Embedding Probabilistic Logic for Machine Reading

aka Towards Two-Way Interaction with Reading Machines

Sebastian Riedel (University College London)
Collaborators

- Tim Rocktäschel
  UCL

- Matko Bosnjak
  UCL

- Ivan Sanchez
  UCL

- Sameer Singh
  UWash

- Limin Yao
  UMass Amherst (Twitter)

- Andrew McCallum
  UMass Amherst

- Ben Marlin
  UMass Amherst
“Who works in London and is interested in NLP?"
"Who works in London and is interested in NLP?"

interest(x,NLP),
workFor(x,y),
in(y,London)

"Sebastian Riedel works in the area of NLP and is now Lecturer at UCL"
“Who works in London and is interested in NLP?"

\[
\text{interest}(x,\text{NLP}), \quad \text{worksFor}(x,y), \quad \text{in}(y,\text{London})
\]

Statistical Relational Learner and Reasoner

"Sebastian Riedel works in the area of NLP and is now Lecturer at UCL"
Benefit: Transitive Reasoning

“Who works in London and is interested in NLP?

\[
\text{interest}(x, \text{NLP}), \text{worksFor}(x,y), \text{in}(y, \text{London})
\]

\[
\text{in}(\text{UCL}, \text{London})
\]

\[
\text{works-in-area-of}(\text{Seb}, \text{NLP})
\]

\[
\text{lecturer-at}(\text{Seb}, \text{UCL})
\]

\[
\text{worksFor}(x,y):
\quad \text{faculty-at}(x,y)
\]

\[
\text{interest}(x,y):
\quad \text{works-in-area-of}(x,y) [0.9]
\]

\[
\text{faculty-at}(x,y):
\quad \text{lecturer-at}(x,y)
\]

Wide universal schema

Syntax

Coreference

Statistical Relational Learner and Reasoner

Statistical NLP

"Sebastian Riedel works in the area of NLP and is now Lecturer at UCL"
Benefit: More Coverage

“Who is faculty in London and interested in NLP?

interest(x,NLP), worksFor(x,y), in(y,London)

faculty-at(x,y): works-for(x,y)

lecturer-at(x,y)

Syntax

Coreference

Wide universal schema

Statistical Relational Learner and Reasoner

Statistical NLP

"Sebastian Riedel works in the area of NLP and is now Lecturer at UCL"
Benefit: Code Reuse

“Who lives in London and is interested in NLP?

interest(x,NLP), worksFor(x,y), in(y,London)

Statistical Relational Learner and Reasoner

[Laò et al., 2011]

in(UCL,London)
works-in-area-of(Seb,NLP)
lecturer-at(Seb,UCL)
worksFor(x,y):
  faculty-at(x,y)

interest(x,y):
  works-in-area-of(x,y)[0.9]

livesIn(x,z):
  worksFor(x,y), locatedIn(y,z) [0.6]

Wide universal schema

Syntax Coreference

“Sebastian Riedel works in the area of NLP and is now Lecturer at UCL“
Joint Inference

“Who lives in London and is interested in NLP?

interest(x,NLP),
worksFor(x,y),
in(y,London)

Wide universal schema

in(UCL,London)
works-in-area-of(Seb,NLP)
lecturer-at(Seb,UCL)
worksFor(x,y):
  faculty-at(x,y)
interest(x,y):
  works-in-area-of(x,y) [0.9]
livesIn(x,z):
  worksFor(x,y),
locatedIn(y,z) [0.6]

Syntax Coreference

Statistical Relational Learner and Reasoner

Statistical NLP

"Sebastian Riedel works in the area of NLP and is now Lecturer at UCL"
Reasoner and Learner

Statistical Relational Learner and Reasoner
Probabilistic Logics

Use (weighted) logics to define graphical models

Examples

Markov Logic
[Richardson and Domingos, 2006]

Bayesian Logic Programs
[Kersting, 2007]
Probabilistic Logics

Use (weighted) logics to define graphical models
Matrix Factorization

Think of database as a matrix or tensor

<table>
<thead>
<tr>
<th>lecturer-at</th>
<th>prof-at</th>
<th>works-for</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

