WWW 2019

ActiveLink: Deep Active Learning for Link Prediction in Knowledge Graphs

Natalia Ostapuk, Jie Yang, Philippe Cudré-Mauroux

University of Fribourg
Fribourg, Switzerland

OUTLINE

1. Knowledge Graph Introduction
2. Query strategy
3. Model updating
4. Experiment

Introduction

Knowledge Representation

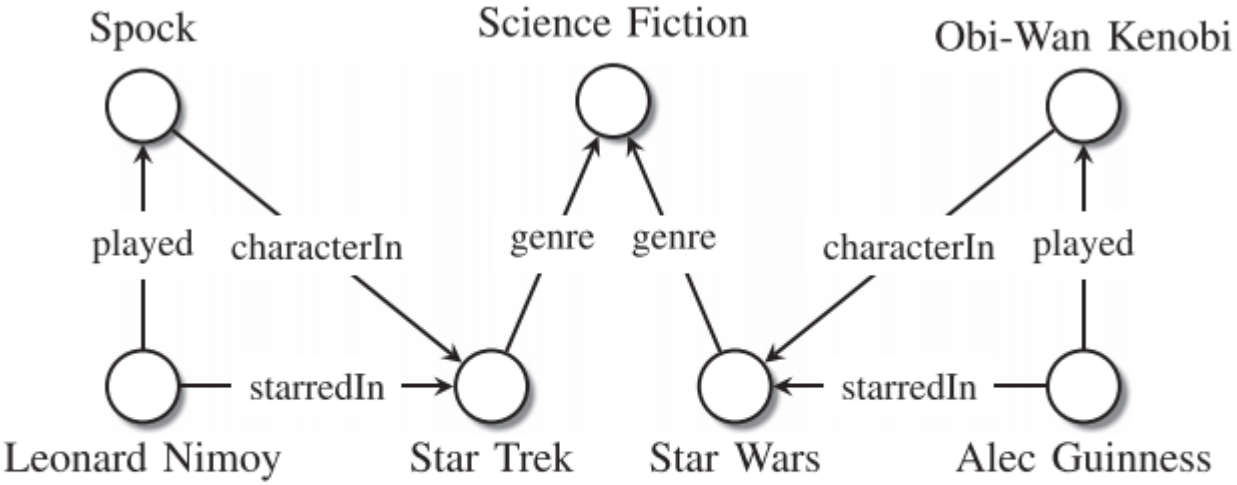


Fig. 1. ***Sample knowledge graph. Nodes represent entities, edge labels represent types of relations, and edges represent existing relationships.***

Introduction

Knowledge Representation

Header entity | relation | Tail entity

subject	*predicate*	*object*
(LeonardNimoy,	*profession,*	*Actor)*
(LeonardNimoy,	*starredIn,*	*StarTrek)*
(LeonardNimoy,	*played,*	*Spock)*
(Spock,	*characterIn,*	*StarTrek)*
(StarTrek,	*genre,*	*ScienceFiction)*

(s, p , o) or
(h, r, t) or
(e1, r, e2)

TKDE17- Knowledge Graph Embedding-A Survey of Approaches and Applications

Proc. IEEE15 A Review of Relational Machine Learning for Knowledge Graphs

Introduction

Tasks in KG

- KG Embedding
- Link Prediction
- Triple Classification
- Entity Classification
- Entity Resolution
- Relation Extraction
- Question Answering
- Recommender Systems

In-KG Applications (KG refinement)

Out-of-KG Applications

Introduction

KG Embedding

A typical KG embedding technique generally consists of three steps:

- representing entities and relations
- defining a scoring function
- learning entity and relation representations

$$\min_{\Theta} \sum_{\tau \in \mathbb{D}^+ \cup \mathbb{D}^-} \log\left(1 + \exp(-y_{hrt} \cdot f_r(h,t))\right)$$

$$\min_{\Theta} \sum_{\tau^+ \in \mathbb{D}^+} \sum_{\tau^- \in \mathbb{D}^-} \max(0, \gamma - f_r(h,t) + f_{r'}(h',t'))$$

Translational Distance Models

distance-based scoring functions

TransE: $f_r(h,t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{1/2}$

Semantic Matching Models

similarity-based scoring functions

RESCAL : $f_r(h,t) = \mathbf{h}^\top \mathbf{M}_r \mathbf{t}$

Introduction

Tasks in KG

In-KG Applications (KG refinement)

- Link Prediction: predicting h given (r, t) or t given (h, r), with the former denoted as (?, r, t) and the latter as (h, r, ?).
- Triple Classification: verifying whether an unseen triple fact (h, r, t) is true or not.
- Entity Classification: aims to categorize entities into different semantic categories, e.g., AlfredHitchcock is a Person.
- Entity Resolution: verifying whether two entities refer to the same object.

