Multi-Hop Knowledge Graph Reasoning with Reward Shaping

Victoria Lin, Richard Socher, Caiming Xiong
{xilin, rsocher, cxiong}@salesforce.com

EMNLP 2018
Which directors has Tom Hanks collaborated with?
Which directors has Tom Hanks collaborated with?
Structured Query Answering

Knowledge Graph

Topic entity

Topic relation

Hanks collaborator?
Structured Query Answering

- Tom Hanks
- Meryl Streep
- United States
- California
- Steven Spielberg
- The Post
- Phyllida Lloyd

The diagram shows connections between entities such as location, residence, and collaboration. The relationship between Tom Hanks and another collaborator is marked as "Incomplete."
Knowledge Graph Embeddings

Highly accurate & Efficient

Lack interpretability

Why Spielberg is a collaborator of Hanks?

Tab 1. ConvE query answering performance on the UMLS benchmark dataset (Kok and Domingos 2007)

ConvE

MRR

0.957 (max = 1)

Why Spielberg is a collaborator of Hanks?

DistMult (Yang et al. 2015), ComplEx (Trouillon et al. 2016), ConvE (Dettmers et al. 2018).
Multi-Hop Reasoning Models

Reasoning over discrete structures

United States

California

locates_in

live_in

produced_in

born_in

cast_in

collaborator?

Steven Spielberg

The Post

Tom Hanks

collaborator?

Phyllida Lloyd

Meryl Streep
Multi-Hop Reasoning Models

Sequential decision making

Steve Spielberg

California

United States

Tom Hanks

The Post

collaborator?

Phyllida Lloyd

Meryl Streep

born_in
locate_in
live_in
produced_in
director
cast_in
collaborator
Multi-Hop Reasoning Models

- Tom Hanks

- Steven Spielberg

- The Post

- Meryl Streep

- Phyllida Lloyd

- California

- United States

- born_in

- locate_in

- live_in

- produced_in

- director

- cast_in

- collaborator
Multi-Hop Reasoning Models

Tom Hanks

The Post

Stephen Spielberg

United States

California

locate_in

live_in

produced_in

born_in

director

cast_in

The Post

collaborator?

Steve Spielberg

Meryl Streep

Phyllida Lloyd

collaborator?
Multi-Hop Reasoning Models

Tom Hanks

cast_in

The Post
director

Stephen Spielberg

<END>
Multi-Hop Reasoning Models

MINERVA (Das et al. 2018)
Multi-Hop Reasoning Models

Interpretable

Significant performance gap

<table>
<thead>
<tr>
<th></th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvE</td>
<td>0.957</td>
</tr>
<tr>
<td>RL</td>
<td>0.825</td>
</tr>
</tbody>
</table>

Tab 2. ConvE and RL (MINERVA) query answering performance on the UMLS benchmark dataset (Kok and Domingos 2007)
Multi-Hop Reasoning Models: Ideal Case

Interpretable

Significant performance gap

<table>
<thead>
<tr>
<th></th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvE</td>
<td>0.957</td>
</tr>
<tr>
<td>RL</td>
<td>0.825</td>
</tr>
</tbody>
</table>

Tab 2. ConvE and RL (MINERVA) query answering performance on the UMLS benchmark dataset (Kok and Domingos 2007)
Challenges

Incompleteness

Training example

16
Challenges

Incompleteness

No reward

Overfit to the observed answers

Training example

United States

California

locate_in

born_in

director

produced_in

live_in

Tom Hanks

Phyllida Lloyd

collaborator?

collaborator?

collaborator?