13
Matrix Factorization

*Embed entity (pairs) in low dimensional vector spaces*

<table>
<thead>
<tr>
<th>lecturer-at</th>
<th>prof-at</th>
<th>works-for</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>
Matrix Factorization

Embed relations in low dimensional vector spaces

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & ? & ? \\
? & ? & ? \\
? & ? & ? \\
\end{array}
\]

\begin{array}{ccc}
1 & 1 & 1 \\
1 & ? & ? \\
? & ? & ? \\
? & ? & ? \\
\end{array}

\text{lecturer-at} \quad \text{prof-at} \quad \text{works-for}
Matrix Factorization

Find a matrix-matrix product that approximates observed DB

\[
\begin{pmatrix}
1 & 1 & 1 \\
1 & & 1 \\
1 & & 1 \\
1 & & 1
\end{pmatrix}
\approx
\begin{pmatrix}
? & ? & ? \\
? & ? & ? \\
? & ? & ? \\
? & ? & ?
\end{pmatrix}
\times
\begin{pmatrix}
lecturer-at & prof-at & works-for \\
? & ? & ? \\
? & ? & ? \\
? & ? & ?
\end{pmatrix}
\]
Matrix Factorization

Or a non-linear function of this product

\[
\begin{pmatrix}
1 & 1 \\
1 & 1 \\
1 & 1 \\
1 & 1 \\
\end{pmatrix} 
\approx \text{sigmoid}
\]
Matrix Factorization

Low rank forces some 0 cells to become non-zero => prediction

\[
\begin{pmatrix}
1 & 1 & 1 \\
1 & .9 & \\
1 & 1 & .9 \\
1 & &
\end{pmatrix}
\approx \text{sigmoid}
\]

[Nickel, Bordes, …]
Overview

Matrix Factorization Models

“Talking to Reading Machines”

Injecting Knowledge

Extracting Knowledge

Data

sigmoid

×

“lecturers are employees!”

sigmoid

×

“lecturers are employees?”

sigmoid

×
Universal Schema Matrix

Schema contains structured and unstructured (~OpenIE) relations

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20
Goal: Learn to Complete

Schema contains structured and unstructured (~OpenIE) relations

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Model N: Baseline Classifier

[ Mintz et al 2009, ... ]

Standard supervised relation extractor ...

\[
p( y_{emp}^x, y_{emp} = 1 | )
\]
Model N: Classifier

Standard supervised relation extractor ...

\[ p(y_{\text{emp}}^x,y^x,\cdot) = 1 | \mathbf{f}^{x,y}_{\text{emp}} \]
Model N: Classifier

[Mintz et al 2009,...]

Standard supervised relation extractor ...

\[ p(y_{\text{emp}}^{x,y} = 1 | f_{\text{emp}}^{x,y}, w_{\text{emp}}) \]
Model N: Classifier

[Mintz et al 2009,...]

Standard supervised relation extractor ...
Model N: Classifier

... for each pattern

\[ p(y_{prof}^{x,y} = 1 | f_{prof}^{x,y}, w_{prof}) \propto \exp[< f_{prof}^{x,y}, w_{prof}>] \]
Model F: Latent Feature (Factorization)

[Collins et al, 2001]

Model the probability of a pair \((x, y)\) being in relation “prof”

\[
p(\nu_{y^{prof}}^{x,y} = 1|\nu^{x,y}, \nu_{prof}) \propto \exp[<\nu^{x,y}, \nu_{prof}>]
\]
Model F: Latent Feature (Factorization)

[Collins et al, 2001]

Per tuple latent feature vector

\[ p(y_{prof}^{x,y} = 1 | \mathbf{v}^{x,y}, \mathbf{w}_{prof}) \propto \exp[< \mathbf{v}^{x,y}, \mathbf{w}_{prof}>] \]
Model F: Latent Feature (Factorization)

[Collins et al, 2001]

Per tuple latent feature vector

\[
p(y_{\text{prof}}^{x,y} = 1|v^{x,y}, w_{\text{prof}}) \propto \exp[<v^{x,y}, w_{\text{prof}}>]\]

Model F: Latent Feature (Factorization)

[Collins et al, 2001]

