



DRUM:End-To-End Differentiable Rule Mining On Knowledge Graphs

2024-08-20

presenter : Sooho Moon

DMAIS

Ali Sadeghian(*1), Mohammadreza Armandpour(*2), Patrick Ding(2), Daisy Zhe Wang(1),

1-Department of Computer Science, University of Florida

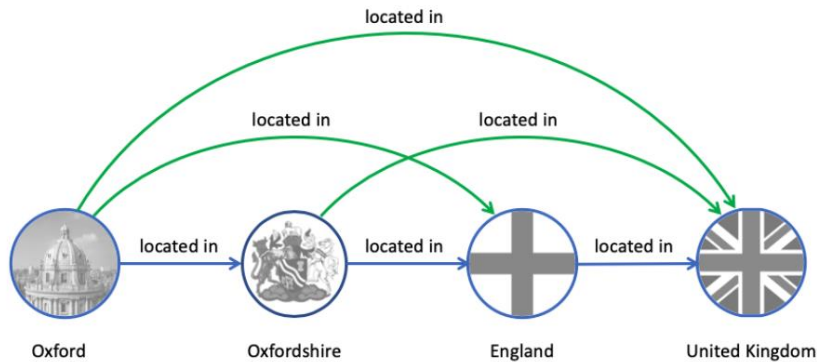
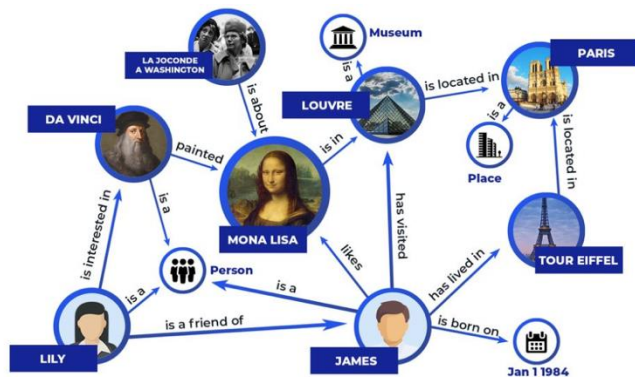
2-Department of Statistics, Texas A&M University

NeurIPS 2019

- Main interest
- Problem with previous works
- Architecture
- Experiments
- Conclusion

■ Knowledge Graph(KG) reasoning

- **Knowledge graph** : 그래프의 특수한 형태로 entity(노드)가 relation(간선)으로 연결되어 있는 구조인 데이터(Knowledge Base, KB)
- **KG reasoning** : Entity와 relation의 특성들을 학습해 데이터에 대한 추론을 목적으로 함



Problem with previous works

▪ Most prominent directions for KB reasoning

1. Representation learning(e.g., TransE(2013), ComplEx(2016))

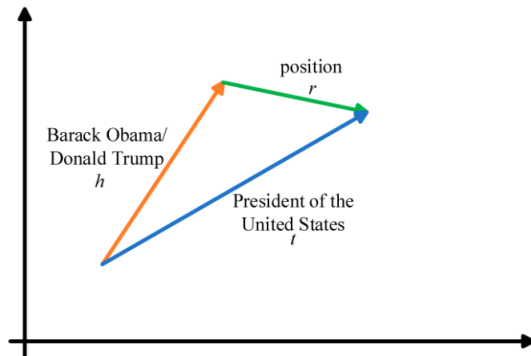
entity와 relation을 latent space로 임베딩해 관계를 학습

장점

- Triple 하나씩 model에 입력되므로 학습 과정이 단순함
- 연속 공간인 latent space에서 학습하므로 gradient base 알고리즘들 사용 가능

단점

- Transductive learning
- 결과들이 해석 가능하지 않음(e.g., 이 벡터는 왜 여기로 임베딩 되었는가?)



