} |\phi(l)|^2.$$

where $\mathcal{L}^-$ is a sample set of the facts not observed in UKG, and α is a hyper-parameter to weigh unobserved fact confidence penalization. Similar to previous works, we sample those facts by corrupting the head and the tail for observed facts to generate $\mathcal{L}^-$ during training.

In terms of constraints, let $\mathcal{R}_{\text{tr}}$ be the set of transitive relations, $\mathcal{R}_{\text{c}}$ be the set of composite relation groups, and w_{tr} and w_{c} be the regularization coefficients. We add the following regularization to impose our constraints on relations:

$$\mathcal{J}_2 = w_{\text{tr}} \sum_{r \in \mathcal{R}_{\text{tr}}} L_{\text{tr}}(r) + w_{\text{c}} \sum_{(r_1,r_2,r_3) \in \mathcal{R}_{\text{c}}} L_{\text{c}}(r_1, r_2, r_3).$$

Combining both learning objectives, the learning process optimizes the joint loss $J = J_1 + J_2$.

4.4 Inference

Once BEUrRE is trained, the model can easily infer the confidence of a new fact (h, r, t) based on the confidence score function $\phi(h, r, t)$ defined in Section 4.1. This inference mechanism easily supports other types of reasoning tasks, such as inferring the plausibility of a new fact, and ranking multiple related facts. The experiments presented in the next section will demonstrate the ability of BEUrRE to perform those reasoning tasks.

5 Experiments

In this section we present evaluation of our model on two UKG reasoning tasks, i.e. confidence prediction and fact ranking. More experimentation details are in Appendices.

Dataset	#Ent.	#Rel.	#Rel. Facts	Avg(s)	Std(s)
CN15k	15,000	36	241,158	0.629	0.232
NL27k	27,221	404	175,412	0.797	0.242

Table 1: Statistics of the datasets. *Ent.* and *Rel.* stand for entities and relations. Avg(s) and Std(s) are the average and standard deviation of confidence.

Dataset	Transitivity	Composition
CN15k	causes	N/A
NL27k	locationAtLocation	(atheletePlaysForTeam, teamPlaysSport) → atheletePlaysSport

Table 2: Examples of relations with logical constraints.

5.1 Experiment settings

Datasets. We follow Chen et al. (2019) and evaluate our models on CN15k and NL27k benchmarks, which are subsets of ConceptNet (Speer et al., 2017) and NELL (Mitchell et al., 2018) respectively. Table 1 gives the statistics of the datasets. We use the same split provided by Chen et al. (2019): 85% for training, 7% for validation, and 8% for testing. We exclude the dataset PPI5k, the subgraph of the protein-protein interaction (PPI) network STRING (Szklarczyk et al., 2016), where the supporting scores of PPI information are indicators based on experimental and literary verification, instead of a probabilistic measure.

Logical constraints. We report results of both versions of our model with and without logical constraints, denoted as BEUrRE (rule+) and BEUrRE respectively. For a fair comparison, we incorporate into BEUrRE (rule+) the same set of logical constraints as UKGE (Chen et al., 2019). Table 2 gives a few examples of the relations on which we impose constraints.

Baselines. We compare our models with UKG embedding models as well as deterministic KG embedding models.

UKG embedding models include UKGE (Chen et al., 2019) and URGE (Hu et al., 2017). While UKGE has multiple versions incorporated with different regression functions, we report the results of the best performing one with the logistic function. We also include results for both settings with and without constraints, marked as UKGE (rule+) and UKGE in result tables respectively.

Dataset	CN15k		NL27k	
Metrics	MSE	MAE	MSE	MAE
URGE	10.32	22.72	7.48	11.35
UKGE	9.02	20.05	2.67	7.03
BEUrRE	7.80	20.03	2.37	7.12
UKGE(rule+)	8.61	19.90	2.36	6.90
BEUrRE(rule+)	**7.49**	**19.88**	**2.01**	**6.89**

Table 3: Results of fact confidence prediction ($\times 10^{-2}$).

Variants	uncons.	rule+
Metrics	MSE ($\times 10^{-2}$)	
BEUrRE	7.80	7.49
—w/o Gumbel distribution	8.13	8.14
—Single relation-specific transform	7.81	7.60

Table 4: Ablation study results on CN15k. *uncons.* represents the unconstrained setting, and *rule+* denotes the logically constrained setting.

URGE was originally designed for probabilistic homogeneous graphs and cannot handle multi-relational graphs, so accordingly we ignore relation information when embedding a UKG. UOKGE (Boutouhami et al., 2020) cannot serve as a baseline because it requires additional ontology information for entities that is not available to these UKGs.

