Review of TransG

A Generative Model for Knowledge Graph Embedding

Omar Gutiérrez

Contents

Contents

Overview

- Traditional knowledge bases are symbolic and logic.
- Thus, numerical machine learning methods are not compatible with these databases.
- **Solution**: Knowledge Graph Embedding (KGE).
- The idea of KGE is convert entities and relations into continuous vector spaces.
- Several solutions have been proposed: TransE, Latent Factor Model, RESCAL, etc.
- **TransG** was proposed in 2015.
- Its main contribution is introduce and solve the issue of **multiple relation semantics**.

Translation-Based Embedding Methods

- A family of embedding models is known as translation-based.
- **TransE** is the most significant work.
  - Translation principle:

$$h + r \approx t$$

  - Score function:

$$f_r(h, t) = \|h_r + r - t_r\|_2^2$$

- Several modifications have been proposed:
- TransH, TransR, CTransR, TransM, and KG2E

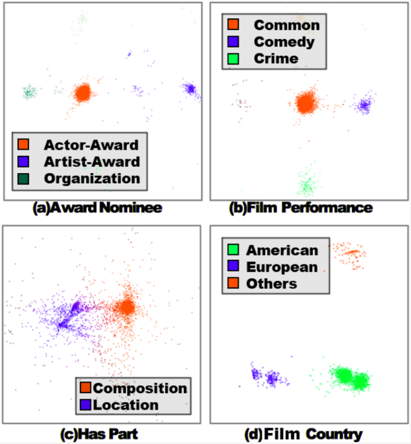
Other Embedding Methods

- **Structured and unstructured embedding**
- **Neural Network based Embedding**: Single Layer Model, Neural Tensor Network
- **Factor models and matrix factorization**: Latent Factor Model, RESCAL

Multiple Relation Semantics

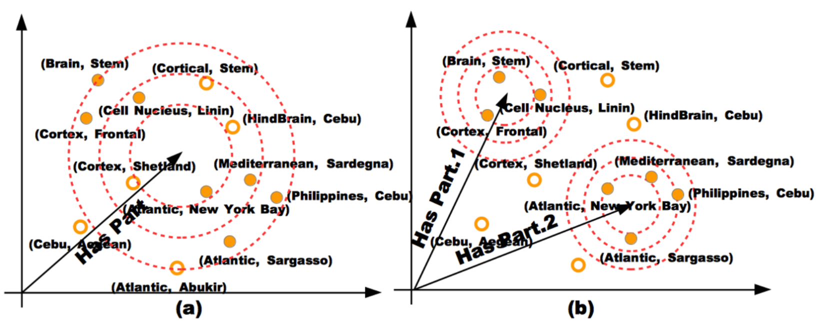
- A relation may have **multiple meanings**, depending on the associated entity pairs:
  - Composition: `<Table, HasPart, Leg>`
  - Location: `<Atlantics, HasPart, NewYorkBay>`
- Why?
  - **Simplification:** Abstracting multiple similar relations into one specific is a common trick
  - **Nature of knowledge:** Ambiguous information
- The ambiguity of knowledge means a semantic mixture

Multiple Relation Semantics II



- This PCA dimension reduction visualization shows TransE embedding vectors
- With TransE there is only one cluster whose center is the vector r

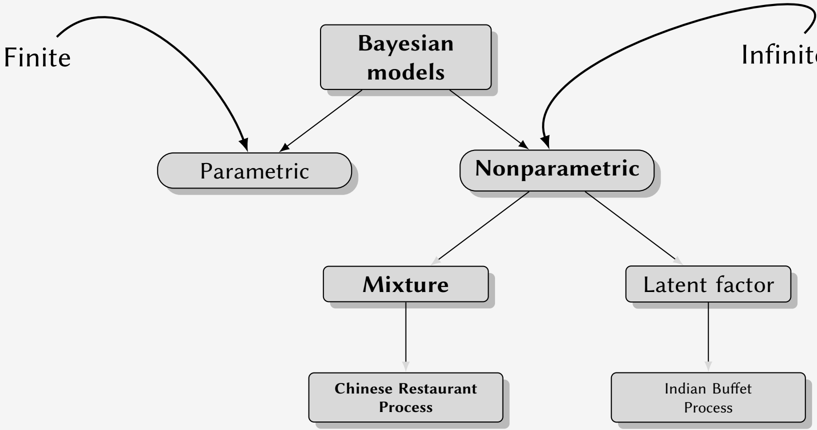
Multiple Relation Semantics III



Contents

Bayesian nonparametric models

TransG is a **Bayesian nonparametric infinite mixture model**



Bayesian parametric vs nonparametric models

- Traditional approach (finite)
  - The number of parameters θ (e.g. clusters) is prespecified
  - We have a prior distribution over parameters $P(\theta)$
  - For example, in the Gaussian mixture model, each cluster will be modelled using a parametric model (e. g. Gaussian)
- Bayesian nonparametric models
  - We assume that there is an **infinite** number of latent clusters
  - A finite number of clusters is *inferred* from data
  - The number of clusters grow as new data points are observed

