Probabilistic Knowledge Graph Construction:

Compositional and Incremental Approaches

Dongwoo Kim, Lexing Xie & Cheng Soon Ong, Australian National University, Data61

Motivation

- How to measure **uncertainty** of unknown triples given knowledge graph.
- How to incorporate a **graph structure** of knowledge graph into low-rank factorisation to improve unknown triple prediction.
- How to maximise total number of triples while keeping high prediction performance in **incremental knowledge population**.

Contributions

- Propose a **probabilistic formulation** of bilinear tensor factorisation that allows us to predict the uncertainty of unknown triples.
- Incorporate a path structure of knowledge graph into factorisation by modelling a **composition of relations**.
- Develop an incremental population method that searches the factorised space, trading of exploration and exploitation using **Thompson sampling**.

1 Compositional Relational Model

A triple, e.g. {*Obama, president of, US*}, is a basic unit of a knowledge graph. A collection of knowledge triples can be represented as a 3d tensor where each dimension represent entity, relation, and entity, respectively. The goal of statistical relational models is to factorise this tensor to obtain low dimensional representations of entities and relations.

Probabilistic bilinear factorisation model

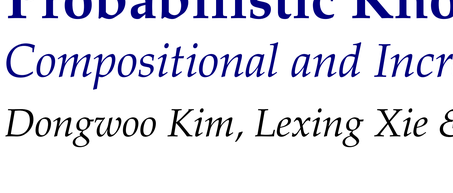

Figure 1: Illustration of widely used bilinear factorisation model, RESCAL, where entities are embedded into D-dimensional latent space.

$$e_i, e_j \in \mathbb{R}^D, \quad R_k \in \mathbb{R}^{D \times D}$$

We reformulate popular RESCAL model in a probabilistic way by placing isotropic normal prior over entity vectors and relation matrices. For observations, we use the normal distribution as in figure 1 (PNORMAL) and logistic function (PLOGIT).

Compositional triples

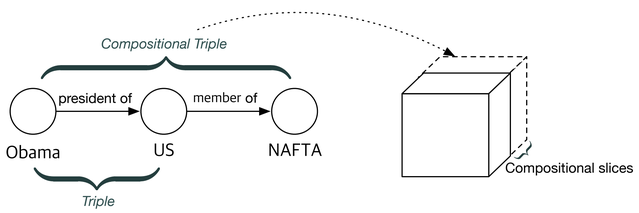

Figure 2: Composition of multiple triples can be interpreted as a **compositional triple**. We augment the original tensor with these additional triples to incorporate explicit path structure into factorisation.

Adding compositional slices will increase the number of relation parameters exponentially. Therefore, we propose two different structures to model the compositionality in the latent embedding spaces.

1. Multiplicative compositionality (PCOMP-MUL):

$$x_{icj} \sim \mathcal{N}(e_i^T R_{c_1} R_{c_1} \dots R_{c_n} e_j, \sigma^2) \tag{1}$$

2. Additive compositionality (PCOMP-ADD):

$$x_{icj} \sim \mathcal{N}(e_i^T \frac{1}{c_n}(R_{c_1} + R_{c_1} + \dots + R_{c_n}) e_j, \sigma^2), \tag{2}$$

where c is a compositional relation with composition of relations $\{c_1, c_2, ..., c_n\}$.

2 Knowledge Completion

We first evaluate our model to **predict unknown triples** given observations to measure the performance of proposed models with all non compositional and compositional variants. For this experiments, we divide datasets into training and testing, and then measure ROC-AUC scores on the test set.

Datasets We use three benchmark datasets for experiments.

Dataset	# rel	# entities	# triples	sparsity
Kinship	26	104	10,790	0.038
UMLS	49	135	6,752	0.008
Nation	56	14	2,024	0.184

Table 1: Description of datasets. Sparsity denotes the ratio of valid triples to invalid triples.

