LENA: Locality-Expanded Neural Embedding for Knowledge Base Completion

Fanshuang Kong*, Richong Zhang*, Yongyi Mao⋆, Ting Deng*

*Beihang University, China
⋆University of Ottawa, Canada

Jan 12^{nd}, 2019

Knowledge Bases

- Knowledge base(KB) : a collection of factual data
  - "Ontario is a province of Canada"
  - "Toronto is the capital of Ontario"
  - "Jim Carrey's birth place is Newmarket (Ontario)."
  - "Jim Carrey played Stanley Ipkiss in movie *The Mask*"
  - "The genre of movie *The Mask* is comedy"
  - ...
- Data are "cross-linked" in a KB
- Examples:
  - DBpedia(2007-present). From Wikipedia
  - YAGO(2008-present). From Wikipedia/Wordnet/Geonames
  - Freebase(2007-2016). From Wikipedia/NNDB/MusicBrainz/Fashion Model Directory/...

- DBpedia
  - 2007 - present
  - by Leipzig Univ./Univ. of Mannheim/OpenLink Software
  - from Wikipedia
- YAGO
  - 2008 - present
  - by Max-Planck Institute for Computer Science
  - from Wikipedia/Wordnet/Geonames
- Freebase
  - 2007 - present
  - by Metaweb $\rightarrow$ Google
  - from Wikipedia/NNDB/MusicBrainz/Fashion Model Directory/...

Knowledge Bases

- KBs are accumulating enormous amount of knowledge
  - Freebase contains 3 billion records involving 50 million entities
- KBs facilitate new applications
  - Information retrieval
  - Knowledge mining
- Rich research problems in KB
  - Construction of KB
  - Quality improvements
  - Question answering
  - Knowledge discovery
  - ...
- KB Embedding: a generic methodology for all those

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

Prior Art: Embedding of Binary Relations

- Models
  - TransE
  - TransH
  - TransR
  - PTransE
  - Unstructured Model
  - Neural Tensor Network Model ...

Assumption

Relations are sufficient to aggregate all structural information.

- This assumption may not hold sufficiently well!
  - Entities have varying local or global connectivity statistics.
  - Relations involve varying number of factual triples?

Example of Neighbourhood Information

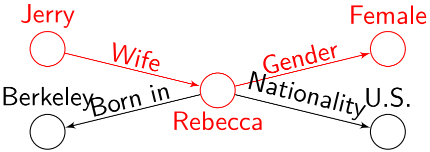

Figure: A subgraph of Rebecca.

- "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

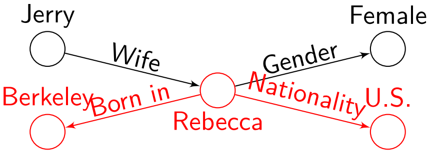

Figure: A subgraph of Rebecca.

- "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

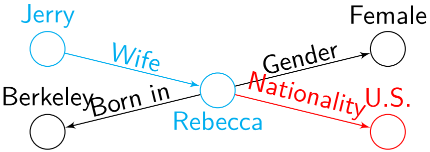

Figure: A subgraph of Rebecca.

- “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.

Roadmap of This Talk

1. Locality-Expanded Neural Embedding
2. Experiments
3. Concluding Remarks

Roadmap of This Talk

1. Locality-Expanded Neural Embedding
2. Experiments
3. Concluding Remarks

Model

Probabilistic Model

$$p(t|h,r) = \frac{\exp(s(h,r,t))}{\sum_{t' \in \mathcal{N}} \exp(s(h,r,t'))}. \quad (1)$$

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

$$\mathbf{x} := D_{\mathrm{E}} x \quad (2)$$

$$\mathbf{r} := D_{\mathrm{R}} r \quad (3)$$

Score Function

$$\begin{aligned} s(h,r,t) := & \langle v^{\mathrm{E}}(h,r,t) + \mathbf{r} + b_{\mathrm{E}}, C_{\mathrm{E}}\mathbf{t} \rangle \\ & + \langle v^{\mathrm{R}}(h,r,t) + \mathbf{h} + b_{\mathrm{R}}, C_{\mathrm{R}}\mathbf{t} \rangle \end{aligned} \quad (4)$$

Neighbourhood Graph

Neighbourhood

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

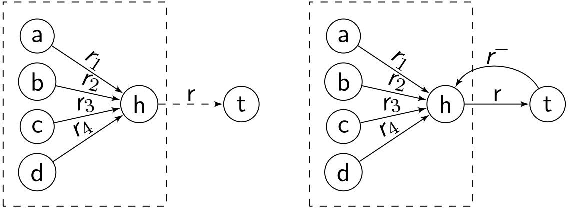

Figure: Example of neighbourhood graphs $\mathcal{G}(h, r, t)$ (the subgraphs in the dashed boxes) of triple (h, r, t). Triples in $\mathcal{G}$ are represented by a solid edge, and triples (e.g., candidate triples) not in $\mathcal{G}$ are represented by a dashed edge.