Meryl Streep

The Post

cast_in

cast_in

cast_in

cast_in
Challenges

Path Diversity

1. locate_in California
2. live_in United States
3. director The Post
4. born_in Tom Hanks
5. cast_in collaborator?
6. produced_in Steven Spielberg
7. cast_in collaborator? Meryl Streep
8. collaborator Phyllida Lloyd
Challenges

Path Diversity

False positive (spurious) paths ① ②

Overfit to the spurious paths

1. Meryl Streep
2. Tom Hanks
3. Phyllida Lloyd
4. Steven Spielberg
5. United States
6. California
7. The Post
8. collaborator?
Proposed Solutions

- Incompleteness
- Reward Shaping
- Policy Gradient
- Path Diversity
- Action Dropout
Reinforcement Learning Framework
Reinforcement Learning Framework

Environment  State  Action  Transition  Reward

\[ e_s \rightarrow r_q \]

\[ e_s \]
Reinforcement Learning Framework

Environment  State  Action  Transition  Reward

$e_s \xrightarrow{r_q} e^1$

$e_s \xrightarrow{r^1} e^1$

$e_s \xrightarrow{r^2} e^2$

$\vdots$

$e_s \xrightarrow{r^N} e^N$

$A_t$
Reinforcement Learning Framework

Environment → State → Action → Transition → Reward

$e_s \rightarrow r_q \rightarrow e^1 \rightarrow \cdots \rightarrow e^N$

$e_s \rightarrow r^1 \rightarrow e^1 \rightarrow \cdots \rightarrow e^2 \rightarrow \cdots \rightarrow e^N$

$e_s \rightarrow r^N \rightarrow e^N$

$A_t$
Reinforcement Learning Framework

Environment  State  Action  Transition  Reward

$e_s \rightarrow r_q \rightarrow e_1$

$e_s \rightarrow r_1 \rightarrow e_1$
Reinforcement Learning Framework

Environment → State → Action → Transition → Reward

$e_s \rightarrow rq \rightarrow e_1 \rightarrow r_2 \rightarrow e_2$
Reinforcement Learning Framework

Environment → State → Action → Transition → Reward

- $e_s \rightarrow r_q \rightarrow e_1$
- $e_s \rightarrow r_1 \rightarrow e_1$
- $r_t \rightarrow e_t$
- $r_T \rightarrow e_T$

Max # steps
Reinforcement Learning Framework

Environment $e_s$ $r_q$ $e_1$

State $r_1$ $r_t$ $e_t$

Action $r_T$ $e_T$

Reward

Predicted answer
Reinforcement Learning Framework

$e_s \rightarrow r_q \rightarrow e_1$

$e_s \rightarrow r_1 \rightarrow e_1$ ... $r_t \rightarrow e_t$ ... $r_T \rightarrow e_T$

$R_b(s_T) = 1\{(e_s, r_q, e_T) \in G\}$
Reinforcement Learning Framework

Environment → State → Action → Transition → Reward

$e_s \rightarrow r_q \rightarrow \ldots \rightarrow r_t \rightarrow e_t \rightarrow \ldots \rightarrow r_T \rightarrow e_T$

Learn *which action to choose* given a state

$R_b(s_T) = 1\{(e_s, r_q, e_T) \in G\}$
Policy Gradient

Policy function: \( \pi_{\Theta}(a_t \mid s_t) \)

Probability of choosing an action given the current state:

- \( \pi_{\Theta}(a_t^1 \mid s_t) \)
- \( \pi_{\Theta}(a_t^2 \mid s_t) \)
- \( \ldots \)
- \( \pi_{\Theta}(a_t^N \mid s_t) \)

Action at time step \( t \): \( a_t^i = (r_t^i, e_t^i) \)
Policy Gradient

Policy function

Design choice: include search history in the state representation

current state

MINERVA (Das et al. 2018)
Policy Gradient

Policy function

$\pi_{\Theta}(a_t | s_t)$

Our model extensions are applicable to any parameterization of $\pi_{\Theta}$

Design choice: include search history in the state representation

MINERVA (Das et al. 2018)
REINFORCE Training

Sample according to $\pi_{\Theta}(a_t | s_t)$

Update $\Theta$ by maximizing the expected reward

REINFORCE (Williams, 1992)
REINFORCE Training

*False-negative & nearly-correct entities ➞ true-negatives*

```
Tom Hanks
  cast_in
    The Post
      cast_in
        Meryl Streep
          collaborator
            Phyllida Lloyd
              +1

Tom Hanks
  cast_in
    United States
      director
        Steven Spielberg
          live_in
            Steven Spielberg
              0

Tom Hanks
  born_in
    California
      locate_in
        United States
          live_in
            Steven Spielberg
              0
```

