Web-based Knowledge Extension

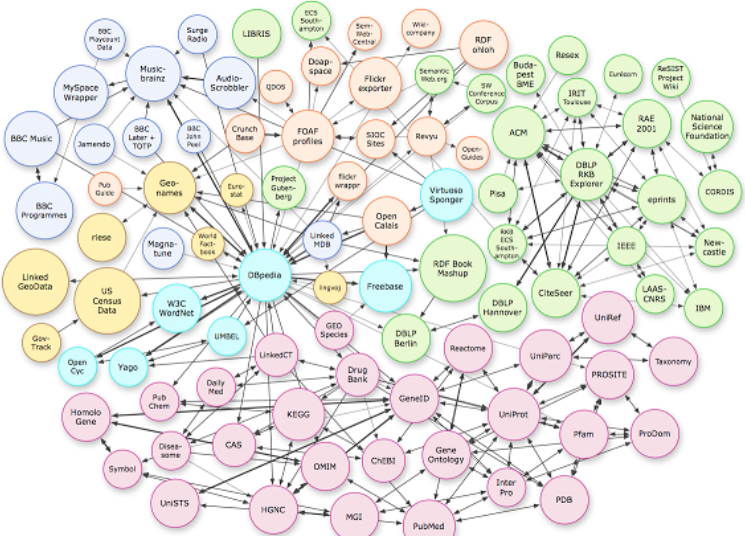
Richong Zhang

zhangrc@act.buaa.edu.cn

Beihang University, China

May 15, 2019

Linked Data



- As of 2012, Google has over 570 million objects and 18 billion facts.
- As of 2016, Google has over 70 billion facts.

Knowledge Bases

- Emerging knowledge bases (KB): Freebase, YAGO, DBpedia, etc.
- A KB contains a collection of facts in the form of (h, r, t)
  - h: head/subject entity
  - r: relation
  - t: tail/object entity

Example (Factual Triple in KB)

(Beijing, isCapitalOf, China)

- h: Beijing
- r: isCapitalOf
- t: China

- KB can be interpreted as edge-labelled graph

entity	$\Rightarrow$	vertex
triple	$\Rightarrow$	edge
relation in triple	$\Rightarrow$	edge label

Example

$$\text{Triple } (h, r, t) \Rightarrow \text{edge } h \xrightarrow{r} t$$

- KB can be interpreted as edge-labelled graph

entity	$\Rightarrow$	vertex
triple	$\Rightarrow$	edge
relation in triple	$\Rightarrow$	edge label

Example

$$\text{Triple } (h, r, t) \Rightarrow \text{edge } h \xrightarrow{r} t$$

Our Belief

"Such a structured knowledge representation is fundamental for developing AI."

- Can AI retrieve information from KB?
- Can AI generalize knowledge from KB? ("link prediction")
- Can AI reason using KB?

KB Embedding

- relations/entities $\Rightarrow$ representations in a Euclidean space
- preserves intra-relational and inter-relational structures

Idea

In the Euclidean space,

$$\overrightarrow{\text{Ottawa}} \text{ w.r.t. } \overrightarrow{\text{Canada}} \equiv \overrightarrow{\text{Beijing}} \text{ w.r.t. } \overrightarrow{\text{China}}$$
$$\overrightarrow{\text{CND}} \text{ w.r.t. } \overrightarrow{\text{Canada}} \equiv \overrightarrow{\text{RMB}} \text{ w.r.t. } \overrightarrow{\text{China}}$$

- KB embedding converts discrete topology to a continuous one
- $\Rightarrow$ avoids combinatorial complexity of algorithms
- $\Rightarrow$ potentially benefits all areas of KB research

Binary Relations

instance: "Beijing is the capital of China"

$$\Downarrow$$

as a triple: (Beijing, isCapitalOf, China)

$$\Downarrow$$

as a length-2 vector: $(\text{Beijing}, \text{China}) \in \text{isCapitalOf} \subseteq \mathcal{N} \times \mathcal{N}$

$\mathcal{N}$: the set of all entities

Insight

A binary relation is a subset of the cartesian product $\mathcal{N} \times \mathcal{N}$.