Per tuple latent feature vector

\[ p(y_{\text{prof}}^{x,y} = 1|v^{x,y}, w_{\text{prof}}) \propto \exp[<v^{x,y}, w_{\text{prof}}>] \]

\[ = \text{sigmoid}(<v^{x,y}, w_{\text{prof}}>) \]
Model F: Latent Feature (Factorization)

Transitive Reasoning

\[ \text{X-is-historian-at-Y} \quad \text{X-is-professor-at-Y} \quad \text{X-museum-at-Y} \quad \text{X-teaches-history-at-Y} \quad \text{employee(X,Y)} \]
Model F: Latent Feature (Factorization)

Transitive Reasoning

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>?</td>
</tr>
<tr>
<td>w</td>
<td>w</td>
<td>w</td>
<td>w</td>
<td>w</td>
</tr>
</tbody>
</table>
Model F: Latent Feature (Factorization)

Transitive Reasoning

\[ \text{X-is-historian-at-Y} \quad \text{X-is-professor-at-Y} \quad \text{X-museum-at-Y} \quad \text{X-teaches-history-at-Y} \quad \text{employee(X,Y)} \]
Model F: Latent Feature (Factorization)

Transitive Reasoning

X-is-historian-at-Y  X-is-professor-at-Y  X-museum-at-Y  X-teaches-history-at-Y  employee(X,Y)
Model F: Latent Feature (Factorization)

Transitive Reasoning

\[
\begin{array}{ccccc}
\text{X-is-historian-at-Y} & \text{X-is-professor-at-Y} & \text{X-museum-at-Y} & \text{X-teaches-history-at-Y} & \text{employee(X,Y)} \\
\end{array}
\]
Model F: Latent Feature (Factorization)

Transitive Reasoning

- **X-is-historian-at-Y**
- **X-is-professor-at-Y**
- **X-museum-at-Y**
- **X-teaches-history-at-Y**
- **employee(X,Y)**

![Diagram](image-url)
Model F: Latent Feature (Factorization)

Transitive Reasoning

X-is-historian-at-Y  X-is-professor-at-Y  X-museum-at-Y  X-teaches-history-at-Y  employee(X,Y)

r  r  r

w  w  w  w

v  v  v  v

0.8
Model F: Latent Feature (Factorization)

Bootstrapping without fantasy

\[ \begin{align*}
\text{X-is-historian-at-Y} & \quad \text{X-is-professor-at-Y} & \quad \text{X-museum-at-Y} & \quad \text{X-teaches-history-at-Y} & \quad \text{employee(X,Y)} \\
\text{r} & \quad \text{r} & \quad \text{r} & \quad \text{r} & \quad \text{r} \\
\text{w} & \quad \text{w} & \quad \text{w} & \quad \text{w} & \quad \text{w} \\
\end{align*} \]
## Model E: Selectional Preferences

Relations have **entity type restriction**

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model E: Selectional Preferences

Relations have **entity type restriction**

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

0 0 0
Model E: Selectional Preferences

Argument Slot 1 weight vector ...

\[ p(y_{\text{prof}}^x, y_{\text{prof}}^y = 1 | \ldots) \propto \exp[ < v^x, w_{\text{prof}}^1 > + < v^y, w_{\text{prof}}^2 > ] \]
Model E: Selectional Preferences

... dot-product with feature vector of entity 1

\[ p(y^x_{\text{prof}} = 1 \mid \ldots) \propto \exp[<v^x, w^1_{\text{prof}}> + <v^y, w^2_{\text{prof}}>] \]
Model E: Selectional Preferences

Argument Slot 2 weight vector ...

\[ p(y^x_{\text{prof}} = 1 | \ldots) \propto \exp[<v^x, w^1_{\text{prof}}> + <v^y, w^2_{\text{prof}}>] \]
Model E: Selectional Preferences

... dot-product with feature vector of entity 2

\[ p(y_{\text{prof}}^x, y_{\text{prof}}^y = 1 | \ldots) \propto \exp[<v^x, w^1_{\text{prof}}> + <v^y, w^2_{\text{prof}}>] \]
Combinations

models capture different aspects of the data, combine them (e.g., NF)

\[ p(y_{\text{emp}}^{x,y} = 1 | \ldots) \propto \exp[<f_{\text{emp}}^{x,y}, w_{\text{emp}}^{N}> + <v^{x,y}, w_{\text{emp}}^{F}>] \]
Evaluation (Structured)

[Mintz 09; Yao 11; Surdenau 12]

Evaluate *average precision* per Freebase relation.