Problem with previous works

▪ Most prominent directions for KB reasoning

2. Rule mining(e.g., TensorLog(2016), NeuralLP(2017))

relation의 결합으로 구성된 논리적 규칙들의 패턴을 학습

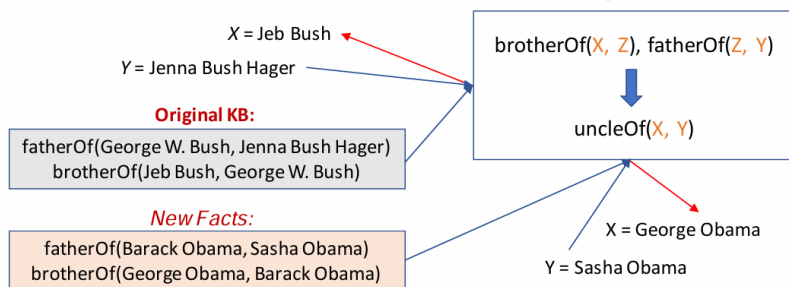
장점

- Inductive learning
- 해석 가능한 추론을 도출함

단점

- Rule의 structure(이산적 문제), parameter(연속적 문제)를 동시에 알아내야 함

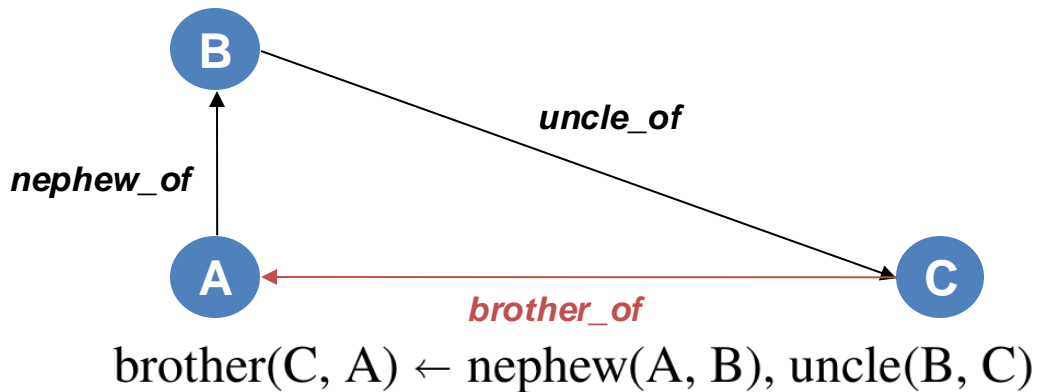
Question: Which person X is uncle of Y ?



Problem with previous works

■ Why rule mining?

- Transductive, inductive setting 모두에 적용 가능
(데이터가 적거나 노이즈가 많은 KB에서도 모델 사용 가능)
- 추론들이 해석 가능하므로 모델과 인간의 관계를 긴밀하게 해줌
(e.g., 디버깅 용이, 결과 신뢰도 향상)



Problem with previous works

■ Previous approaches for training models on rule mining

- Predefined statistical measures 사용해 rule을 평가

support, confidence 등의 휴리스틱한 metric들은 평가 이상의 의미가 없음



- Inductive Logic Programming(ILP) system 접근(e.g., FOIL(1990), MDIE(1995))

다양한 특성 학습이 가능한 시스템이지만 negative example의 필요, scaling issue의 문제가 있음

$$\text{schema of ILP} \left\{ \begin{array}{l} \forall e \in E^+ : B \wedge H \models e \\ \forall e \in E^- : B \wedge H \not\models e \end{array} \right.$$

Problem with previous works



- Ontological Pathfinding(OP) 접근(e.g., AMIE+(2015))

support, confidence metric을 기반으로 parallelization, partitioning 기술들을 사용해 속도 향상

However, 여전히 metric과 이산적 path counting 고유의 한계를 극복하지 못함



- End-to-end differentiable model(e.g., Neural LP(2017))

최초의 neural network + LSTM 기반 rule mining 모델

structure(이산), scoring(연속) 문제를 동시에 다루어 의의가 큼



DRUM(2019)