Deterministic KG embedding models TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), and TuckER (Balazevic et al., 2019) have demonstrated high performance on reasoning tasks for deterministic KGs, and we also include them as baselines. These models cannot predict confidence scores for uncertain facts, so we compare our method with them only on the ranking task. Following Chen et al. (2019), we only use facts with confidence above the threshold $\tau = 0.85$ to train deterministic models.

Model configurations. We use Adam (Kingma and Ba, 2014) as the optimizer and fine-tune the following hyper-parameters by grid search based on the performance on the validation set, i.e. MSE for confidence prediction and normalized Discounted Cumulative Gain (nDCG) for fact ranking. Hyper-parameter search range and the best hyper-parameter configurations are given in Appendix A.1. Training terminates with early stopping based on the same metric with a patience of 30 epochs. We repeat each experiment five times and report the average results.

5.2 Confidence Prediction

This task seeks to predict the confidence of new facts that are unseen to training. For each uncertain fact (l, s_l) in the test set, we predict the confidence of l and report the mean squared error (MSE) and mean absolute error (MAE).

Results. Results are reported in Table 3. We compare our models with baselines under the unconstrained and logically constrained (marked with *rule+*) settings respectively. Under both settings, BEUrRE outperforms the baselines in terms of MSE on both datasets.

Under the unconstrained setting, BEUrRE improves MSE of the best baseline UKGE by 0.012 (ca. 14% relative improvement) on CN15k and 0.003 (ca. 11% relative improvement) on NL27k. The enhancement demonstrates that box embeddings can effectively improve reasoning on UKGs. It is worth noting that even without constraints in learning, BEUrRE can still achieve comparable MSE and MAE to the logically constrained UKGE (rule+) on both datasets and even outperforms UKGE (rule+) on CN15k. Considering that constraints of relations in CN15k mainly describe transitivity, the aforementioned observation is consistent with the fact that box embeddings are naturally good at capturing transitive relations, as shown in the recent study (Vilnis et al., 2018).

With logical constraints, BEUrRE (rule+) further enhances the performance of BEUrRE and reduces its MSE by 0.0031 (ca. 4% relative improvement) on CN15k and 0.0036 (ca. 15% relative improvement) on NL27k. This is as expected, since logical constraints capture higher-order relations of facts and lead to more globally consistent reasoning. We also observe that BEUrRE (rule+) brings larger gains over BEUrRE on NL27k, where we have both transitivity constraints and composition constraints, than on CN15k with only transitivity constraints incorporated.

In general, with box embeddings, BEUrRE effectively improves reasoning on UKGs with better captured fact-wise confidence. Furthermore, the results under the logically constrained setting show the effectiveness of improving reasoning with higher-order relations of uncertain facts.

Ablation Study. To examine the contribution from Gumbel distribution to model box boundaries and the effectiveness of representing relations as two

Dataset	CN15K		NL27k	
Metrics	linear	exp.	linear	exp.
TransE	0.601	0.591	0.730	0.722
DistMult	0.689	0.677	0.911	0.897
ComplEx	0.723	0.712	0.921	0.913
RotatE	0.715	0.703	0.901	0.887
TuckER	0.736	0.724	0.877	0.870
URGE	0.572	0.570	0.593	0.593
UKGE	0.769	0.768	0.933	0.929
BEUrRE	0.796	0.795	0.942	0.942
UKGE(rule+)	0.789	0.788	0.955	0.956
BEUrRE(rule+)	**0.801**	**0.803**	**0.966**	**0.970**

Table 5: Mean nDCG for fact ranking. *linear* stands for linear gain, and *exp.* stands for exponential gain.

separate transforms for head and tail boxes, we conduct an ablation study based on CN15k. The results for comparison are given in Table 4. First, we resort to a new configuration of BEUrRE where we use smoothed boundaries for boxes as in (Li et al., 2019) instead of Gumbel boxes. We refer to boxes of this kind as soft boxes. Under the unconstrained setting, using soft boxes increases MSE by 0.0033 on CN15k (ca. 4% relative degradation), with even worse performance observed when adding logical constraints. This confirms the finding by Dasgupta et al. (2020) that using Gumbel distribution for boundaries greatly improves box embedding training. Next, to analyze the effect of using separate transforms to represent a relation, we set the tail transform g_r as the identity function. For logical constraint incorporation, we accordingly update the constraint on transitive relation r as $P_{\text{Box}}(u \mid f_r(u)) = 1, u \in \Omega_{\text{Box}}$, which requires that u always contains $f_r(u)$, i.e. the translation vector of f_r is always zero and elements of the scaling vector are always less than 1. Although there is little difference between using one or two transforms under the unconstrained setting, under the logically constrained setting, the constraint is too stringent to be preserved with only one transform.