Prior Art of Modelling

- TransE [Bordes et al., 2013]
- TransH[Wang et al., 2014]
- TransR [Lin et al., 2015]
- ProjE[Shi and Weninger, 2017]
- ComplEx[Trouillon et al. 2016]
- ConvE[Dettmers et al., 2017]
- Analogy [Liu et al., 2017]...

Prior Art of Modelling

- TransE [Bordes et al., 2013]
- TransH[Wang et al., 2014]
- TransR [Lin et al., 2015]
- ProjE[Shi and Weninger, 2017]
- ComplEx[Trouillon et al. 2016]
- ConvE[Dettmers et al., 2017]
- Analogy [Liu et al., 2017]...

Note

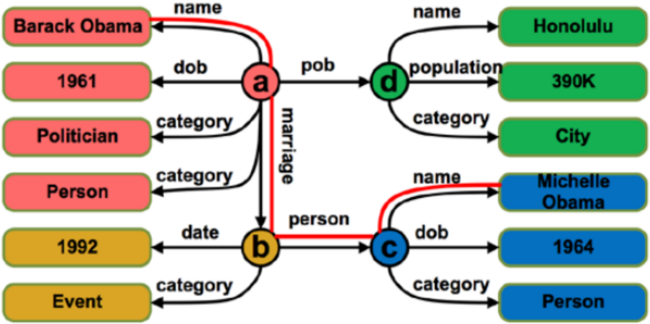
- All assume binary relations
- Trained and tested on FB15K, WN18, FB15K-237, WN18RR

Multi-fold Relation

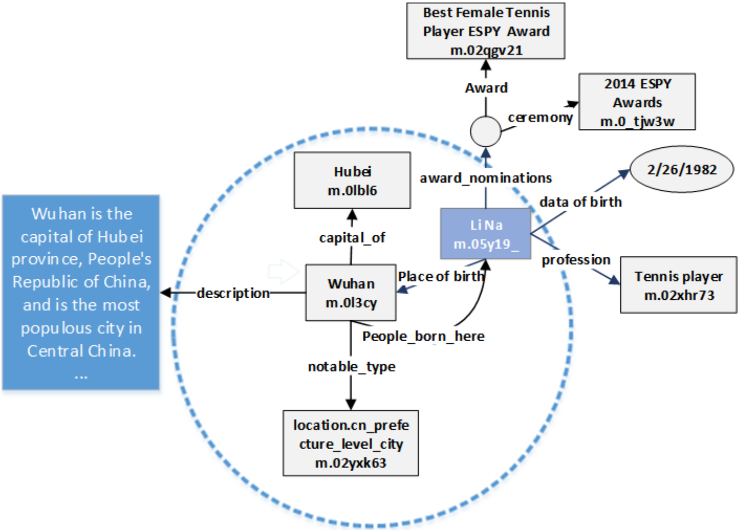
Example

How to represent the fact: “Obama and Michelle were married on October 3, 1992 at Trinity United Church of Christ in Chicago, Illinois.”.

Multi-fold Relation Example



Knowledge Base: Detailed Example



Multi-Fold Relation: Definition

instance: "Jackie Chan played Lee in Rush Hour"

$$\Downarrow$$

$$(\text{JackieChan}, \text{Lee}, \text{RushHour}) \in \text{MovieActing} \subseteq \mathcal{N}^3 := \mathcal{N} \times \mathcal{N} \times \mathcal{N}$$

Insight (Codd, 1970)

A J-fold (or J-ary) relation is a subset of the J-fold cartesian product $\mathcal{N}^J := \underbrace{\mathcal{N} \times \mathcal{N} \times \ldots \times \mathcal{N}}_{J \text{ times}}$.

Definition (Multi-Fold Relation)

Let $\mathcal{M}$ be a set of *roles* in the KB, and a *(multi-fold) relation* R on $\mathcal{N}$ with roles $\mathcal{M}$ is a subset of $\mathcal{N}^{\mathcal{M}}$.

This Talk

Motivation:

- Non-binary relations are ubiquitous.
- Around 75% and 71% people entities do not have nationality and birth place.
- Over 1/3 Freebase entities participate in non-binary relations.