■ Background

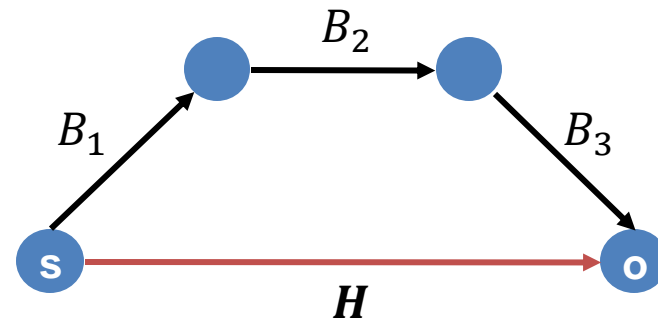
- Definitions

$\text{KB} : G = \{(s, r, o) | s, o \in \mathcal{E}, r \in \mathcal{R}\}$

First order logical rule : $\mathbf{B} \Rightarrow H, \mathbf{B} = \bigwedge_i B_i(\cdot, \cdot)$

H : head predicate

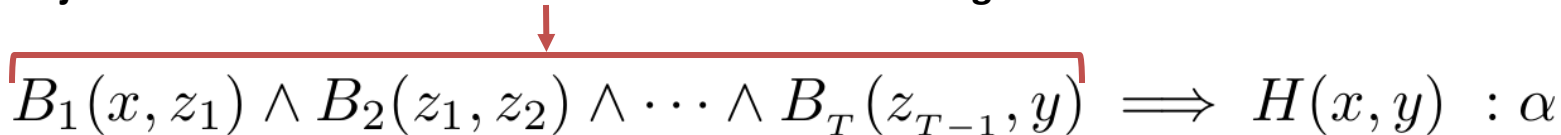
B_i : body atom (e.g., *livesIn*(\cdot, \cdot))



■ Background

- Rule Mining

Objective : Learn “closed and connected” first-order logical Horn clauses from KB


$$B_1(x, z_1) \wedge B_2(z_1, z_2) \wedge \cdots \wedge B_T(z_{T-1}, y) \implies H(x, y) : \alpha$$

— T : length of rule

— z_i : entity variable

— α : confidence value of H

특정 rule의 B_i 찾기 위해 이산 공간 탐색 + α 학습 위해 연속 공간 탐색

- Differentiable formulation

n 이 entity의 개수일 때 각 entity를 one-hot vector로 표현 : $\{v_1, \dots, v_n\}$

$$v_2 \\ = [0, 1, 0, \dots, 0]^T$$

특정 relation B_r 에 대한 entity 인접 행렬 : A_{B_r}

$$A_{B_r} = \begin{pmatrix} 0 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \rightarrow B_r(v_1, v_n)$$

- Differentiable formulation

$$B_1(x, z_1) \wedge B_2(z_1, z_2) \wedge \cdots \wedge B_T(z_{T-1}, y)$$



$$\mathbf{v}_x^T \cdot \mathbf{A}_{B_1} \cdot \mathbf{A}_{B_2} \cdots \mathbf{A}_{B_T} \cdot \mathbf{v}_y$$

위 수식의 결과(스칼라)는 x 에서 y 로 B_{r_i} 를 타고 이어질 수 있는 경로의 개수와 일치

■ Objective function

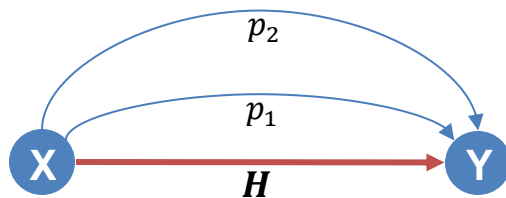
- 각 head relation H 에 대해 다음 식을 maximize 하는 적절한 α 를 찾는 것

$$\omega_H(\alpha) \doteq \sum_s \alpha_s \prod_{k \in p_s} \mathbf{A}_{B_k}$$

학습하고자 하는 rule에 대해 indexing 하는 s , s 에 따른 body atom 집합 p_s

$$O_H(\alpha) \doteq \sum_{(x, H, y) \in KG} \mathbf{v}_x^T \omega_H(\alpha) \mathbf{v}_y$$

