Traversing Knowledge Graphs in Vector Space

Kelvin Guu, John Miller, Percy Liang (2015)

Presented by Ben Striner 10/17/2017

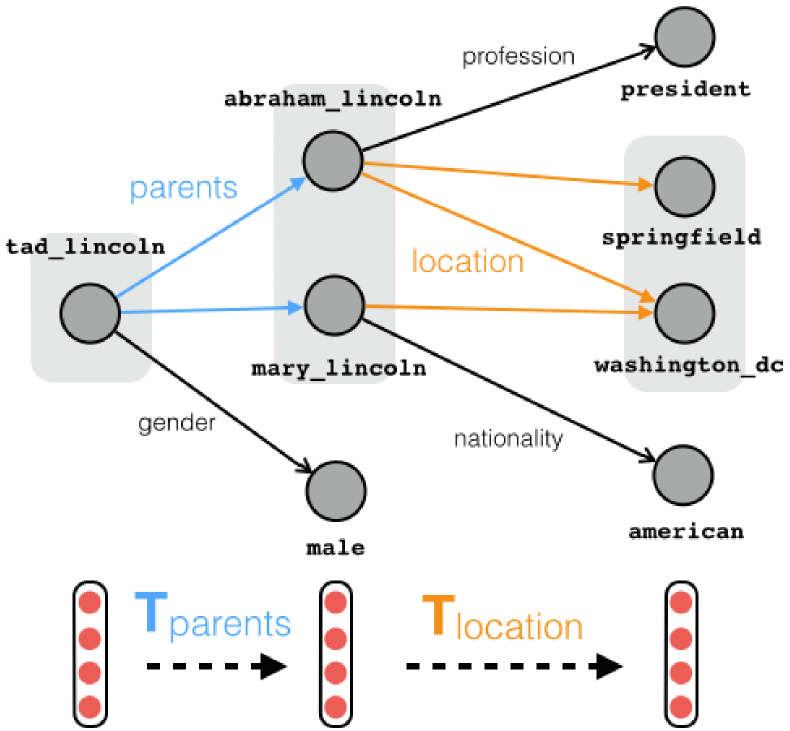
Contents

- What is the dataset?
- What is the task?
- What is their model?
- How does it perform?

Dataset

Knowledge Graph

- Entities are nodes and relationships are labeled edges
- Tuples (s,r,t): (tad_lincoln, parent, abe_lincoln)



Dataset

- WordNet and Freebase
  - Dataset of entities and relationships
  - Some edges withheld for testing

		WordNet	Freebase
Relations		11	13
Entities		38,696	75,043
Base	Train	112,581	316,232
	Test	10,544	23,733
Paths	Train	2,129,539	6,266,058
	Test	46,577	109,557

Table 1: WordNet and Freebase statistics for base and path query datasets.

Tasks

Tasks

- Path Query
  - Given start point and series of relationships predict target
  - Tad_lincoln/parent/location = DC
  - Using multi-step paths generated by walking base dataset
- Knowledge Base Completion
  - Predict whether an edge exists or not
  - Formulated as single-edge path query
  - Using base dataset

Path Query

- A query (q) consists of an "anchor entity" (s) and a path (p)
- A path is a series of relationships $p = (r_1, \ldots, r_k)$
- The answer is the "denotation" $[\![q]\!]$
- Defined recursively

$$[\![s]\!] \stackrel{\text{def}}{=} \{s\}, \tag{1}$$

$$[\![q/r]\!] \stackrel{\text{def}}{=} \{t : \exists s \in [\![q]\!], (s, r, t) \in \mathcal{G}\}. \tag{2}$$

Path Query Evaluation

- C: Candidate answers “type match”
  - Participate in final relationship at least once
  - For example, all entities that are the target of a “located at” relationship would identify most valid locations
- N(q): Incorrect candidate answers

$$\mathcal{C}\left(s/r_1/\cdots/r_k\right) \stackrel{\text{def}}{=} \{t \mid \exists e, (e, r_k, t) \in \mathcal{G}\} \quad (3)$$

$$\mathcal{N}\left(q\right) \stackrel{\text{def}}{=} \mathcal{C}\left(q\right) \setminus [\![q]\!]. \quad (4)$$

Mean Quantile

- Evaluate fraction of incorrect answers are ranked after correct answer

$$\frac{\left|\left\{t' \in \mathcal{N}(q) : \mathrm{score}(q, t') < \mathrm{score}(q, t)\right\}\right|}{|\mathcal{N}(q)|} \quad (13)$$

Knowledge Base Completion Evaluation

- Evaluate accuracy versus negative samples
- For comparison to previous work (Socher)