We take a fundamental look at the following questions:

- How to represent multi-fold relational data?
- How to embed multi-fold relational data?
- How to complete multi-fold relational data?

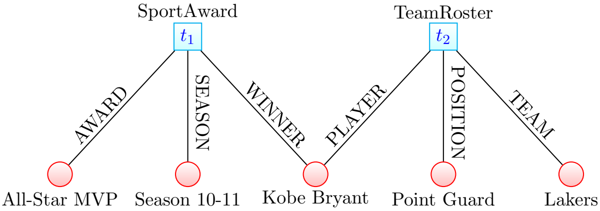
Roadmap of This Talk

Representation of *m-fold* Relations

J. Wen et. al: On the Representation and Embedding of Knowledge Bases beyond Binary Relations. IJCAI 2016

Representation:Instance Graph

circle vertex	$\leftrightarrow$	entity
square vertex	$\leftrightarrow$	instance
edge	$\leftrightarrow$	entity participating in an instance
edge label	$\leftrightarrow$	role
instance vertex annotation	$\leftrightarrow$	relation type

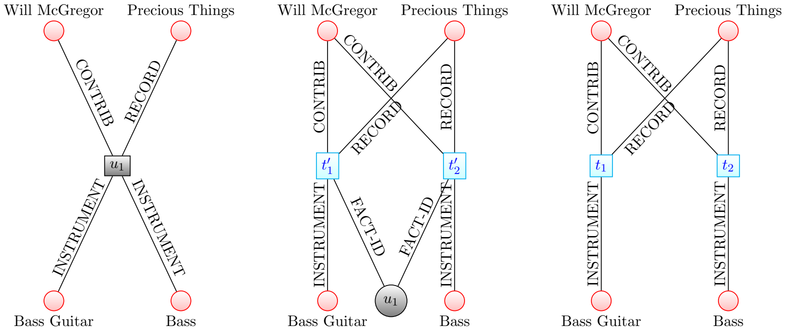


$t_1 \in$ SportAward: "Kobe Bryant is the All-Star MVP for 2010-2011"
$t_2 \in$ TeamRoster: "Kobe Bryant is a Point Guard in Lakers"

Fact Representation to Instance Representation

Example

“Will McGregor played bass and bass guitar in Precious Things.”

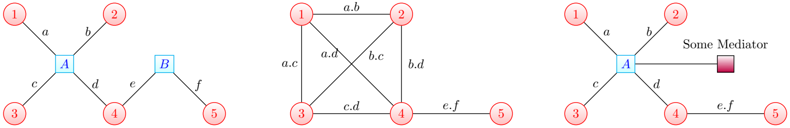


Fact Graph $\mathcal{F}$ Inst. Graph $T_{\text{id}}(\mathcal{F})$ Inst. Graph $T(\mathcal{F})$

Lemma

Fact representation $\mathcal{F}$ can be recovered from $T_{\text{id}}(\mathcal{F})$.

Freebase Representation



Fact graph   S2C-Conversion on $A\&B$ Freebase

Lemma

After applying S2C conversions to a graph, in general the graph is no longer recoverable.

Insight

Freebase contains equivalent information as a fact representation. Edges (triples) in Freebase have three different semantics. Not clean!

Insight

Fact graphs and instance graphs are both superior to S2C-converted graphs or Freebase representation, with instance graphs more

Standard dataset: FB15K [Bordes et al., 2013]

- Extracted from Freebase
- S2C applied to every CVT, converting non-binary relations to binary
  - $\Rightarrow$ loss of structural information
- Mediators are not filtered out and have participated in S2C conversions.
  - $\Rightarrow$ introduced additional noise

Insight

FB15K is not suitable for embedding KBs containing non-binary relations.