- 최종적으로 H 에 대해 optimize 하고 싶은 함수



■ Neural LP의 한계

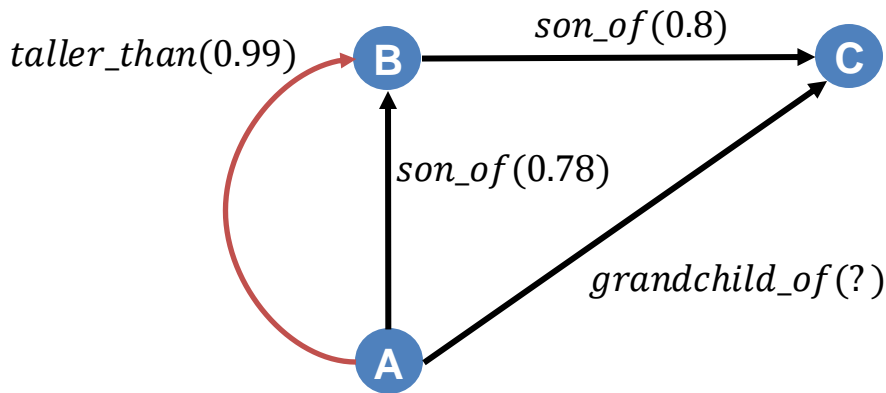
- 기존 수식 $O_H(\alpha)$ 은 over-parameterization 문제에 봉착함($O(|\mathcal{R}|^T)$)
따라서 수식 $w_H(\alpha)$ 를 아래와 같이 수정함

$$\Omega_H^I(\mathbf{a}) \doteq \prod_{i=1}^T \left(\sum_{k=0}^{|\mathcal{R}|} a_{i,k} \mathbf{A}_{B_k} \right) \quad \mathbf{A}_{B_0} = I_n: \text{길이 } T \text{ 이하의 모든 rule들을}$$

학습할 수 있도록 해줌

- 위 수식은 Neural LP의 근간 수식과 거의 일치하고 파라미터 개수는 $T(|\mathcal{R}| + 1)$
- 하지만 DRUM 저자들은 위 수식을 optimize 하는 것은 필연적으로 높은 confidence를 가지는 틀린 rule을 학습함을 증명함

■ Neural LP의 한계 증명(Theorem 1)

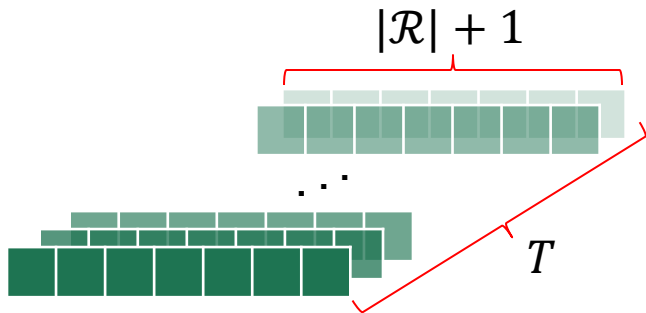


“특정 깊이에서의 무의미한 relation의 confidence가 지나치게 클 수 있음”

- 실험적으로도 위 정리를 따르는 틀린 rule들이 Neural LP를 통해 발견됨 (결과는 experiments에서)

■ DRUM

- 다시 돌아와 길이가 T 이하인 rule들에 대한 confidence 개수는 $(|\mathcal{R}| + 1)^T$
- 이는 각 axis 크기가 $|\mathcal{R}| + 1$ 인 T 차원 텐서로 볼 수 있음
- 즉 $B_{r_1} \wedge B_{r_2} \wedge \cdots \wedge B_{r_T}$ 에 해당하는 confidence는 (r_1, r_2, \dots, r_T) 위치에 저장되는 **confidence value tensor**라고 할 수 있음