Models

Modelling Traversal and Membership

- Traversal operator determines the set that can be reached from xs
- Membership operator determines if xt is in the set reached from xs
- Defined recursively

$$\mathbb{T}_{r_i}(v) = v^\top W_{r_i}$$

$$\mathbb{M}(v, x_t) = v^\top x_t$$

$$\text{score}(s/r, t) = \mathbb{M}(\mathbb{T}_r(x_s), x_t) \tag{7}$$

$$[\![s]\!]_{\mathrm{V}} \stackrel{\text{def}}{=} x_s, \tag{8}$$

$$[\![q/r]\!]_{\mathrm{V}} \stackrel{\text{def}}{=} \mathbb{T}_r\left([\![q]\!]_{\mathrm{V}}\right). \tag{9}$$

$$\text{score}(q, t) = \mathbb{M}([\![q]\!]_{\mathrm{V}}, [\![t]\!]_{\mathrm{V}}). \tag{10}$$

Objective Function

- Use Max-Margin loss against incorrect answers that "type match"
- Use paths of different lengths

$$J(\Theta) = \sum_{i=1}^{N} \sum_{t' \in \mathcal{N}(q_i)} \left[1 - \text{margin}(q_i, t_i, t')\right]_+ ,$$

$$\text{margin}(q, t, t') = \text{score}(q, t) - \text{score}(q, t'),$$

$$\Theta = \{\mathbb{M}\} \cup \{\mathbb{T}_r : r \in \mathcal{R}\} \cup \left\{x_e \in \mathbb{R}^d : e \in \mathcal{E}\right\} .$$

Experimental models

- Model Traversal and Membership functions three ways
  - Bilinear
  - TransE
  - Bilinear-Diag
- Also use NTN (for some experiments)

Bilinear Model (Nickel et al., 2011)

- Traditional queries can be answered by chaining matrix multiplication
  - Entities are one-hot indicator vectors
  - Relationships are adjacency matrices
  - xA = all nodes connected to node x by adjacency A (vector)
  - y^Tx = 1 if x is in set y else 0 (scalar)
- Build a similar model with continuous learned representations

$$\text{score}(s/r, t) = x_s^\top W_r x_t. \tag{5}$$

$$\text{score}(q, t) = x_s^\top W_{r_1} \dots W_{r_k} x_t. \tag{6}$$

TransE (Bordes et al., 2013)

- Every entity and relationship embedded as a vector
- Traversal is addition, Membership is L2 distance

$$\mathrm{score}(s/r, t) = -\|x_s + w_r - x_t\|_2^2. \qquad (11)$$

$$\mathbb{M}(v, x_t) = -\|v - x_t\|_2^2 \qquad (12)$$

$$\mathbb{T}_r(x_s) \quad = \quad x_s \ + \ w_r$$

$$\mathrm{score}(q, t) = -\|x_s + w_{r_1} + \cdots + w_{r_k} - x_t\|_2^2.$$

Bilinear-Diag Model (Yang et al., 2015)

- Same as Bilinear model but matrices are diagonal
  - Can no longer interpret weight matrices as adjacency matrices
  - Same number of parameters as TransE
- https://arxiv.org/pdf/1412.6575.pdf

Experiments

Results

- Models trained on single edges perform poorly even when all edges have been seen during training

		Bilinear			Bilinear-Diag			TransE		
Path query task		SINGLE	COMP	(%red)	SINGLE	COMP	(%red)	SINGLE	COMP	(%red)
WordNet	MQ	84.7	89.4	**30.7**	59.7	90.4	**76.2**	83.7	93.3	**58.9**
	H@10	43.6	54.3	**19.0**	7.9	31.1	**25.4**	13.8	43.5	**34.5**
Freebase	MQ	58.0	83.5	**60.7**	57.9	84.8	**63.9**	86.2	88	**13.0**
	H@10	25.9	42.1	**21.9**	23.1	38.6	**20.2**	45.4	50.5	**9.3**
KBC task		SINGLE	COMP	(%red)	SINGLE	COMP	(%red)	SINGLE	COMP	(%red)
WordNet	MQ	76.1	82.0	**24.7**	76.5	84.3	**33.2**	75.5	86.1	**43.3**
	H@10	19.2	27.3	**10.0**	12.9	14.4	**1.72**	4.6	16.5	**12.5**
Freebase	MQ	85.3	91.0	**38.8**	84.6	89.1	**29.2**	92.7	92.8	**1.37**
	H@10	70.2	76.4	**20.8**	63.2	67.0	**10.3**	78.8	78.6	-0.9

Table 2: **Path query answering and knowledge base completion.** We compare the performance of single-edge training (SINGLE) vs compositional training (COMP). MQ: mean quantile, H@10: hits at 10, %red: percentage reduction in error.