Window Attention

Attention Weights

$$\alpha_{\Gamma}^{\mathrm{E}}(l) := \frac{\exp\langle \gamma_r^{\mathrm{E}}, \mathbf{h}(l;\Gamma)\rangle}{\sum_{j=0}^{L} \exp\langle \gamma_r^{\mathrm{E}}, \mathbf{h}(j;\Gamma)\rangle} \tag{5}$$

$$\alpha_{\Gamma}^{\mathrm{R}}(l) := \frac{\exp\langle \gamma_r^{\mathrm{R}}, \mathbf{r}(l;\Gamma)\rangle}{\sum_{j=0}^{L} \exp\langle \gamma_r^{\mathrm{R}}, \mathbf{r}(j;\Gamma)\rangle} \tag{6}$$

- Both attention parameters γ_r^{E} and γ_r^{R} are dependent of the r

Soft-selection of entities and relations

$$\begin{aligned} v^{\mathrm{E}}(\Gamma) &:= \alpha_{\Gamma}^{\mathrm{E}}(0)\mathbf{h} + \sum_{l=1}^{L} \alpha_{\Gamma}^{\mathrm{E}}(l)\mathbf{h}(l;\Gamma) \\ v^{\mathrm{R}}(\Gamma) &:= \alpha_{\Gamma}^{\mathrm{R}}(0)\mathbf{r} + \sum \alpha_{\Gamma}^{\mathrm{R}}(l)\mathbf{r}(l;\Gamma). \end{aligned} \tag{7}$$

Cross Window Pooling

Apply a pooling operation on $v^{\mathrm{E}}(\Gamma)$'s and $v^{\mathrm{R}}(\Gamma)$'s across all windows.

$$v^{\mathrm{E}}(h,r,t) := \mathbf{max\_pooling}\left\{v^{\mathrm{E}}(\Gamma) : \Gamma \in \widetilde{\mathcal{H}}_L(h,r,t)\right\}; \tag{8}$$

$$v^{\mathrm{R}}(h,r,t) := \mathbf{max\_pooling}\left\{v^{\mathrm{R}}(\Gamma) : \Gamma \in \widetilde{\mathcal{H}}_L(h,r,t)\right\}. \tag{9}$$

Objective Function

$$\Theta^* := \arg\min_{\Theta} \sum_{(h,r)\in\mathcal{K}} \sum_{t\in\mathcal{T}(h,r)} \left(-\frac{1}{|\mathcal{T}(h,r)|}\log p(t|h,r)\right). \tag{10}$$

Roadmap of This Talk

Dataset Statistics and Experimental Settings

- Datasets: FB15K, WN18, FB15K-237 and WN18-RR
- In FB15K and WN18 dataset, many testing triples are reciprocal to training triples

Table: The statistics of datasets used in this study.

Datasets	entities	relations	triples(train/test/valid)
FB15K	14,951	1,345	483,142 / 59,071 / 50,000
WN18	40,943	18	141,442 / 5,000 / 5,000
FB15K-237	14,541	237	272,115 / 20,466 / 17,535
WN18-RR	40,943	11	86,835 / 3,134 / 3,034

Experiment Protocols

- Compute the loss of each triple (h, r, x) under the model where x ranges over all entities in the KB
- Rank these losses from low to high
- Obtain the rank for $x = t$ as the rank for testing case (h, r, t)
- Use the standard metrics Mean rank (MR), top-10 hit (HIT@10, or simply HIT), reciprocal rank (MRR) and their corresponding filtered version metrics, FMR, FHIT and FMRR, to evaluate the model performances.

Table: The hyper-parameters of LENA.

	FB15K	FB15K-237	WN18	WN18-RR
L	3	3	5	3
H	90	90	90	60

Results: Link Prediction Performance

Models	FB15K-237					
	MR	F-MR	MRR	F-MRR	HIT	F-HIT
TransE	367	194	12.1	20.8	28.4	42.0
TransH	357	186	12.5	21.5	29.3	43.3
DistMult	453	255(254)	14.0	22.7(24.1)	27.6	40.7(41.9)
ComplEx	456	245(339)	12.8	22.5(24.7)	26.4	41.2(42.8)
Analogy	468	274	14.3	23.3	27.4	40.2
ProjE	360	193	16.0	29.8	29.3	47.7
ConvE	483	269(246)	15.3	31.1(31.6)	28.4	48.1(49.1)
R-GCN+	-	-	[15.6]	[24.9]	-	[41.7]
LENA δ=0.1	**328**	**174**	**17.5**	31.0	**32.5**	**49.9**
LENA δ=0.25	**345**	**170**	**16.8**	**31.8**	**31.6**	**50.4**
LENA δ=0.5	364	**175**	**16.3**	**32.0**	**30.8**	**50.4**
Models	**WN18-RR**					
	MR	F-MR	MRR	F-MRR	HIT	F-HIT
TransE	3542	3529	10.8	12.4	32.9	35.3
TransH	3894	3881	11.0	12.7	33.2	35.2
DistMult	7753	7643(5110)	28.1	39.1(43.0)	40.4	41.9(49.0)
ComplEx	8303	8299(5261)	28.1	39.0(44.0)	40.1	41.3(51.0)
Analogy	8221	8075	27.6	38.9	39.5	41.0
ProjE	3732	3718	27.8	38.2	46.9	50.0
ConvE	4810	4795(5277)	31.1	42.5(46.0)	47.1	49.8(48.0)
LENA δ=0.1	**3028**	**3014**	28.7	35.7	**48.6**	**51.1**
LENA δ=0.25	**3276**	**3262**	30.2	41.5	**48.3**	**51.5**
LENA δ=0.5	**3300**	**3285**	28.3	42.5	**48.5**	**51.4**

