Injecting Logical Background Knowledge into Embeddings for Relation Extraction

Tim Rocktäschel, Sameer Singh and Sebastian Riedel

2015-08-29
最先端 NLP 勉強会
Sho Yokoi <yokoi@ecei.tohoku.ac.jp>

※ 図表はすべて，論文および論文著者の作成したスライドからの引用です

Table of Contents

Table of Contents

Problem: Relation Extraction

- ideitifying relations between named entities
- motivation:

Freebase is incomplete

- **Missing Facts:** `placeOfBirth` attribute is missing for 71% of the people (Dong et al., 2014)
- **Missing Entities:** Contains no information about `UCL Machine Reading Lab`
- **Missing Relations:** May contain `profAt(John Shawe-Taylor, UCL)` but not `givesLecturesAt(John Shawe-Taylor, UCL)`
- **Machine reading and reasoning to the rescue!**

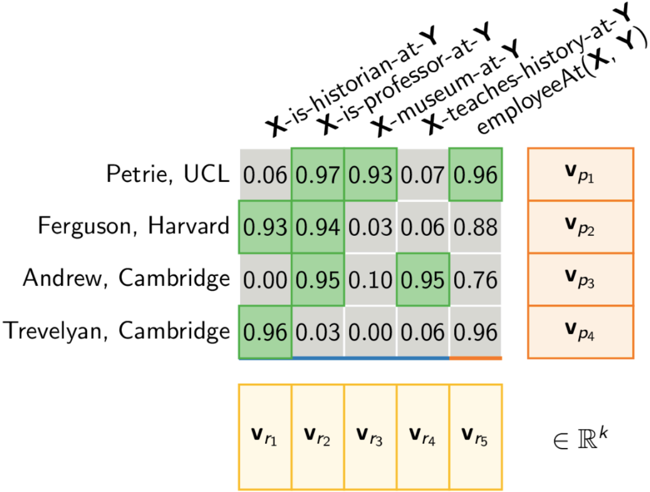
Table of Contents

[Riedel+ 13]

	X-is-historian-at-Y	X-is-professor-at-Y	X-museum-at-Y	X-teaches-history-at-Y	employeeAt(X, Y)
Petrie, UCL	?	1.00	1.00	?	1.00
Ferguson, Harvard	1.00	1.00	?	?	?
Andrew, Cambridge	?	1.00	?	1.00	?
Trevelyan, Cambridge	1.00	?	?	?	?
	Textual Patterns				Freebase

- rows: entity-pairs of interest
- cols: textual patterns + Freebase relations

[Riedel+ 13]



- matrix Factorization = low-rank embedding

Approach 1: Matrix Factorization (Low-rank Embedding)

- ✔ generalization
- ✔ tractable (computation is easy)
- ✘ hard to fix mistakes
- ✘ data sparsity

Approach 2: Logical Inference (Rule-based)

- ✔ easy to modify and improve
- ✔ data sparsity
- ✔ easy to fix mistakes
- ✘ generalization
- ✘ intractable (e.g. Markov Logic Networks)

Table of Contents

やりたいこと

- matrix factorization (low-rank embedding) に logical inference を組み込む

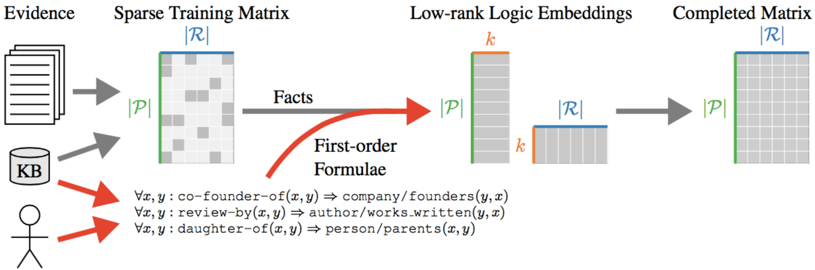


Table of Contents

Research Questions

- How to inject logical background knowledge into embedding?
- Does injection of logic formulae into the embeddings of entity-pairs and lelations provide any ***benefits***?
  - 行列分解に FOL で書かれた知識を導入することでどの程度予測精度が向上するか?
  - データの疎性に対して頑健か?
  - 「ならば」の非対称性を捉えた学習になっているか?