JF17K Dataset: Construction

- Download full Freebase
- Remove entities involved in very few triples
- Remove triples involving *String*, *Enumeration Type* and *Numbers*
- Construct fact representation
- Remove meta-relation containing a single role
- Randomly select 10,000 facts from each meta-relations containing more than 10,000 facts $\Rightarrow$ fact representation $\mathcal{F}$.
- Construct two instance representations $T_{\text{id}}(\mathcal{F})$ and $T(\mathcal{F})$
- Filter $T(\mathcal{F})$ so that each entity participates in at least 5 instances $\Rightarrow$ instance representation $\mathcal{G}$
- Filter $T_{\text{id}}(\mathcal{F})$ correspondingly $\Rightarrow$ instance representation $\mathcal{G}_{\text{id}}$
- Construct S2C($\mathcal{G}$) $\Rightarrow$ instance representation $\mathcal{G}_{\text{s2c}}$

Note

JF17K contains three consistent datasets: $\mathcal{G}$, $\mathcal{G}_{\text{id}}$, $\mathcal{G}_{\text{s2c}}$
Available at `github.com/wenjf/multi-relational_learning`.

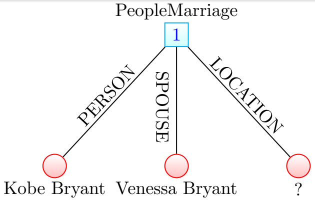
JF17K: Statistics

	$\mathcal{G}^{\checkmark}/\mathcal{G}^{\checkmark}_{\mathrm{id}}$	$\mathcal{G}^{\checkmark}_{\mathrm{s2c}}$	$\mathcal{G}^{?}/\mathcal{G}^{?}_{\mathrm{id}}$	$\mathcal{G}^{?}_{\mathrm{s2c}}$
# of entities	17629	17629	12282	12282
# of instances/triple types	181	381	159	336
# of instances/triples	139997	254366	22076	52933

Note

JF17K has similar scale/statistics as FB15K.

Knowledge Completion via Type-Augmented Embedding



Zhang et. al: Embedding of Hierarchically Typed Knowledge Bases. AAAI 2018

Embedding of Hierarchically Typed Knowledge Bases

Hierarchical Types

- Exploit entity type information in knowledge base embedding.
- A framework augments a “typeless” embedding model.

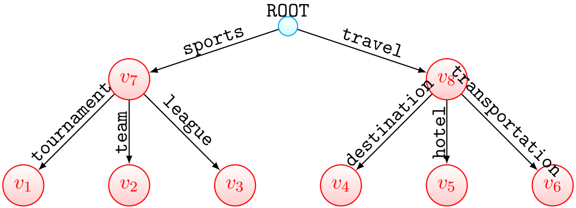


Figure: The tree Γ of types in the toy example.

Type as a set of entities

- Every node in the tree can be understood as a “type”.
- Every node in the tree is interpreted as a constraint.

Type Space

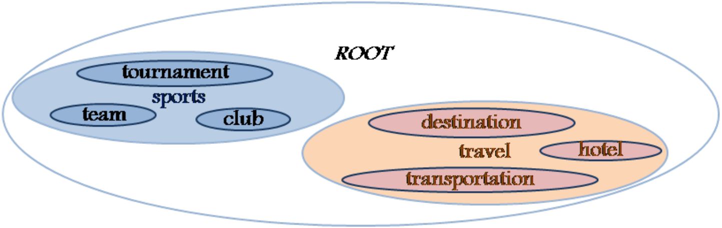


Figure: Type space in our model

Note

- Each Type is a set of entities.
- Each type is a constraint on the entity.
- Each type (tree node) is mapped to a subset of the embedding space.
- Each such subset is chosen as an affine subspace.

Entity-Type Cost

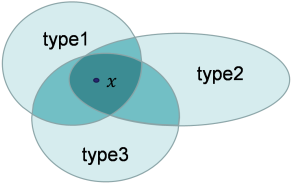


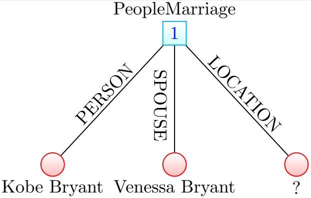
Figure: Entity-type model

Note

- An entity lives in the intersection of the subspaces.
- An entity x should satisfy all its type, and the entity-type cost function defined as follows:

$$G(\phi, \Omega) := \sum_{x \in \mathcal{N}} \left(\sum_{v \in L(x)} g_v(\phi(x)) + \sum_{v' \in L^-(x)} [T_{\mathrm{ET}} - g_{v'}(\phi(x))]_+ \right) \tag{1}$$

Knowledge Completion via Locality-Expanded Embedding



F. Kong et. al: LENA: Locality-expanded Neural Embedding AAAI 2019

Example of Neighbourhood Information

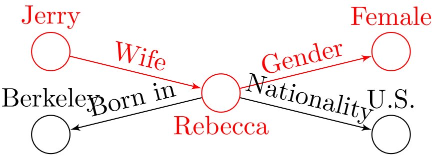


Figure: A subgraph of Rebecca.