Results: Link Prediction Performance

Models	FB15K					
	MR	F-MR	MRR	F-MRR	HIT	F-HIT
TransE	194	54	16.6	31.6	48.4	73.9
TransH	193	54	16.7	31.9	48.5	74.0
DistMult	282	113(97)	24.7$\vert$24.2$\vert$	70.8(65.4)	48.9	83.0(82.4)
ComplEx	278	119	25.4$\vert$24.2$\vert$	71.6(69.2)	49.9	83.5(84.0)
Analogy	273	114	25.5$\langle 25.3 \rangle$	72.3(72.5)	50.1	83.9(85.4)
ProjE	164	53	29.0	62.0	53.8	80.0
ConvE	189	48(64)	27.3	69.0(74.5)	52.4	85.4(87.3)
Gaifman	-	{75}	-	-	-	{84.2}
R-GCN+	-	-	[26.2]	[69.6]	-	[84.2]
LENA δ=0.1	**153**	50	**30.7**	59.5	**55.9**	79.6
LENA δ=0.25	**154**	**42**	**29.7**	63.7	**54.7**	81.9
LENA δ=0.5	**161**	**39**	28.6	65.8	53.4	83.1

Models	WN18					
	MR	F-MR	MRR	F-MRR	HIT	F-HIT
TransE	320	307	28.7	39.3	77.5	92.3
TransH	327	314	29.0	39.4	77.8	92.6
DistMult	654	642(902)	52.7$\vert$53.2$\vert$	73.9(82.2)	77.6	93.6(93.6)
ComplEx	737	735	64.5$\vert$58.7$\vert$	94.2(94.1)	82.2	94.5(94.7)
Analogy	725	717	65.6$\langle 65.7 \rangle$	94.2(94.2)	83.3	94.6(94.7)
ProjE	281	266	58.1	82.6	81.5	95.2
ConvE	434	417(504)	53.3	94.4(94.2)	79.6	95.5(95.5)
Gaifman	-	{352}	-	-	-	{93.9}
R-GCN+	-	-	[56.1]	[81.9]	-	[96.4]
LENA δ=0.1	**254**	**242**	**66.4**	89.8	**84.2**	95.6
LENA δ=0.25	**276**	**261**	65.1	92.7	82.4	95.6
LENA δ=0.5	312	296	62.2	93.8	81.4	95.5

Behaviour of Attention

Rank Promotion

$$\text{rp}(h, r, t) := \text{rank}^{\text{ProjE}}(h, r, t) - \text{rank}^{\text{LENA}}(h, r, t),$$

where $\text{rank}^{\text{ProjE}}(h, r, t)$ and $\text{rank}^{\text{LENA}}(h, r, t)$ are the rank values of (h, r, t) given by ProjE and LENA.

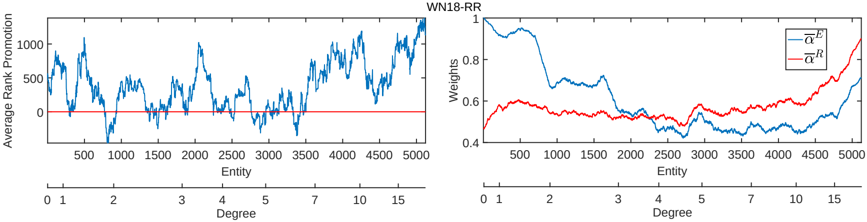

Figure: α^{E} and α^{R} vs the degree of entities.

Behaviour of Attention

Table: Examples of identified informative neighbors.

Testing triple	Informative Neighbors	α^E	α^R
Marriott International, Liabilities_Currency, U.S. Dollar	Marriott International, Region, Maryland	0.996	0.501
James Arness, Place_Lived, Minneapolis	James Arness, People_Born_Here, Minneapolis	0.9797	0.0001
Bob Dylan, Instruments_Played, Bass Guitar	Bob Dylan, Instrumentalists, Guitar	0.977	1.59e-06
Hepatitis, Symptom_of, Jaundice	Hepatitis, Risk_Factor, Alcoholism	1.532e-06	0.999

Roadmap of This Talk

Concluding Remarks

- The embeddings of a triple (h, r, t) may be insufficient for predicting its factual existence.
- Extracting and combining information from larger graph neighbourhoods can therefore improve link-prediction performance.
- We show that attention mechanisms are an effective means of achieving such information extraction and combining.
- LENA has broken a number of performance records, over a range of datasets.

Thank you!

https://github.com/fskong/LENA