Table of Contents

Notation [1]

- $\langle\texttt{constant}\rangle ::= e_i, e_j \in \mathcal{E}$
  - named entities
  - e.g. Nolan
- $\langle\texttt{predicate}\rangle ::= r_m \in \mathcal{R}$
  - binary relations between the entities
  1. textual patterns (e.g. *#2-co-founder-of-#1*)
  2. Freebase relations (e.g. `company/founders`)
- $\langle\texttt{term}\rangle ::= \langle\texttt{constant}\rangle \mid \langle\texttt{varriable}\rangle$
  - using function-free first-order logic
  - no function symbols

Notation [2]

- $\langle\texttt{ground atom}\rangle ::= r_m(e_i, e_j)$
  - predicates applied to constants
  - e.g. `directorOf`(NOLAN, INTERSTELLAR)
- $\boldsymbol{w} = \left\{r_m(e_i, e_j)\right\}$: possible world
  - a set of ground atoms
  - = training data

Table of Contents

Table of Contents

MF: Matrix Factorization

- $\{0,1\}^{|\mathcal{P}|\times|\mathcal{R}|}$ の行列を，$\mathbb{R}^{|\mathcal{P}|\times k}$ の行列と $\mathbb{R}^{k\times|\mathcal{R}|}$ の行列の積に分解する
- training data
  - $\boldsymbol{w} = \left\{r_m(e_i, e_j)\right\}$: possible world
- model to learn
  - $\boldsymbol{V} = \left\{\boldsymbol{v}_{(e_i,e_j)}\right\} \cup \{\boldsymbol{v}_{r_m}\}$: model
  - $\boldsymbol{v}_{(e_i,e_j)} \in \mathbb{R}^k$: entity pair (e_i, e_j) のベクトル表現
  - $\boldsymbol{v}_{r_m} \in \mathbb{R}^k$: relation r_m のベクトル表現
- objective function

$$p(\boldsymbol{w}|\boldsymbol{V}) = \prod_{r_m(e_i,e_j)\in\boldsymbol{w}} \sigma(\boldsymbol{v}_{r_m}\cdot\boldsymbol{v}_{(e_i,e_j)}) \prod_{r_m(e_i,e_j)\notin\boldsymbol{w}} \left(1-\sigma(\boldsymbol{v}_{r_m}\cdot\boldsymbol{v}_{(e_i,e_j)})\right)$$

Table of Contents

Pre: Pre-Factorization Inference

1. トレーニングデータを，logical formulae に従って拡充
    - logical formula: e.g. $\forall x, y : r_s(x, y) \implies r_t(x, y)$
2. 行列分解

Table of Contents

Joint: Joint Optimization

- $\mathcal{F}$
  - a logical formula
- $\mathfrak{F}$
  - a training set of logical formulae

1. ground atoms (facts)
2. logical background knowledge

Objective Function

$$\min_{\boldsymbol{V}} \left(\sum_{\mathcal{F} \in \mathfrak{F}} \mathcal{L}([\mathcal{F}]) + \lambda \sum_{\boldsymbol{v} \in \boldsymbol{V}} \|\boldsymbol{v}\|_2^2 \right)$$

- $[\mathcal{F}] := p(\mathcal{F} | \boldsymbol{V})$
  - the marginal probability that the formula $\mathcal{F}$ is true under the model
  - 論文では $p(\boldsymbol{w}|\boldsymbol{V})$ と記載
- $\mathcal{L}([\mathcal{F}]) := -\log([\mathcal{F}])$

optimize using AdaGrad → We need to calculate

- the marginal probabilities $[\mathcal{F}]$
- the gradients of the losses $\partial \mathcal{L}([\mathcal{F}]) / \partial \boldsymbol{v}$