- "Rebecca is the wife of Jerry" is relevant to "Rebecca's gender is female"
- "Rebecca was born in Berkeley" is useful for predicting "the Nationality of Rebecca is U.S."
- "Rebecca is the wife of Jerry" is irrelevant to "the nationality of Rebecca is U.S."

Insight

The "modelling locality" can be expanded from edges to larger graph neighbourhoods.

Example of Neighbourhood Information

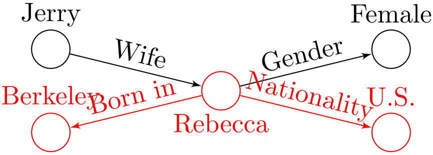


Figure: A subgraph of Rebecca.

- "Rebecca is the wife of Jerry" is relevant to "Rebecca's gender is female"
- "Rebecca was born in Berkeley" is useful for predicting "the Nationality of Rebecca is U.S."
- "Rebecca is the wife of Jerry" is irrelevant to "the nationality of Rebecca is U.S."

Insight

The "modelling locality" can be expanded from edges to larger graph neighbourhoods.

Example of Neighbourhood Information

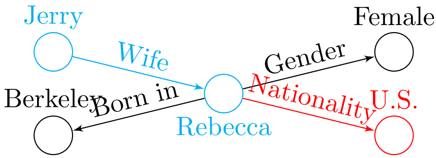


Figure: A subgraph of Rebecca.

- “Rebecca is the wife of Jerry” is relevant to “Rebecca’s gender is female”
- “Rebecca was born in Berkeley” is useful for predicting “the Nationality of Rebecca is U.S.”
- “Rebecca is the wife of Jerry” is irrelevant to “the nationality of Rebecca is U.S.”

Insight

The “modelling locality” can be expanded from edges to larger graph neighbourhoods.

Model

Probabilistic Model

$$p(t|h,r) = \frac{\exp(s(h,r,t))}{\sum_{t' \in \mathcal{N}} \exp(s(h,r,t'))}. \tag{2}$$

Embedding

- We embed entities and relations both as vectors in $\mathbb{R}^k$.
- D_{E} and D_{R} are $k \times |\mathcal{N}|$ matrix
- $x \in \mathcal{N}$ and $r \in \tilde{\mathcal{R}}$ are one-hot vectors

$$\mathbf{x} := D_{\mathrm{E}} x \tag{3}$$

$$\mathbf{r} := D_{\mathrm{R}} r \tag{4}$$

Score Function

$$\begin{aligned} s(h,r,t) := &\langle v^{\mathrm{E}}(h,r,t) + \mathbf{r} + b_{\mathrm{E}}, C_{\mathrm{E}}\mathbf{t}\rangle \\ &+ \langle v^{\mathrm{R}}(h,r,t) + \mathbf{h} + b_{\mathrm{R}}, C_{\mathrm{R}}\mathbf{t}\rangle \end{aligned} \tag{5}$$

Neighbourhood Graph

Neighbourhood

$$\mathcal{G}(h,r,t) := \{e \in \mathcal{G} : t(e) = h, e \neq (t, r^-, h)\}.$$

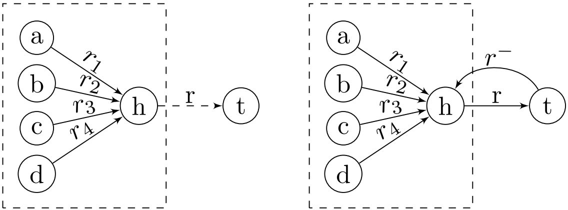
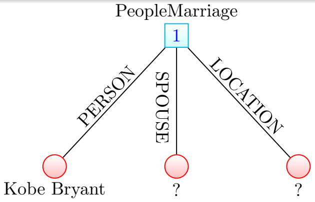


Figure: Example of neighbourhood graphs $\mathcal{G}(h,r,t)$ (the subgraphs in the dashed boxes) of triple (h,r,t). Triples in $\mathcal{G}$ are represented by a solid edge, and triples (e.g., candidate triples) not in $\mathcal{G}$ are represented by a dashed edge.