Case study. To investigate whether our model can encode meaningful probabilistic semantics, we present a case study about box volumes. We examine the objects of the *atLocation* predicate on CN15k and check which entity boxes have larger volume and cover more entity boxes after the relation transformation. Ideally, geographic entities with larger areas or more frequent mentions should be at the top of the list. When using the BEUrRE(rule+) model, the top 10 in all entities are *place, town, bed, school, city, home, house, capital, church, camp*, which are general concepts. Among the observed objects of the *atLocation* predicate, the entities that have the least coverage are *Tunisia, Morocco, Algeria, Westminster, Veracruz, Buenos Aires, Emilia-Romagna, Tyrrhenian sea, Kuwait, Serbia*. Those entities are very specific locations. This observation confirms that the box volume effectively represents probabilistic semantics and captures specificity/granularity of concepts, which we believe to be a reason for the performance improvement.

5.3 Fact Ranking

Multiple facts can be associated with the same entity. However, those relevant facts may appear with very different plausibility. Consider the example about Honda Motor Co. in Section 1, where it was mentioned that *(Honda, competeswith, Toyota)* should have a higher belief than *(Honda, competeswith, Chrysler)*. Following this intuition, this task focuses on ranking multiple candidate tail entities for a query $(h, r, \underline{?t})$ in terms of their confidence.

Evaluation protocol. Given a query $(h, r, \underline{?t})$, we rank all the entities in the vocabulary as tail entity candidates and evaluate the ranking performance using the normalized Discounted Cumulative Gain (nDCG) (Li et al., 2009). The gain in retrieving a relevant tail t_0 is defined as the ground truth confidence $s_{(h,r,t_0)}$. Same as Chen et al. (2019), we report two versions of nDCG that use linear gain and exponential gain respectively. The exponential gain puts stronger emphasis on the most relevant results.

Results. We report the mean nDCG over the test query set in Table 5. Although the deterministic models do not explicitly capture the confidence of facts, those models are trained with high-confidence facts and have a certain ability to differentiate high confidence facts from lesser ones. URGE ignores relation information and yields worse predictions than other models. UKGE explicitly models uncertainty of facts and is the best performing baseline.

The proposed BEUrRE leads to more improvements under both the unconstrained and logically constrained settings. Under the unconstrained setting, BEUrRE offers consistently better performance over UKGE. Specifically, on CN15k, BEUrRE leads to 0.027 improvement in both linear nDCG and exponential nDCG. On NL27k, it offers 0.009 higher linear nDCG and 0.013 higher exponential nDCG. Similar to the results on the confidence prediction task, even unconstrained BEUrRE is able to outperform the logically constrained UKGE (rule+) on CN15k without incorporating any constraints of relations. This further confirms the superior expressive power of box embeddings.

In summary, box embeddings improve accuracy and consistency of reasoning and BEUrRE delivers better fact ranking performance than baselines.

6 Conclusion

This paper presents a novel UKG embedding method with calibrated probabilistic semantics. Our model BEUrRE encodes each entity as a Gumble box representation whose volume represents marginal probability. A relation is modeled as two affine transforms on the head and tail entity boxes. We also incorporate logic constraints that capture the high-order dependency of facts and enhance global reasoning consistency. Extensive experiments show the promising capability of BEUrRE on confidence prediction and fact ranking for UKGs. The results are encouraging and suggest various extensions, including deeper transformation architectures as well as alternative geometries to allow for additional rules to be imposed. In this context, we are also interested in extending the use of the proposed technologies into more downstream tasks, such as knowledge association (Sun et al., 2020) and event hierarchy induction (Wang et al., 2020). Another direction is to use BEUrRE for ontology construction and population, since box embeddings are naturally capable of capturing granularities of concepts.

Ethical Considerations

Real-world UKGs often harvest data from open data sources and may include biases. Reasoning over biased UKGs may support or magnify those biases. While not specifically addressed in this work, the ability to inject logical rules could be one way to mitigate bias, and the ability to interpret the learned representation probabilistically allows the investigation of potential learned biases.

All the datasets used in this paper are publicly available and free to download. The model proposed in the paper aims to model uncertainty in knowledge graphs more accurately, and the effectiveness of the proposed model is supported by the empirical experiment results.

Acknowledgment

We appreciate the anonymous reviewers for their insightful comments and suggestions.

This material is based upon work sponsored by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, and by Air Force Research Laboratory under agreement number FA8750-20-2-10002. We also thank our colleagues within IESL at UMass Amherst, for their helpful discussions. Michael, Shib and Xiang were supported in part by the Center for Intelligent Information Retrieval and the Center for Data Science, in part by the IBM Research AI through the AI Horizons Network, in part by the University of Southern California subcontract No. 123875727 under Office of Naval Research prime contract No. N660011924032 and in part by the University of Southern California subcontract no. 89341790 under Defense Advanced Research Projects Agency prime contract No. FA8750-17-C-0106. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

References

Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems, NeurIPS*.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 5184–5193. Association for Computational Linguistics.