■ DRUM

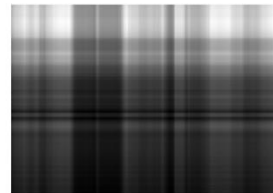
- $\Omega_H^I(\mathbf{a})$ 는 confidence value tensor의 rank one approximation으로 해석할 수 있음
- **Rank L approximation**(rank one 포함)은 L이 클수록 텐서를 더 정확히 근사 가능



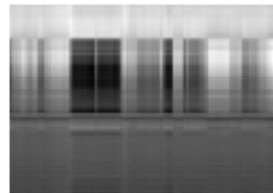
$$\Omega_H^L(\mathbf{a}, L) \doteq \sum_{j=1}^L \left\{ \prod_{i=1}^T \sum_{k=0}^{|\mathcal{R}|} a_{j,i,k} \mathbf{A}_{B_k} \right\}$$



(a) Original image \mathbf{A} .



(b) Rank-1 approximation $\hat{\mathbf{A}}(1)$.



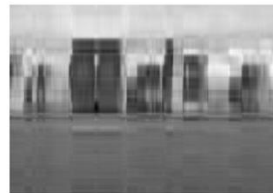
(c) Rank-2 approximation $\hat{\mathbf{A}}(2)$.



(d) Rank-3 approximation $\hat{\mathbf{A}}(3)$.



(e) Rank-4 approximation $\hat{\mathbf{A}}(4)$.



(f) Rank-5 approximation $\hat{\mathbf{A}}(5)$.

■ DRUM이 해결해야할 문제들

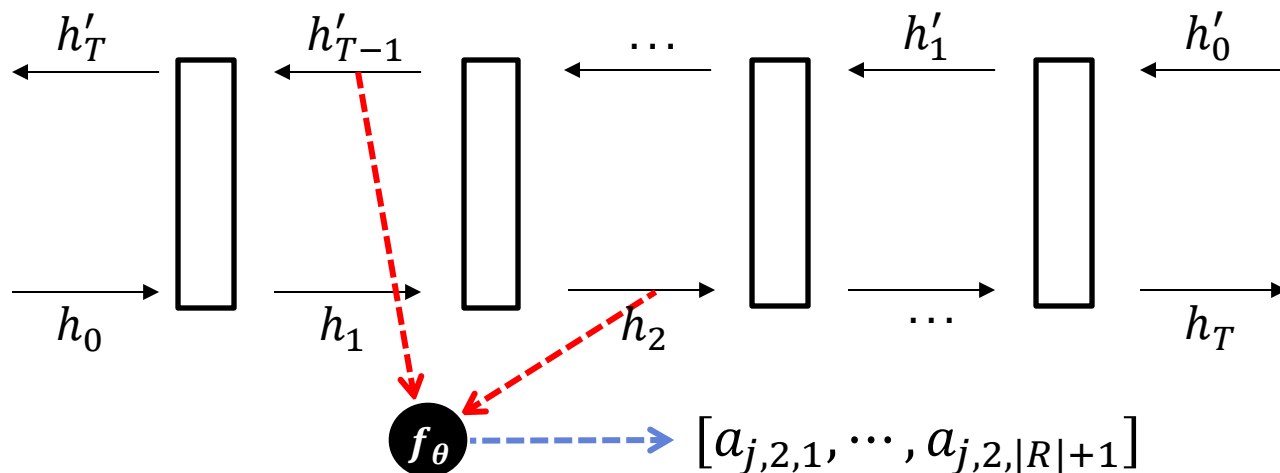
1. 하나의 h 에 대해서는 파라미터 개수가 $LT(|\mathcal{R}| + 1)$ 이지만 전체 개수는 $O(|\mathcal{R}|^2)$ 이므로 여전히 무거움
2. 또한 Ω_H^L 를 직접 optimize 하는 것은 특정 head의 학습 결과를 다른 head까지 영향을 주지 못한다는 한계가 존재
- + 3. 어떤 relation은 같은 rule에 올 수 없음(e.g., $\text{father_of}(X, \dots) + \text{wife_of}(X, \dots)$)