Chinese Restaurant Process

- Imagine a restaurant with an infinite number of tables,
- and a sequence of customers entering the restaurant and sitting down.
- The first customer enters and sits at the first table
- The second customer enters and sits...
  - at the first table with probability $\frac{1}{1+\alpha}$
  - at the second table with probability $\frac{\alpha}{1+\alpha}$
- We realize that CRP is a form of clustering: K groups and each group with size N_k

Generative Process

1. For an entity $e \in E$:
   1.1 Draw each entity embedding mean vector from a standard normal distribution as a prior: $\mathbf{u_e} \backsim \mathcal{N}(\mathbf{0}, \mathbf{1})$.
2. For a triple $(h, r, t) \in \Delta$:
   2.1 Draw a semantic component from Chinese Restaurant Process for this relation: $\pi_{r,m} \sim CRP(\beta)$.
   2.2 Draw a head entity embedding vector from a normal distribution: $\mathbf{h} \backsim \mathcal{N}(\mathbf{u_h}, \sigma_h^2\mathbf{E})$.
   2.3 Draw a tail entity embedding vector from a normal distribution: $\mathbf{t} \backsim \mathcal{N}(\mathbf{u_t}, \sigma_t^2\mathbf{E})$.
   2.4 Draw a relation embedding vector for this semantics: $\mathbf{u_{r,m}} = \mathbf{t} - \mathbf{h} \backsim \mathcal{N}(\mathbf{u_t} - \mathbf{u_h}, (\sigma_h^2 + \sigma_t^2)\mathbf{E})$.

- Translation principle from $h + r \approx t$ becomes

$$m^*_{(h,r,t)} = \arg\max_{m=1\ldots M_r} \left(\pi_{r,m} e^{-\frac{||\mathbf{u_h} + \mathbf{u_{r,m}} - \mathbf{u_t}||_2^2}{\sigma_h^2 + \sigma_t^2}} \right)$$

$$\mathbf{h} + \mathbf{u}_{\mathbf{r},\mathbf{m}^*_{(\mathbf{h},\mathbf{r},\mathbf{t})}} \approx \mathbf{t}$$

- Notice that r is has a selective translation

Contents

Experiments

- The experiments were carried out in benckmark datasets subsets of Freebase and Wordnet
  - WN18, WN11, FB13, FB15K
  - Raw setting:
  - Filter setting: corrupted triples are filtered from test dataset.

Link Prediction

Datasets	**WN18**				**FB15K**			
Metric	Mean Rank		HITS@10(%)		Mean Rank		HITS@10(%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	263	251	75.4	89.2	243	125	34.9	47.1
TransH	401	388	73.0	82.3	212	87	45.7	64.4
TransR	238	225	79.8	92.0	198	77	48.2	68.7
CTransR	**231**	**218**	79.4	92.3	199	75	48.4	70.2
PTransE	N/A	N/A	N/A	N/A	207	58	51.4	84.6
KG2E	362	348	80.5	93.2	183	69	47.5	71.5
TransG	357	345	**84.5**	**94.9**	**152**	**50**	**55.9**	**88.2**

Figure: Evaluation results on link prediction

- Lower **Mean Rank** and higher **HITS@10** mean better performance

Triple Classification

Methods	WN11	FB13	AVG.
NTN	70.4	87.1	78.8
TransE	75.9	81.5	78.7
TransH	78.8	83.3	81.1
TransR	85.9	82.5	84.2
CTransR	85.7	N/A	N/A
KG2E	85.4	85.3	85.4
TransG	**87.4**	**87.3**	**87.4**

Figure: Triple classification accuracy

Contents

Conclusions

- A solution for a new issue, commonly seen in knowledge graph, multiple relation semantics was presented with TransG
- Authors claim **substantial improvements** against the state-of-the-art baselines
  - Still, TransE has 234 references and TransG has 1 reference
  - O(TransG) = (M x O(TransE)), TransE is one of the fastest methods for
- We still have many letters in the alphabet to write more versions of Translation-Based methods ;)
  - For example, use another BNP to model the issue presented here.