<table>
<thead>
<tr>
<th>Relation</th>
<th>MI09</th>
<th>YA11</th>
<th>SU12</th>
<th>N+F+E</th>
</tr>
</thead>
<tbody>
<tr>
<td>employee</td>
<td>0.67</td>
<td>0.64</td>
<td>0.7</td>
<td>0.79</td>
</tr>
<tr>
<td>containedby</td>
<td>0.48</td>
<td>0.51</td>
<td>0.54</td>
<td>0.69</td>
</tr>
<tr>
<td>parents</td>
<td>0.24</td>
<td>0.27</td>
<td>0.58</td>
<td>0.39</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Weighted MAP</td>
<td>0.48</td>
<td>0.52</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>MAP</td>
<td>0.32</td>
<td>0.42</td>
<td>0.56</td>
<td>0.63</td>
</tr>
</tbody>
</table>

~45 minutes to train our models on 4000 relations, ~50k entity pairs
Injecting Knowledge

“lecturers are employees!”
Injecting Knowledge

“a liquid turns into a solid when its temperature is lowered below its freezing point"
Injecting Knowledge: Rules

\[ \forall x, y: \text{birthplace}(x, y) \Rightarrow \text{bornIn}(x, y) \]
Goal: Predict Unseen Cells ...

<table>
<thead>
<tr>
<th>native-of</th>
<th>'s birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
... By Using Rules and Data

$$\forall x, y: \text{birthplace}(x, y) \Rightarrow \text{bornIn}(x, y)$$
Baselines

\[ \forall x, y: \text{birthplace}(x, y) \Rightarrow \text{bornIn}(x, y) \]

<table>
<thead>
<tr>
<th>native-of 's birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>.8</td>
</tr>
<tr>
<td>1</td>
<td>.8</td>
<td>.8</td>
</tr>
</tbody>
</table>

- **Rules only**
- **Rules after learning**
- **Rules before learning**
Pre-Injection may not add data at all

<table>
<thead>
<tr>
<th>native-of</th>
<th>birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

∀x,y: birthplace(x,y) ⇒ bornIn(x,y)

Rules before learning
Pre-Injection may not add data at all

\[ \forall x, y: \text{birthplace}(x, y) \Rightarrow \text{bornIn}(x, y) \]

<table>
<thead>
<tr>
<th>native-of's birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rules before learning
Idea: Iterate

<table>
<thead>
<tr>
<th>native-of</th>
<th>'s birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

∀x,y: birthplace(x,y) ⇒ bornIn(x,y)

Inference with model
Apply rules
... and learn again
Our approach

[Rocktaeschel et al 15]

- **Directly** optimise to **fulfil formulae in expectation**
- formulae have **compositional expectations**

\[
E_{v,w}[\text{birthplace}(\text{Seb}, \text{HH})] = E_{v,w}[y_{\text{birthplace}}^{\text{Seb,HH}}] = \text{sigm}(<v^{\text{Seb,HH}}, w_{\text{birthplace}}>)
\]

\[
E_{v,w}[r(X_1, X_2)] = \text{sigm}(<v^{X_1,X_2}, w_r>)
\]

\[
E_{v,w}[A \land B] = E_{v,w}[A] \times E_{v,w}[B]
\]

\[
E_{v,w}[\neg A] = 1 - E_{v,w}[A]
\]

\[
E_{v,w}[A \Rightarrow B] = 1 - (E_{v,w}[A] \times (1 - E_{v,w}[B]))
\]
Our approach

[Rocktaeschel et al 15]

- **Directly** optimise to **fulfil formulae in expectation**
- formulae have **compositional expectations**
- **quantification** through **grounding**

\[
E_{v,w}[\forall x. f(x)] = E_{v,w}[f(X_1) \land \ldots \land f(X_n)] \\
= E_{v,w}[f(X_1)] \times \ldots \times E_{v,w}[f(X_n)]
\]
General Framework

[Rocktaeschel et al 15]

- Find embeddings $v$ and $w$ that...