Implementation Details

- For each query, sample 10 negative entities
- Entity vectors constrained to unit ball
- Gradient clipping, Minibatch of 300, AdaGrad
- Train on length 1 until convergence, then train on full
- Explicitly parameterized inverse relationships (parent = child^-1)
  - Exclude trivial queries where exact inverse was in training
- Experiment with parameterizing entities with word embeddings

Generating Paths

- Generate queries by random walks
  - Uniform sample of path length and start point
  - Uniform sample of available relationships
  - Uniform sample of next node given that relationship
- Large amounts of training data generated

Deduction vs Induction

- Deduction
  - Entities and relations seen during training, just not the exact query
- Induction
  - Required edge not seen during training

Path query task		WordNet		Freebase	
		Ded.	Ind.	Ded.	Ind.
Bilinear	SINGLE	96.9	66.0	49.3	49.4
	COMP	**98.9**	**75.6**	**82.1**	**70.6**
Bi-Diag	SINGLE	56.3	51.6	49.3	50.2
	COMP	**98.5**	**78.2**	**84.5**	**72.8**
TransE	SINGLE	92.6	71.7	85.3	72.4
	COMP	**99.0**	**87.4**	**87.5**	**76.3**

Table 3: **Deduction and induction.** We compare mean quantile performance of single-edge training (SINGLE) vs compositional training (COMP). Length 1 queries are excluded.

Pretrained word vectors

- Using pretrained word vectors can improve performance

Accuracy	WordNet		Freebase	
	EV	WV	EV	WV
NTN	70.6	86.2	87.2	**90.0**
Bilinear COMP	77.6	**87.6**	86.1	89.4
TransE COMP	**80.3**	84.9	**87.6**	89.6

Table 5: Model performance in terms of accuracy. EV: entity vectors are separate (initialized randomly); WV: entity vectors are average of word vectors (initialized with pretrained word vectors).

Composition improves performance

- Although a perfect model trained on single steps should work on multiple steps, it doesn't
- Cascading errors cause problems but composition helps

Cascading Errors

- Models suffer from cascading errors
- Trained on up to 5 steps

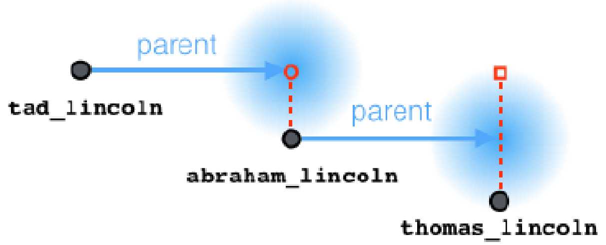


Figure 2: **Cascading errors visualized for TransE.** Each node represents the position of an entity in vector space. The relation `parent` is ideally a simple horizontal translation, but each traversal introduces noise. The red circle is where we expect Tad's parent to be. The red square is where we expect Tad's grandparent to be. Dotted red lines show that error grows larger as we traverse farther away from Tad. Compositional training pulls the entity vectors closer to the ideal arrangement.

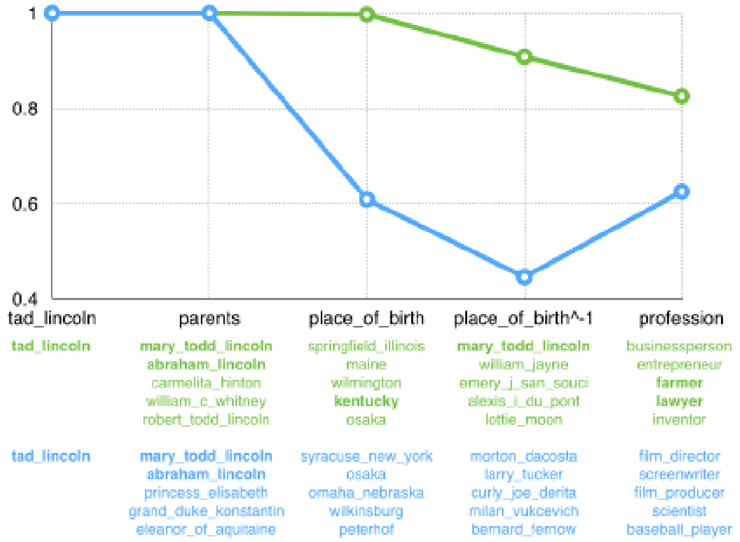


Figure 3: Reconstruction quality (RQ) at each step of the query `tad_lincoln/parents/place_of_birth/place_of_birth`$^{-1}$`/profession`. COMP experiences significantly less degradation in RQ as path length increases. Correspondingly, the set of 5 highest scoring entities computed at each step using COMP (green) is significantly more accurate than the set given by SINGLE (blue). Correct entities are bolded.

Questions/Discussion