Knowledge Instance Re-construction



Zhang et. al: Scalable Instance Reconstruction in Knowledge Bases via Relatedness Affiliated Embedding. WWW 2018

Link Prediction is not Enough!

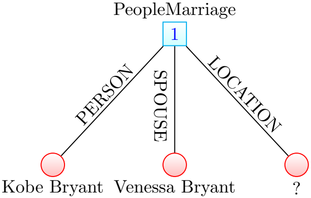


Figure: Link prediction vs. instance reconstruction

Instance Reconstruction

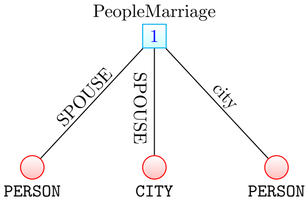
Definition

- To recover an instance in which entities are missing from all but one role.

Note

- Recover t from x^* , x^* will be referred to as the *key* for the problem
- The complexity of instance reconstruction is $O(\mathcal{N}^{m-1})$
- The primary challenge is to develop a scalable reconstruction algorithm

Step 1: Filtering



Schema-based Filtering

- Leveraging the type requirements on the entities dictated by the schema of a relation to reduce the number of entities to be considered in forming instances.

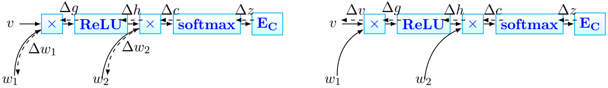


Figure: (left) Update of MLP; (right) Update of Entity Embedding

Relatedness Filtering

- To predict if two entities are related.
- Returns the set of entity pairs (x, y) that the two entities in any pair are thought to be related.
- iterative learn the embedding model and MLP classification model.

Step 2: Splicing

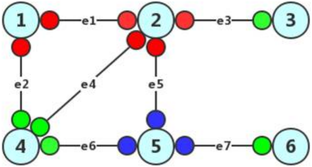


Figure: Edge-end-Colored Graph

Edge-End-Colored Graph (EECG)

- Two end points of each edge are colored by two distinct colors
- Every entity is interpreted as a vertices.
- Every related entity pair can be interpreted as an edge.
- The role under relation is interpreted as the color .

Step 2: Splicing

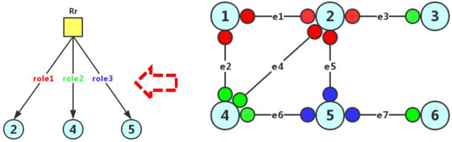


Figure: Color-matched Clique

Color-matched Clique (CMC)

- A sub-EECG and complete.
- The color set of every vertex is a singleton set.
- Every two vertices have different colors.

Observations

Freebase Types (23425 types)

(m.01xxvxx,type.object.id,“/freebase/type_profile/featured$_{t}opics''$)
(m.011nd5wd,type.object.id,“/freebase/type_profile/ownership”)
/astronomy/star/planet$_{s}isusedforlistingplanetsaroundastar$

Domains

/business - the ID of the Business domain
/music - the Music domain
/film - the Film domain

Properties

Property	Expected Type	Description
actor	Person	An actor, e.g. in tv, radio etc.
countryOfOrigin	Country	The country of the principal offices
director	Person	A director of e.g. tv, radio etc.

Concluding Remarks

Representation

- Relations in the real world are often multi-fold.
- FB15K is no longer suited for embedding multi-fold relations.

Embedding

- Direct modelling is a superior framework.
- Type and structural information is useful for embedding.

Reconstruction

- KB containing n-ary relations are challenged by instance reconstruction problem.
- SIR algorithm has significantly reduced complexity for solving instance reconstruction.

Thank you