- Maximize **log expectation** of a set of formulae $f$

\[
\text{arg max}_{v,w} \sum_f \log(E_{v,w}[f])
\]

- **Generalizes** regular (binary) matrix factorization with logistic loss

- Get gradients by **back-propagation** through log($E[.]$) tree

- Optimize via **SGD** / Adagrad etc.
Experiments

[Rocktaeschel et al 15]

- “Zero-shot” learning

- Given: a lot of surface form data, but no Freebase relations

- Goal: given few (36) Freebase rules, learn to Freebase relations
## Experiments: Zero-Shot Learning

Remove Freebase data from training set …

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Experiments: Zero-Shot Learning

and learn only from surface form relations, and rules

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Zero-Shot Learning Results (MAP)

[Rocktaeschel et al 15]
Learning Curve

[Rocktaeschel et al 15]
Generating Data?

<table>
<thead>
<tr>
<th>native-of’s birthplace</th>
<th>bornIn</th>
<th>livesIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Hasn’t worked yet
➢ Row embeddings overtrain
➢ At test time premise appears with other relations
Challenge 1: Injecting Symbolic Rules

“lecturers are employees!” ➔ sigmoid ➔ [ ]
Challenge 2: Extracting Explanations

"lecturers are employees!"

sigmoid

$x, y$

$A$

$B$

$R$

$P$

$F$

$A = \{ (x, y) \mid P(x, y) \}$

$\left[ \begin{array}{c} \text{San Jose} \\ \text{Sikorsky Aircraft} \\ \text{California} \end{array} \right]$
Challenge 2: Extracting Explanations

"I returned Sebastian because we know he is a lecturer at UCL, which is in London, so he most likely lives in London..."
“Knowledge Extraction”

Learn a more interpretable model from distributed representations (for interpretation, not for use)

- Neural Networks => if-then rules (Thrun, 95)
- Neural Networks => Decision Trees (Craven, 96)

Open Questions

- Go beyond classification: joint models
- Use to provide proofs of complex predictions
- Integrate into a dialog between human and machine
Explaining Matrix Factorization

[Sanchez et al. 2015, KRR]

Data $|\mathcal{P}|$ $\Rightarrow$ $k$ $\approx$ Matrix Factorization

Bayesian Network

User

First-order Logic

$\forall p : r_1(p) \Rightarrow r_2(p)$

$\forall p : r_2(p) \Rightarrow r_4(p)$

$\forall p : r_3(p) \Rightarrow r_4(p)$
Extracting Bayesian Networks

Learn Embeddings from Data

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extracting Bayesian Networks

Generate data from embeddings (threshold or sample)

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Extracting Bayesian Networks

Learning a tree shaped Bayesian Network

<table>
<thead>
<tr>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Benefits of Bayesian Network Trees

[Sanchez et al. 2015, KRR]

- Provide a **joint model** over all relations
  - more **compact** (than one decision tree per relation)
  - more **faithful** to the **joint** MF model

- **Probabilistic** interpretation, captures probabilistic nature of MF

- Very **scalable**
  - Learning: **Prim Algorithm** to find Maximum Spaning Tree over mutual information
  - Inference: Belief Propagation in **non-loopy** graph
Faithfulness

![Averaged 11-point Precision/Recall](image)

Figure 1: Fidelity of the descriptive models to the MF model.

That is, the ranked list of facts produced by the logical rules matches the predictions of the MF poorly. This might be due to their deterministic nature: since their response is in a binary domain they are not able to provide a confidence value in $[0,1]$. This makes it difficult to capture the ranking behavior of the MF model.

The decision trees provide more sensible confidence scores and hence rankings. This is reflected in better average precision curves. The BN model outperforms the other models substantially. We believe that this is a consequence of its probabilistic formulation, and the ability to capture the joint nature of the MF model better.