Introduction

Tasks in KG

Out-of-KG Applications

- Relation Extraction — aims to extract relational facts from plain text where entities have already been detected.
- Question Answering — Q: Who directed Psycho? A: AlfredHitchcock (AlfredHitchcock, DirectorOf, Psycho)
- Recommender Systems — leverages heterogeneous information in a KG to improve the quality of recommender system

WWW 2019

ActiveLink: Deep Active Learning for Link Prediction in Knowledge Graphs

Natalia Ostapuk, Jie Yang, Philippe Cudré-Mauroux

University of Fribourg
Fribourg, Switzerland

Problem Statement

Knowledge Graph $\mathcal{G} = \{(s, r, o)\} \subset \mathcal{E}, \mathcal{R}, \mathcal{E}$

where $\mathcal{E}$ is the entity set, $\mathcal{R}$ is the relation set, and where a triple (s, r, o) represents a relationship $r \in \mathcal{R}$ between a subject $s \in \mathcal{E}$ and an object $o \in \mathcal{E}$

Link Predictor $p(y|(s, r, o), \Theta) = \text{softmax}\,(f_r^{\Theta}(s, o))$

Active Learning $(s, r, o)^* = \underset{(s,r,o)\in\mathcal{G}}{\arg\max}\ \phi((s, r, o))$

Uncertainty Sampling

Problem:

Standard deep learning tools for regression and classification do not capture model uncertainty. A model can be uncertain in its predictions even with a high softmax output.

Solution:

Performing dropout during the forward pass when making predictions.

"equivalent to the prediction when the parameters are sampled from a variational distribution of the true posterior"

ICML'16 **Gal** and **Ghahramani,** Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Uncertainty Sampling

Entropy based uncertainty

$$\phi((s, r, o)) = H[y|(s, r, o), \mathcal{D}_{train}]$$
$$= -\sum_{C\in\{0,1\}} p(y = C|(s, r, o), \mathcal{D}_{train}) \log p(y = C|(s, r, o), \mathcal{D}_{train})$$
$$= -\sum_{C\in\{0,1\}} (\frac{1}{T}\sum_{t} \hat{p}_C^t) \log(\frac{1}{T}\sum_{t} \hat{p}_C^t) \quad (8)$$

T次预测
的均值

Representativeness

Clustering based

1. K-Means for clustering entities, C clusters
2. **foreach** C*j* ∈ C **do**
3. Compute *redundancy_score*(C*j*);
4. C*k* ← Pick *k* clusters from C based on the scores
5. **foreach** Cj ∈ Ck **do**
6. (s,r,o) ← Pick a triple from Cj based on uncertainty;
7. Add (s,r,o) to training set;

the redundancy score (row 3) is calculated as the averaged similarity between all data samples — specifcally, cosine similarity between entity embeddings — in the cluster and those in the existing training pool.

Meta-Incremental Training

Problem:

Fine-tune the model using the new data samples only is likely to bias the model to the small amount of newly selected data

Solution:

makes use of the triples selected in previous iterations within a certain time window w.

i.e. $\{\mathcal{T}_{i-w}, \mathcal{T}_{i-w+1}, \ldots, \mathcal{T}_{i-1}\}$

$$\Theta_i = \Theta_{i-1} - \beta \Delta_\Theta \sum_{l=i-w}^{i} \mathcal{L}(f_{\Theta'_l}, \mathcal{T}_l)$$

$$\min_{\Theta} \sum_{l=i-w}^{i} \mathcal{L}(f_{\Theta'_l}, \mathcal{T}_l) = \sum_{l=i-w}^{i} \mathcal{L}(f_{\Theta_{i-1} - \alpha \Delta_\Theta \mathcal{L}(f_{\Theta_{i-1}}, \mathcal{T}_l)}, \mathcal{T}_l)$$

Δ_Θ is the gradient of Θ with respect to the loss.

Experiment

Dataset:

- FB15K-237 [32] is a subset of Freebase
- WikiMovie is a subset of Wikidata.

Table 1: Descriptive statistics of the datasets.

