Probabilistic Knowledge Graph Construction:

Compositional and Incremental Approaches

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

Motivating Questions

- 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**?

Contribution

- 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 Probabilistic Relational Model

- **A triple**, e.g. {*Obama*, *president of*, *US*}, is a basic unit of a knowledge graph
- Collection of knowledge triples can be represented as a 3d tensor
- Statistical relational models factorise the tensor into low dimensional entities and relations

Probabilistic bilinear factorisation model

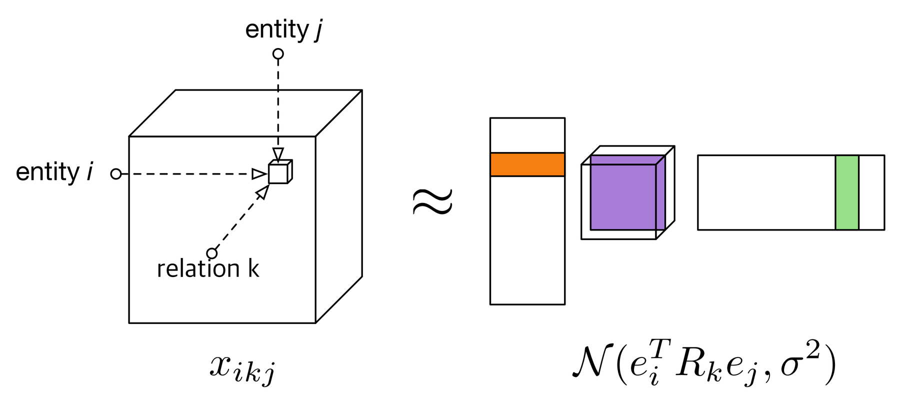

Figure 1: Bilinear factorisation model, RESCAL, where entities are embedded into D-dimensional latent space.

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

We reformulate popular RESCAL model in a probabilistic way by placing isotropic normal prior over entity vectors and relation matrices. For the observation, we design two different models:

1. Normal output (PNORMAL): $x_{ikj} \sim \mathcal{N}(e_i^\top R_k e_j, \sigma_x^2)$
2. Logistic output (PLOGIT): $x_{ikj} \sim \sigma(e_i^\top R_k e_j)$

2 Compositional Relational Model

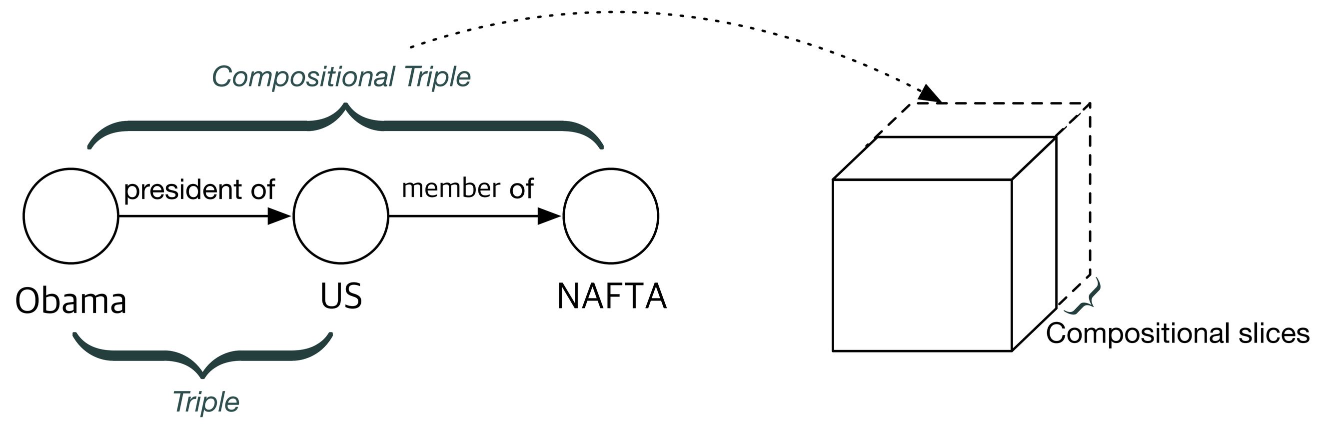

Figure 2: Multiple consecutive triples form a **compositional triple**. We augment the original tensor with compositional additional triples to explicitly incorporate path structure into factorisation.

We design two observation schemes for the compositional model:

1. Multiplicative (PCOMP-MUL): $x_{icj} \sim \mathcal{N}(e_i^\top R_{c_1} R_{c_2} \dots R_{c_n} e_j)$
2. Additive (PCOMP-ADD): $x_{icj} \sim \mathcal{N}(e_i^\top \frac{1}{c_n}(R_{c_1} + \dots + R_{c_n}) e_j)$

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

3 Knowledge Completion

Goal: predict the unobserved part of knowledge graph through the reconstruction of tensor
Method: reconstruct tensor via posterior samples inferred by Gibbs sampling

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.