Interpretability

We show two examples of "explanations" for wrong predictions as produced by the different descriptive models. Figure 2a shows the causes for the MF model to predict the wrong fact arenaStadium (PhiladelphiaEagles, Canton) with confidence of 0.885. We see a BN explanation in Figure 2a: the snippet of the BN that connects observed relations and the prediction. The observed node, playAt, influences the next nodes in the trajectory towards the target node arenaStadium. A clear error of the MF model is indicated by the connection in the BN: beatAt ! arenaStadium.

Following the above example, the decision tree learned specifically for the relation arenaStadium shows a confidence of 1.0 for the same fact. The explanation is the rule if playAt = 1 then p(arenaStadium = 1) = 1.

We think that the interpretability of the decision tree and the BN are quite comparable in this case. By contrast, the logical system does not even predict this fact as it missed the implication.

In Figure 2b we can see the explanation of why the fact reviewMovie (DanielKahneman, Nobel) was predicted by the MF model as true. Given that this text pattern is not a target Freebase relation, no decision tree was learned for it, so no explanation from this model can be sought. On the other side, in the set of logic rules none of the observed neither the target pattern appeared, meaning that their statistical dependence with respect to other patterns was really low.

6 Conclusion

The problem of finding interpretable descriptive models for high-performance latent variable models has been discussed before, but we believe it is time for the community to revisit it. The reasons are both the recent successes of latent variable models, and the increasing complexity of the tasks they address. In particular, in this work we looked at matrix factorization models for knowledge base population, a more complex task than the classification problems considered in existing literature.

As the starting point we proposed three descriptive representations: implication logic rules, decision trees and Bayesian network trees—a representation that has not been previously considered. We found that Bayesian Network trees provide a very competitive combination of fidelity and interpretability.
A “Proof”

- Model observed $playAt(\text{Eagles, Canton})$

- Model wrongly predicted $arenaStadium(\text{Eagles, Canton})$

- The Bayesian Network can provide this “proof”

- Todo: evaluate this in a downstream “debugging” task
Summary

- Do semantics in a **probabilistic relational reasoner**
- Reasoner: **matrix/tensor factorization** (or other LV models)
- Models itself don’t need to be interpretable if we know …
- … how to **Interact with uninterpretable models**
  - **inject** explanations and logical rules
    - Approach: **optimize** embeddings to fulfil formulae
  - **extract** explanations
    - for example: by using an interpretable BN **proxy** model
Thanks
Training
Negative Data

Usually **unavailable** or **sparse**, so...

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Negative Data

…subsample, which can work…

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
but often *does not*

<table>
<thead>
<tr>
<th></th>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Negative Data**
Negative Data

and you need to sample a lot (wasting resources)

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Implicit Feedback

Often users only **click/view/buy** items, or not, but no rating

<table>
<thead>
<tr>
<th></th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
<th>User 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Item 2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ranking

[Rendle et al., 09]

for all (observed, not observed) pairs in column: \( \text{prob(o)} > \text{prob(n)} \)

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ranking

[Rendle et al., 09]

for all (observed, not observed) pairs in a column: \( \text{prob(o)} > \text{prob(n)} \)

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X, Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training: Stochastic Gradient Descent

[Rendle et al., 09]

Sample observed fact...

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training: Stochastic Gradient Descent

[Rendle et al., 09]

Sample unobserved cell for same relation

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X, Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
**Training: Stochastic Gradient Descent**

[Rendle et al., 09]

*Estimate* current beliefs and gradient, *update* parameters accordingly.

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[88]
Training: Stochastic Gradient Descent

Estimate current beliefs and gradient, update parameters accordingly

[Rendle et al., 09]

<table>
<thead>
<tr>
<th>X-is-historian-at-Y</th>
<th>X-is-professor-at-Y</th>
<th>X-museum-at-Y</th>
<th>X-teaches-history-at-Y</th>
<th>employee(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How can we do this?

\( \text{native-of} \ 's \ \text{birthplace} \ \text{bornIn} \ \text{livesIn} \)

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & ? \\
1 & ? & ? \\
\end{array}
\]

\( \text{birthplace}(x,y) \Rightarrow \text{bornIn}(x,y) \)
Overview: Embeddings and …

- **Learning from Data**
  - [NAACL 2013]

- **Injecting Knowledge**
  - [SP 2014, NAACL 2015]

- **Extracting Knowledge**
  - [KRR 2015]