	#Entities	#Relationships	#Triples
FB15K-237	14,541	474	310,116
WikiMovie-300K	36,001	588	286,683
WikiMovie-1M	104,500	788	987,896

Compared methods:

- **Random,** select data randomly.
- **Structured**, a variant of our method that randomly selects triples from each cluster.
- **Uncertainty**, a variant of our method that selects data samples based on entropy.
- **Structured-Uncertainty**, the proposed method.

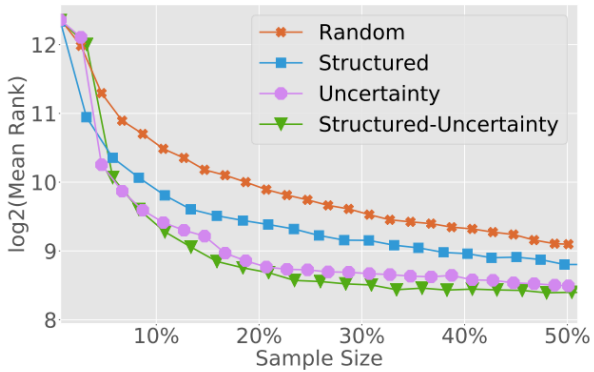
Model training methods:

- Retrain, Incremental, Meta-Incremental

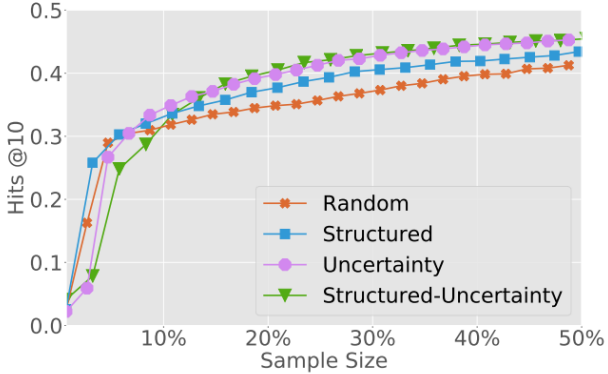
Models: ConvE, MLP

Experiment

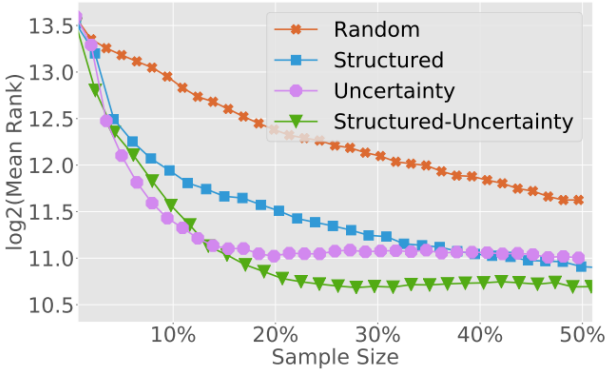
AL



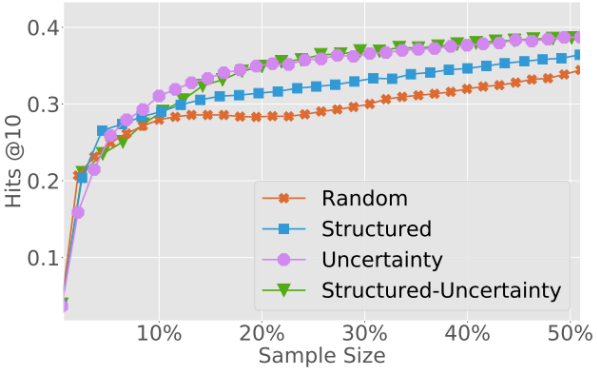
(a) FB15K-237 - Mean Rank



(b) FB15K-237 - Hits@10



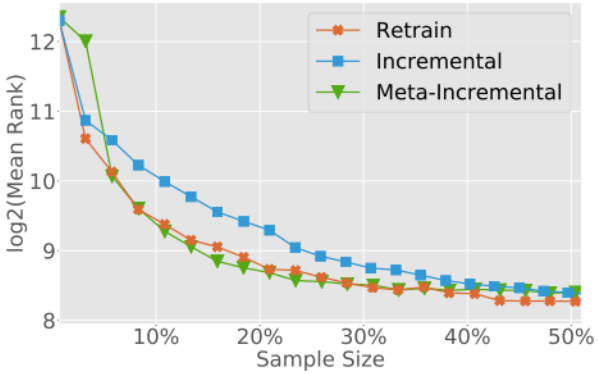
(c) WikiMovie-300K - Mean Rank



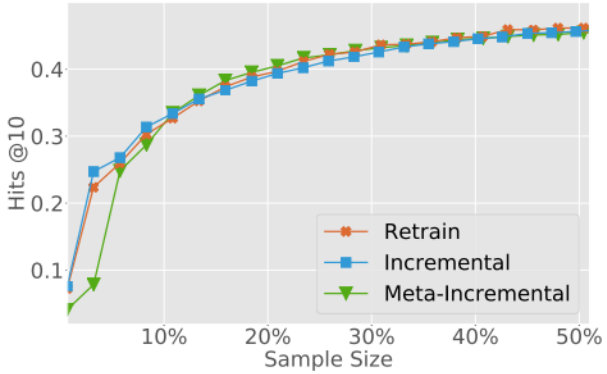
(d) WikiMovie-300K - Hits@10

Experiment

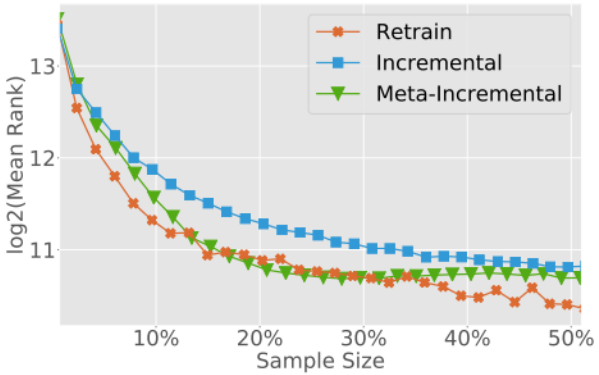