35
Reward Shaping

Unobserved facts

Topic entity

$f_\theta$

Entity returned by search

Topic relation

Tom Hanks \textit{collaborator} Steven Spielberg

Soft correctness
Reward Shaping

Unobserved facts

Topic entity

Topic relation

$f_\theta$

General

Soft correctness

Entity returned by search

Tom Hanks collaborator Steven Spielberg
Reward Shaping

Pre-trained
Value range: (0, 1)
Action Dropout

Intuition: *avoid sticking to* past actions that *had received rewards*
Action Dropout

Intuition: *avoid sticking to past actions that had received rewards*

More likely to be chosen

\[ \pi_\Theta(a_t^i | s_t) \]

- 0.08
- 0.6
- 0.01

Intuition:

Avoid sticking to past actions that had received rewards.
Action Dropout

Randomly offset the \textit{sampling probabilities} \textit{w/} rate $\alpha$ and renormalize

\[
\tilde{\pi}_\Theta(a_t | s_t) \propto \pi_\Theta(a_t | s_t) \cdot m + \epsilon
\]

\[
m_i \sim \text{Bernoulli}(1 - \alpha), \ i = 1, \ldots, N
\]

More likely to be chosen

| $\tilde{\pi}_\Theta(a_t^i | s_t)$ | $\pi_\Theta(a_t^i | s_t)$ |
|--------------------------|--------------------------|
| 0.9                      | 0.08                     |
| 0                        | 0.6                      |
| 0                        | 0.01                     |
Randomly offset the sampling probabilities with rate \( \alpha \) and renormalize.

\[
\tilde{\pi}_\Theta(a_t^i | s_t) \propto \pi_\Theta(a_t^i | s_t) \cdot m + \epsilon
\]

\( m_i \sim \text{Bernoulli}(1 - \alpha), i = 1, \ldots, N \)

More likely to be chosen

\[
\tilde{\pi}_\Theta(a_t^i | s_t) \quad \pi_\Theta(a_t^i | s_t)
\]

0.9 0.08

0 0.01

Force exploration
Action Dropout

Randomly offset the *sampling probabilities* w/ rate $\alpha$ and renormalize

$\tilde{\pi}_\Theta(a_t \mid s_t) \propto \pi_\Theta(a_t \mid s_t) \cdot m + \epsilon$

$m_i \sim \text{Bernoulli}(1 - \alpha), i = 1, \ldots N$

More likely to be chosen

More likely to be chosen

up to $\times 8$ # path traversed

$\tilde{\pi}_\Theta(a_t^i \mid s_t) \propto \pi_\Theta(a_t^i \mid s_t)$

$0.9$$0.08$

$0$$0.01$

Force exploration
# Experiment Setup

## KG Benchmarks

<table>
<thead>
<tr>
<th>Name</th>
<th># Ent.</th>
<th># Rel.</th>
<th># Fact</th>
<th># Degree Avg</th>
<th># Degree Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinship</td>
<td>104</td>
<td>25</td>
<td>8,544</td>
<td>85.15</td>
<td>82</td>
</tr>
<tr>
<td>UMLS</td>
<td>135</td>
<td>46</td>
<td>5,216</td>
<td>38.63</td>
<td>28</td>
</tr>
<tr>
<td>FB15k-237</td>
<td>14,505</td>
<td>237</td>
<td>272,115</td>
<td>19.74</td>
<td>14</td>
</tr>
<tr>
<td>WN18RR</td>
<td>40,945</td>
<td>11</td>
<td>86,835</td>
<td>2.19</td>
<td>2</td>
</tr>
<tr>
<td>NELL-995</td>
<td>75,492</td>
<td>200</td>
<td>154,213</td>
<td>4.07</td>
<td>1</td>
</tr>
</tbody>
</table>

**Evaluation Protocol:** MRR (Mean Reciprocal Rank)
Ablation Studies

Fig 2. Dev set MRR (x100) comparison
Ablation Studies

Fig 2. Dev set MRR (x100) comparison
Ablation Studies

Fig 2. Dev set MRR (x100) comparison
Ablation Studies

Fig 2. Dev set MRR (x100) comparison

Esp. helpful on dense graphs
Main Results

Fig 3. Test set MRR (x100) compared to SOTA multi-hop reasoning and embedding-based approaches
Main Results

Fig 3. Test set MRR (x100) compared to SOTA multi-hop reasoning and embedding-based approaches

拇指向上，表示改进的SOTA多跳推理性能
Main Results

Fig 3. Test set MRR (x100) compared to SOTA multi-hop reasoning and embedding-based approaches
Interpretable Results

Laura Carmichael \( \xrightarrow{\text{profession}} \) ?