Bidirectional RNN

$$\mathbf{h}_i^{(j)}, \mathbf{h}'_{T-i+1} = \text{BiRNN}_j(\mathbf{e}_H, \mathbf{h}_{i-1}^{(j)}, \mathbf{h}'_{T-i}{}^{(j)}), \text{ + gradient clipping, LSTM}$$

$$[a_{j,i,1}, \dots, a_{j,i,|\mathcal{R}|+1}] = f_\theta([\mathbf{h}_i^{(j)}, \mathbf{h}'_{T-i+1}{}^{(j)}]),$$



■ Statistical Relation Learning

- **UMLS**(biomedical relations), **Kinship**(Australian tribes), **Family**(bloodlines of multiple families)

Table 2: Experiment results with maximum rule length 2 and 3

| | | Family | | | | UMLS | | | | Kinship | | | |
|---------|-----------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| | | Hits@ | | | | Hits@ | | | | Hits@ | | | |
| | | MRR | 10 | 3 | 1 | MRR | 10 | 3 | 1 | MRR | 10 | 3 | 1 |
| $T = 2$ | Neural-LP | .91 | .99 | .96 | .86 | .75 | .92 | .86 | .62 | .62 | .91 | .69 | .48 |
| | DRUM-1 | .92 | 1.0 | .98 | .86 | .80 | .97 | .93 | .66 | .51 | .85 | .59 | .34 |
| | DRUM-4 | .94 | 1.0 | .99 | .89 | .81 | .98 | .94 | .67 | .60 | .92 | .69 | .44 |
| $T = 3$ | Neural-LP | .88 | .99 | .95 | .80 | .72 | .93 | .84 | .58 | .61 | .89 | .68 | .46 |
| | DRUM-1 | .91 | .99 | .96 | .85 | .77 | .96 | .92 | .63 | .57 | .88 | .66 | .43 |
| | DRUM-4 | .95 | .99 | .98 | .91 | .80 | .97 | .92 | .66 | .61 | .91 | .71 | .46 |

: comparison with inductive models

■ Statistical Relation Learning

| Datasets | UMLS | | | | Kinship | | | |
|-----------------------------|------|--------|--------|---------|---------|--------|--------|---------|
| | MRR | Hits@1 | Hits@3 | Hits@10 | MRR | Hits@1 | Hits@3 | Hits@10 |
| ConvE | 0.94 | 0.92 | 0.96 | 0.99 | 0.83 | 0.98 | 0.92 | 0.98 |
| ComplEx | 0.89 | 0.82 | 0.96 | 1 | 0.81 | 0.7 | 0.89 | 0.98 |
| MINERVA | 0.82 | 0.73 | 0.90 | 0.97 | 0.72 | 0.60 | 0.81 | 0.92 |
| NTP ¹ | 0.88 | 0.82 | 0.92 | 0.97 | 0.6 | 0.48 | 0.7 | 0.78 |
| NTP- λ ¹ | 0.93 | 0.87 | 0.98 | 1 | 0.8 | 0.76 | 0.82 | 0.89 |
| NTP 2.0 | 0.76 | 0.68 | 0.81 | 0.88 | 0.65 | 0.57 | 0.69 | 0.81 |
| DRUM | 0.81 | 0.67 | 0.94 | 0.98 | 0.61 | 0.46 | 0.71 | 0.91 |

: Comparison with transductive models

Nonetheless, DRUM is much faster(**1.2 min**) vs NTP(- λ)(**+8 hours**) on Kinship