Model training methods



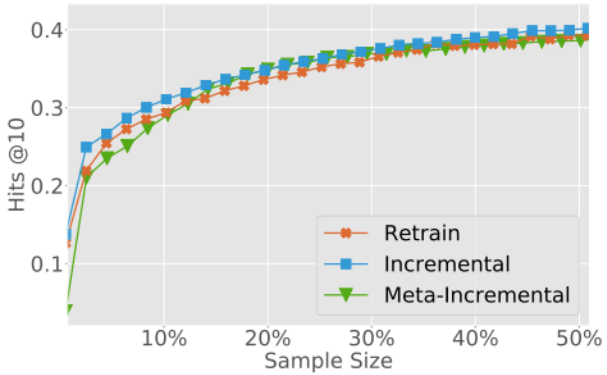
(a) FB15K-237 - Mean Rank



(b) FB15K-237 - Hits@10



(c) WikiMovie-300K - Mean Rank



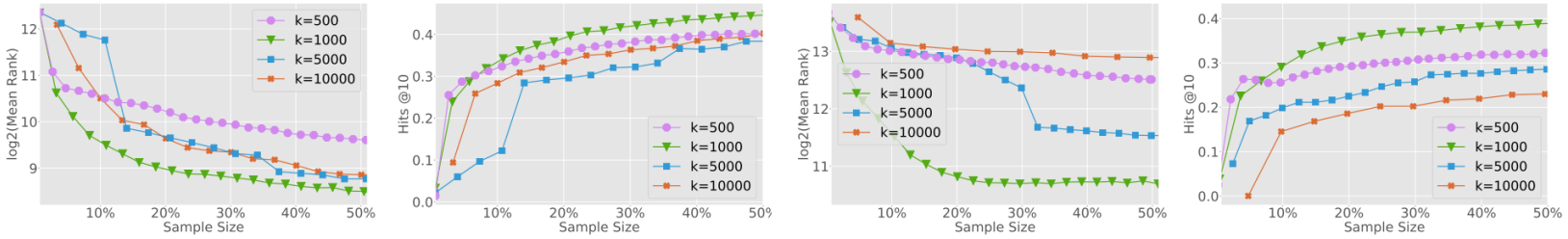
(d) WikiMovie-300K - Hits@10

Experiment

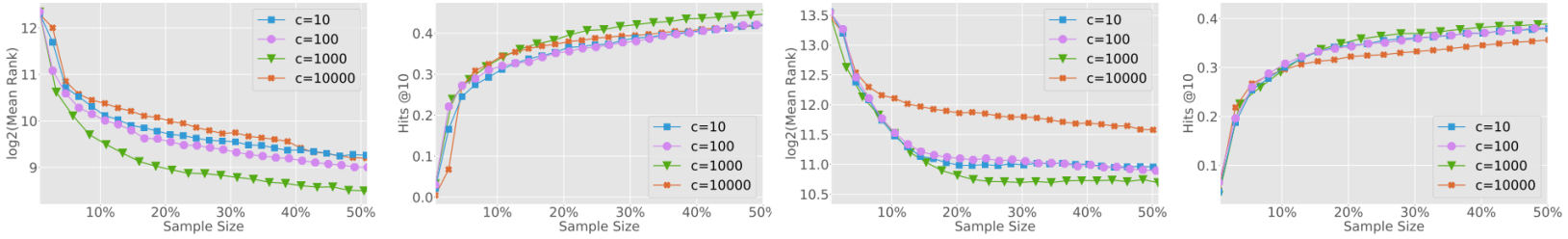
Table 3: Performance of ConvE and MLP trained by ActiveLink and the non-active learning setting on a varying fraction of the FB15K-237 and WikiMovie-300K datasets, measured by both Mean Rank and Hits@10.

Predictor	Fraction	FB15K-237				WikiMovie-300K			
		Mean Rank		Hits@10		Mean Rank		Hits@10	
		Non-Act.	ActiveLink	Non-Act.	ActiveLink	Non-Act.	ActiveLink	Non-Act.	ActiveLink
ConvE	10%	1325.26	409.14	0.255	0.403	6857.77	2150.08	0.219	0.329
	20%	822.88	351.98	0.282	0.440	5253.15	1641.98	0.259	0.365
	30%	605.79	339.48	0.301	0.450	4317.67	1673.97	0.271	0.379
	40%	537.12	329.21	0.326	0.459	3406.10	1658.99	0.296	0.388
	50%	458.91	318.18	0.349	0.464	2832.33	1628.06	0.314	0.396