Unknown triple prediction

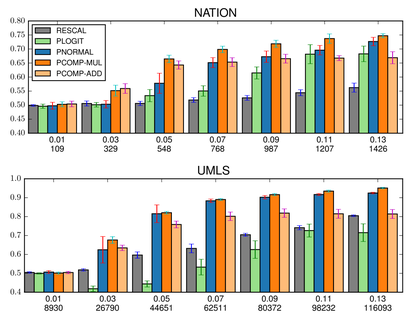

Figure 3: ROC-AUC scores of compositional and non-compositional models. The x-axis denotes the proportion and total number of triples used for training. PNORMAL or PLOGIT generally outperform RESCAL. In general, the multiplicative compositional model (PCOMP-MUL) outperforms the additive compositional model (PCOM-ADD), and performs better the other baseline models.

Compositional path reconstruction

The goal of the compositional models is to factorise triples along with the graph structure as a whole. To show that the model embeds the graph structure into latent space, we evaluate a path reconstruction task where the model predicts a final entity given source entity and sequence of relations.

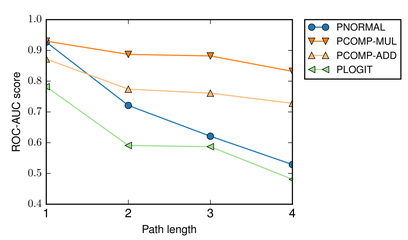

Figure 4: Path reconstruction result of UMLS. The performances of both compositional models remain consistent whereas those of the non-compositional models drop sharply as the length increases.

Visualisation of learned entities

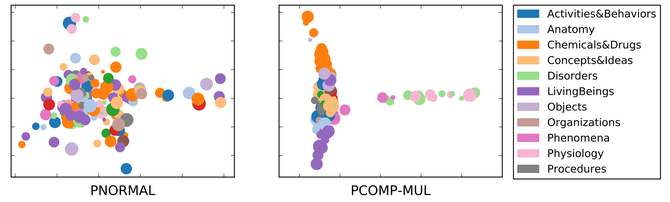

Figure 5: Embedding learned entities of the UMLS dataset into a two-dimensional space through the spectral clustering. Entities with the same type are represented by the same color. The entities with the same type are located closer to each other with the multiplicative compositional model (PCOMP-MUL) than the non-compositional model.

3 Incremental Knowledge Population

The goal of **incremental population** is to maximise the number of positive triples based on the interaction between human experts and labels given a limited amount of budget.

A recent attempt at incremental knowledge population (Jiang et al., 2015), has had difficulties simultaneously achieving high recall and faithful reconstruction. We employ **Thompson sampling**, an approach for solving the multi-armed bandit problem, to find an optimal trade-off between exploration and exploitation during incremental population.

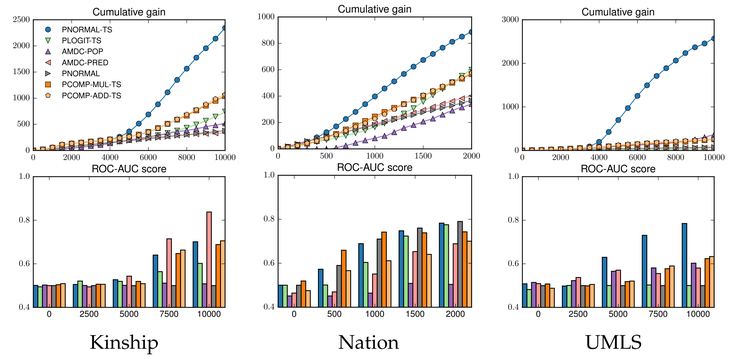

Figure 6: The cumulative gain (upper) and ROC-AUC score (lower) of the Thompson sampling with passive learning and AMDC models. X-axis denotes the number of queries issued. Thompson sampling with PNORMAL model achieves the highest cumulative gain to compare with AMDC and passive learning algorithms and shows comparable performance on ROC-AUC scores. The compositional model performs worse than the non-compositional models.

4 Discussion

Thompson sampling has been studied in the context of multi-armed bandit problems where the goal is to maximise cumulative gains or minimise cumulative regrets over time, whereas its performance on making a predictive model has not been widely discussed so far. Throughout this work, we have empirically shown that maximising cumulative gain entails good predictive models as well.