Knowledge Graph Completion

- WN18RR, FB15k-237 respectively challenging variants of WN18, FB15k

Table 4: Transductive link prediction results. The results are taken from [25, 46] and [40]

| | WN18RR | | | | FB15K-237 | | | |
|-----------------|--------|------|------|------|-----------|------|------|------|
| | MRR | Hits | | | MRR | Hits | | |
| | | @ 10 | @ 3 | @ 1 | | @ 10 | @ 3 | @ 1 |
| R-GCN [36] | – | – | – | – | .248 | .417 | .258 | .153 |
| DistMult [45] | .43 | .49 | .44 | .39 | .241 | .419 | .263 | .155 |
| ConvE [10] | .43 | .52 | .44 | .40 | .325 | .501 | .356 | .237 |
| ComplEx [42] | .44 | .51 | .46 | .41 | .247 | .428 | .275 | .158 |
| Tucker [2] | .470 | .526 | .482 | .443 | .358 | .544 | .394 | .266 |
| ComplEx-N3 [25] | .47 | .54 | – | – | .35 | .54 | – | – |
| RotatE [40] | .476 | .571 | .492 | .428 | .338 | .533 | .375 | .241 |
| Neural LP [46] | .435 | .566 | .434 | .371 | .24 | .362 | – | – |
| MINERVA [8] | .448 | .513 | .456 | .413 | .293 | .456 | .329 | .217 |
| Multi-Hop [27] | .472 | .542 | – | .437 | .393 | .544 | – | .329 |
| DRUM (T=2) | .435 | .568 | .435 | .370 | .250 | .373 | .271 | .187 |
| DRUM (T=3) | .486 | .586 | .513 | .425 | .343 | .516 | .378 | .255 |

Table 9: Transductive link prediction results

| | WN18 | | | |
|-----------|------|------|------|------|
| | MRR | Hits | | |
| | | @ 10 | @ 3 | @ 1 |
| DistMult | .822 | .936 | .914 | .728 |
| ComplEx | .941 | .947 | .936 | .936 |
| Gaifman | – | .939 | – | .761 |
| R-GCN | .814 | .964 | .929 | .697 |
| TransE | .495 | .943 | .888 | .113 |
| ConvE | .943 | .956 | .946 | .935 |
| Neural LP | .94 | .945 | – | – |
| DRUM | .944 | .954 | .943 | .939 |

■ Quality and Interpretability of the Rules

- 특정 rule의 상위 confident 3개에 대한 평가(빨간색은 틀린 논리)

Table 7: Top 3 rules obtained by each system learned on *family* dataset

| | | | |
|-----------|--|---|--|
| Neural LP | brother(B, A) \leftarrow sister(A, B) | <i>wife(C, A) \leftarrow husband(A, B), husband(B, C)</i> | son(C, A) \leftarrow son(B, A), brother(C, B) |
| | brother(C, A) \leftarrow sister(A, B), sister(B, C) | wife(B, A) \leftarrow husband(A, B) | <i>son(B, A) \leftarrow brother(B, A)</i> |
| | brother(C, A) \leftarrow brother(A, B), sister(B, C) | <i>wife(C, A) \leftarrow daughter(B, A), husband(B, C)</i> | <i>son(C, A) \leftarrow son(B, A), mother(B, C)</i> |
| DRUM | brother(C, A) \leftarrow nephew(A, B), uncle(B, C) | wife(A, B) \leftarrow husband(B, A) | son(C, A) \leftarrow nephew(A, B), brother(B, C) |
| | brother(C, A) \leftarrow nephew(A, B), nephew(C, B) | wife(C, A) \leftarrow mother(A, B), father(C, B) | son(C, A) \leftarrow brother(A, B), mother(C, B) |
| | brother(C, A) \leftarrow brother(A, B), sister(B, C) | wife(C, A) \leftarrow son(B, A), father(C, B) | son(C, A) \leftarrow brother(A, B), daughter(B, C) |

: Head relation은 우에서 좌로, body atom은 좌에서 우로 읽음

■ Problems

- 이산 공간과 연속 공간을 동시에 optimize
- Neural LP의 한계
- Transductive system

■ DRUM

- Neural LP를 계승해 neural network를 사용한 Bidirectional RNN 기반 모델
- 학습 결과를 여러 head relation으로 공유 가능
- 간접적으로 score function을 optimize 하여 경량성 제공

Thank You!



Contact: Soho Moon (Email: moonwalk725@cau.ac.kr)