MLP	10%	1386.80	487.64	0.248	0.395	7094.62	2520.71	0.216	0.300
	20%	848.21	374.75	0.283	0.440	5613.55	1536.02	0.248	0.357
	30%	663.18	346.94	0.314	0.457	4463.22	1379.29	0.274	0.376
	40%	547.16	326.11	0.332	0.467	3370.01	1326.25	0.298	0.389
	50%	458.51	322.40	0.354	0.470	2680.69	1372.14	0.325	0.398

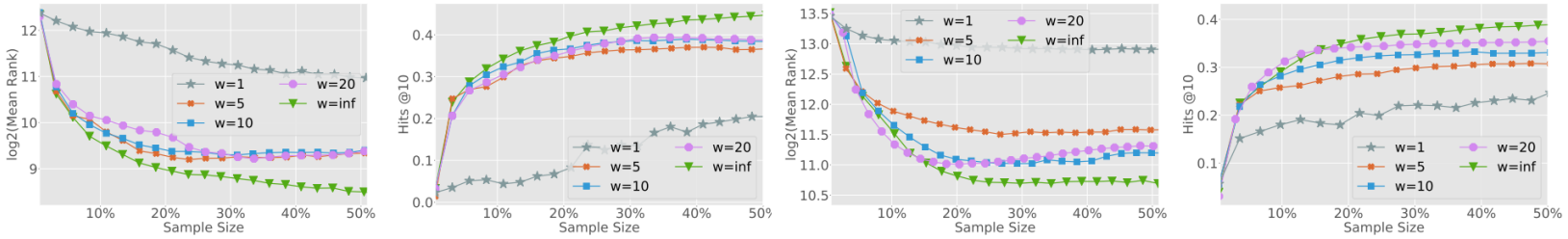
参数敏感性



(a) k - Mean Rank (FB15K-237) (b) k - Hits@10 (FB15K-237) (c) k - Mean Rank (WikiMovie-300K) (d) k - Hits@10 (WikiMovie-300K)



(e) c - Mean Rank (FB15K-237) (f) c - Hits@10 (FB15K-237) (g) c - Mean Rank (WikiMovie-300K) (h) c - Hits@10 (WikiMovie-300K)



(i) w - Mean Rank (FB15K-237) (j) w - Hits@10 (FB15K-237) (k) w - Mean Rank (WikiMovie-300K) (l) w - Hits@10 (WikiMovie-300K)

ure 3: Impact of (a) #samples per iteration k, (b) #clusters c and (c) window size w on the performance of ActiveLink on both [5K-237 and WikiMovie-300K datasets, measured by Mean Rank (shown in semi-log scale) and Hits@10.