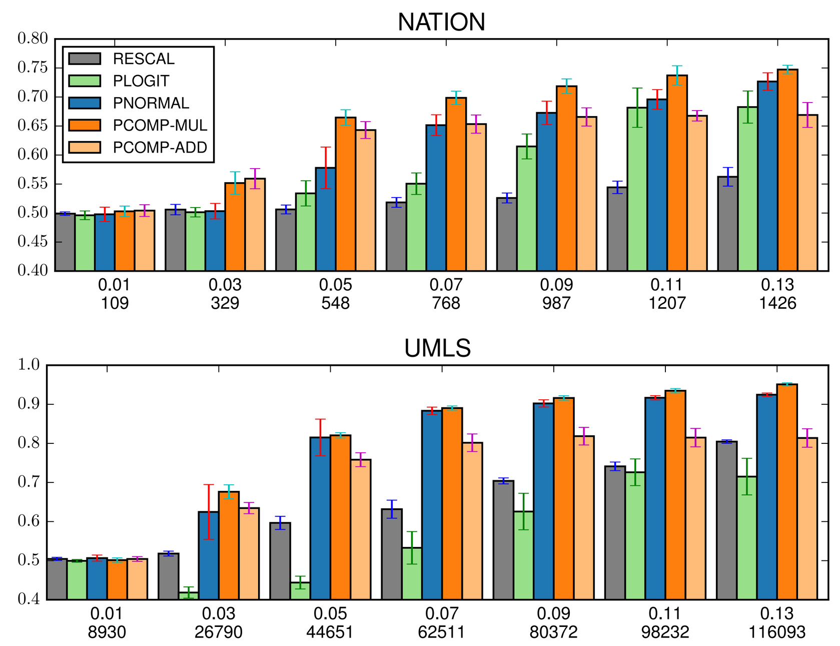

Figure 3: ROC-AUC scores of compositional and non-compositional models. The multiplicative compositional model (PCOMP-MUL) outperforms the other baseline models.

Visualisation of learned entities

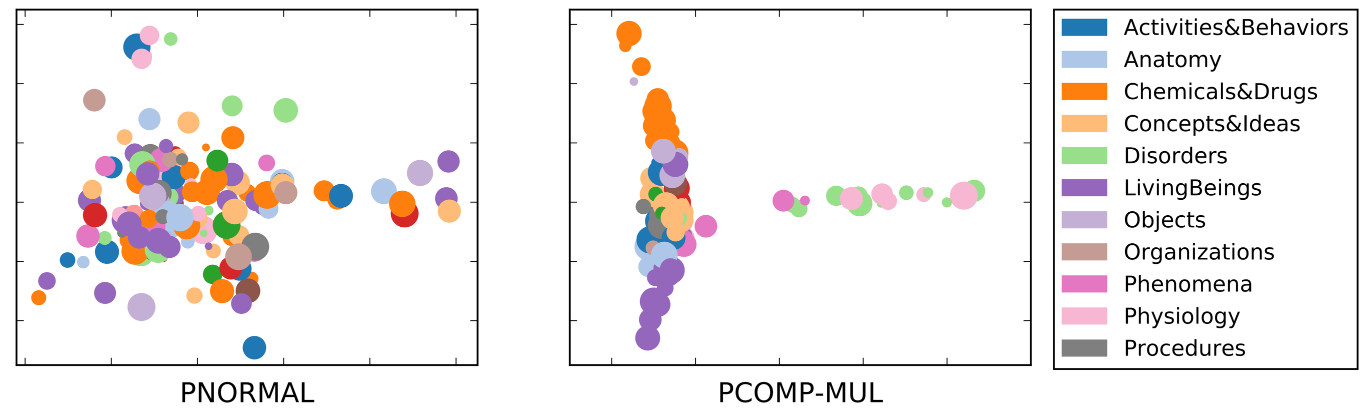

Figure 4: Embedding learned entities of the UMLS dataset into a two-dimensional space through the spectral clustering. The entities of the same type are located closer to each other with the multiplicative compositional model (PCOMP-MUL) than the non-compositional model.

4 Incremental Knowledge Population

Goal: maximise the number of positive triples based on the interaction with human experts given a limited amount of budget
Method: adopt Thompson sampling (TS) from K-armed bandits

- Particle Thompson sampling for population:

1. Sample unobserved triples x_{ikj} from posterior distribution
2. Query the maximum triple & Obtain label from human experts
3. Update posterior using sequential Monte Carlo

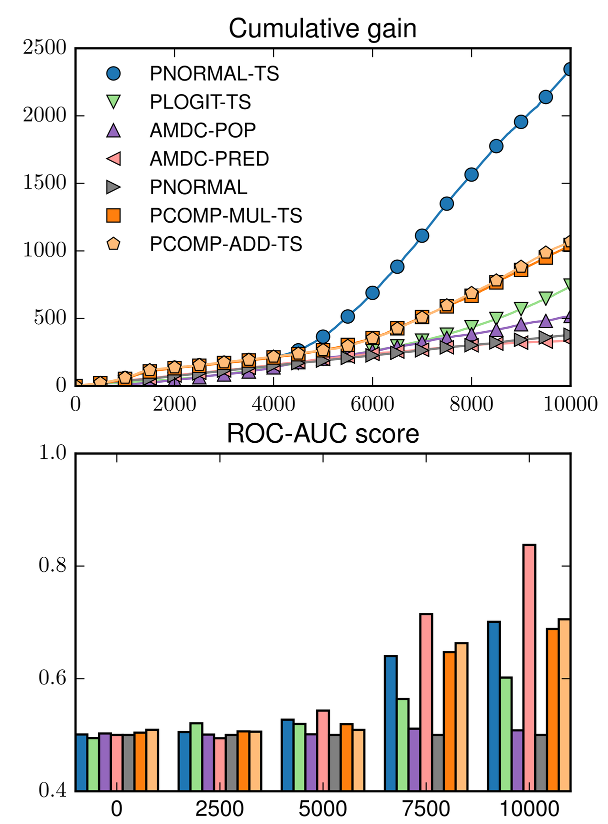

Figure 5: The cumulative gain (upper) and ROC-AUC score (lower) of the Thompson sampling with baseline models. Thompson sampling with PNORMAL model achieves the highest cumulative gain. The compositional model performs worse than the non-compositional models.

- TS has been used to maximise cumulative gains in bandits.
- Maximising cumulative gain entails good predictive models.