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 ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

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 (NIPS)*, pages 2787–2795.

Khaoula Boutouhami, Jiatao Zhang, Guilin Qi, and Huan Gao. 2020. Uncertain ontology-aware knowledge graph embeddings. In *Semantic Technology*, pages 129–136. Springer Singapore.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. 2019. Embedding uncertain knowledge graphs. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 3363–3370. AAAI Press.

Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Lorraine Li, and Andrew McCallum. 2020. Improving local identifiability in probabilistic box embeddings. In *Advances in Neural Information Processing Systems*.

Shib Sankar Dasgupta, Xiang Li, Michael Boratko, Dongxu Zhang, and Andrew McCallum. 2021. Box-to-box transformation for modeling joint hierarchies.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655.

Jiafeng Hu, Reynold Cheng, Zhipeng Huang, Yixang Fang, and Siqiang Luo. 2017. On embedding uncertain graphs. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

Natthawut Kertkeidkachorn, Xin Liu, and Ryutaro Ichise. 2019. Gtranse: Generalizing translation-based model on uncertain knowledge graph embedding. In *Annual Conference of the Japanese Society for Artificial Intelligence*, pages 170–178. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Alice Lai and Julia Hockenmaier. 2017. Learning to predict denotational probabilities for modeling entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 721–730. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Hang Li, Tie-Yan Liu, and Chengxiang Zhai. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*.

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. 2019. Smoothing the geometry of probabilistic box embeddings. *ICLR*.

Farzaneh Mahdisoltani, Joanna Biega, et al. 2015. Yago3: A knowledge base from multilingual Wikipedias. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, B Yang, J Betteridge, A Carlson, B Dalvi, M Gardner, B Kisiel, et al. 2018. Never-ending learning. *Communications of the ACM*.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc.

Dhruvesh Patel, Shib Sankar Dasgupta, Michael Boratko, Xiang Li, Luke Vilnis, and Andrew McCallum. 2020. Representing joint hierarchies with box embeddings. *Automated Knowledge Base Construction*.

Hongyu Ren, Weihua Hu, and Jure Leskovec. 2020. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *8th International Conference on Learning Representations*. OpenReview.net.

DN Sosa, A Derry, M Guo, E Wei, C Brinton, and RB Altman. 2020. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, pages 463–474.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, page 4444–4451. AAAI Press.

Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. 2020. Knowledge association with hyperbolic knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5704–5716.

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 (ICLR)*.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. 2016. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 2071–2080. PMLR.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *ICLR*.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 263–272. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *International Conference on Learning Representations (ICLR)*.

A Appendices

A.1 More Implementation Details

Table 6 lists hyper-parameter search space for obtaining the set of used numbers. We performed grid search to choose the final setting.

Hyper-parameters	Search space
Learning rate lr	{0.001, 0.0001, 0.00001}
Embedding dimension d	{30, 64, 128, 300}
Batch size b	{256, 512, 1024, 2048, 4096}
Gumbel box temperature β	{0.1, 0.01, 0.001, 0.0001}
L_2 regularization λ	{0.001, 0.01, 0.1, 1}

Table 6: Search Space for hyper-parameters

The best hyper-parameter combinations for confidence prediction are $\{lr = 0.0001, b = 1024, d = 64, \beta = 0.01\}$, $b = 2048$ for CN15k and $b = 4096$ for NL27k. L_2 regularization is 1 for box sizes in logarithm scale and 0.001 for other parameters. For fact ranking they are $\{lr = 0.0001, d = 300, b = 4096, \lambda = 0.00001\}$, $\beta = 0.001$ for CN15k and $\beta = 0.0001$ for NL27k. The number of negative samples is fixed as 30. Rule weights are empirically set as $w_{tr} = w_{cp} = 0.1$.

Table 7 lists the hardware specifications of the machine where we train and evaluate all models. On this machine, training BEUrRE for the confidence prediction task takes around 1-1.5 hours. Training BEUrRE for the ranking task takes around 1-2 hours for CN15k and 3 hours for NL27k. For the reported model, on CN15k, BEUrRE has around 2M parameters for confidence prediction and 9M parameters for ranking. On NL27k, BEUrRE has 9M parameters for confidence prediction and 17M for ranking.

Hardware	Specification
CPU	Intel® Xeon® E5-2650 v4 12-core
GPU	NVIDIA® GP102 TITAN Xp (12GB)
RAM	256GB

Table 7: Hardware specifications of the used machine