FB15k-237 (Toutanova and Chen 2016)
Interpretable Results

Laura Carmichael \(\rightarrow\) profession \(\rightarrow\) ?

Richard Bonneville

Penelope Wilton

Laura Carmichael

SAG Award

Don Cheadle

United Kingdom

Ricky Gervais

Elizabeth McGovern

actor/actress

comedienne

Laura Carmichael

/award_winner

/people/person/profession

United Kingdom

/people/person/profession

/award_winner

/people/person/profession

TV producer

FB15k-237 (Toutanova and Chen 2016)
Interpretable Results

FB15k-237 (Toutanova and Chen 2016)
Future directions

- Learn better reward shaping functions
- Investigate similar techniques for other RL paradigms (e.g. Q-learning)
- Extend to more complicated structured queries (e.g. more than one topic entities)
- Extend to natural language QA
Fig 4. Dev set top-1 prediction error overlap of ConvE, Ours and Ours-RS. The absolute error rate of Ours is shown.
BK II - Challenges

Incompleteness

≈ 30% false negative feedback

Fig 1. % of false negative hit in the first 20 epochs of RL training on the UMLS KG benchmark (Kok and Domingos 2007)
Questions for Future Research

1. One natural question to ask is why a perceivable performance gap exists between the embedding-based (EB) model and the RL approach using the same EB model as the reward shaping module (slide 51), especially on FB15k-237 and NELL-995, the two larger and sparser KGs. Since the RL model has full access to the EB model, why could it still lose information? A possible explanation is that for examples where the RL models make mistakes, the topic entity and the target answer are not connected within the specified # hops. Yet our sanity check disproved this — for all examples where only the RL model makes mistakes the topic entity and the target answer were connected by at least one path. (We did not check the quality of these paths.) Conjecture: It is possible that the performance loss comes from the difficulty of RL optimization as it operates over a more complex model space. The RL model + training procedure have much more hyper-parameters than the EB models.

2. In our experiments, very large action dropout rates (0.9 and 0.95) yield good performance on the dense KGs (Kinship and UMLS), but the same strategy does not work for sparser KGs. We observed significant performance drop for FB15k-237, WN18RR and NELL-995 when using very large action dropout rates. And for WN18NN and NELL-995, action dropout rate > 0.1 hurts performance. It is unclear why REINFORCE training on the denser KGs can tolerate a larger shift from the actual policy during path sampling. Conjecture: It seems that the shape of the original policy function ought to be preserved to some degree during training. For Kinship and UMLS, the average node degrees are 85 and 39. In this case on average >= 2 edges remains on when we randomly turned off 95% of the edges. Since other KGs have smaller average node degrees, using a large action dropout rate is equivalent to doing random exploration most of the time.

3. Does EB models define the cap performance in the one-hop KG query answering set up? Could the tasks of path finding and learning KG embeddings be joined together in a way s.t. they can improve each other?

4. Our approach can be viewed as a way to explain pre-trained EB models. Are there better ways to do it?
Acknowledgement upon slides release - I

These slides benefit tremendously from the constructive feedback offered by Salesforce Research team members, including Caiming Xiong, Richard Socher, Yingbo Zhou, Alex Trott, Jin Qu, Lily Hu, Vena Li, Kazuma Hashimoto, Stephan Zheng, Jason Wu (intern) and others.
Acknowledgement upon slides release - II

I am grateful to Prof. Michael Ernst and his co-authors who released the slides of all of their paper presentations.

Especially, the Bradley Hand Font and the Balloon highlighters were borrowed from slides 1 and slides 2.

Victoria Lin
Nov. 4, 2018