for every $\mathcal{F} \in \mathfrak{F}$

Ground Atom

- if $\mathcal{F} = r_m(e_i, e_j)$ then
- $[\mathcal{F}] = \sigma(\boldsymbol{v}_{r_m} \cdot \boldsymbol{v}_{(e_i, e_j)})$

$$\partial[\mathcal{F}]/\partial\mathbf{v}_{(e_i,e_j)} = [\mathcal{F}](1 - [\mathcal{F}])\mathbf{v}_{r_m} \quad (2)$$

$$\partial[\mathcal{F}]/\partial\mathbf{v}_{r_m} = [\mathcal{F}](1 - [\mathcal{F}])\mathbf{v}_{(e_i,e_j)} \quad (3)$$

$$\partial\mathcal{L}([\mathcal{F}])/\partial\mathbf{v}_{(e_i,e_j)} = -[\mathcal{F}]^{-1}\partial[\mathcal{F}]/\partial\mathbf{v}_{(e_i,e_j)} \quad (4)$$

$$\partial\mathcal{L}([\mathcal{F}])/\partial\mathbf{v}_{r_m} = -[\mathcal{F}]^{-1}\partial[\mathcal{F}]/\partial\mathbf{v}_{r_m}. \quad (5)$$

Logical "And" (A Set of Ground Atoms)

- $[\mathcal{A} \wedge \mathcal{B}] = [\mathcal{A}][\mathcal{B}]$
  - provided both formula concern ***non-overlapping*** sets of ground atoms
  - 重なりがある場合は互いに素な集合に分ければよい
- if $\mathcal{F} = \mathcal{F}_1 \wedge \cdots \wedge \mathcal{F}_n$ then
  - $[\mathcal{F}] = \prod_{\mathcal{F}_i \in \mathcal{F}} [\mathcal{F}_i]$
  - $\mathcal{L}([\mathcal{F}]) = \sum_{\mathcal{F}_i \in \mathcal{F}} \mathcal{L}([\mathcal{F}_i])$
- 以後,
  - 論理式 = 論理式が満たすべき ground atoms の集合
  - モデル $\boldsymbol{V}$ に対する論理式の条件付確率 = 論理式をなす ground atom の条件付き確率（$\sigma(\boldsymbol{v}_{r_m} \cdot \boldsymbol{v}_{(e_i,e_j)})$）の総積

…と考える

Complex Logical Formulae

Negation

- $[\neg\mathcal{A}] = 1 - [\mathcal{A}]$

Other Logical Connectives （$\wedge, \neg$ を用いて）

- $[\mathcal{A} \vee \mathcal{B}] = [\neg(\mathcal{A} \wedge \mathcal{B})] = [\mathcal{A}] + [\mathcal{B}] - [\mathcal{A}][\mathcal{B}]$
- $[\mathcal{A} \implies \mathcal{B}] = [\neg\mathcal{A} \vee \mathcal{B}] = \cdots = [\mathcal{A}]([\mathcal{B}] - 1) + 1$

Universal Quantifier

- if $\mathcal{F} = \forall x, y \in \mathcal{E} : \mathcal{G}(x, y)$ then
- $[F] = [\forall x, y \in \mathcal{E} : \mathcal{G}(x, y)] = [\wedge_{x,y\in\mathcal{E}} \mathcal{G}(x, y)]$

More Complex Logical Formulae
以上を再帰的にに適用

“Implications”

- if $\mathcal{F} = \forall x, y \in \mathcal{E} : r_s(x, y) \implies r_t(x, y)$
- $\mathcal{F}_{ij} := r_s(e_i, e_j) \implies r_t(e_i, e_j)$
- $[\mathcal{F}] = \prod_{(e_i, e_j) \in \mathcal{P}} [F_{ij}]$
- $\mathcal{L}([\mathcal{F}]) = \sum_{(e_i, e_j) \in \mathcal{P}} \mathcal{L}([F_{ij}])$

$$[\mathcal{F}_{ij}] = [r_s(e_i, e_j)]\left([r_t(e_i, e_j)] - 1\right) + 1 \tag{6}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}([\mathcal{F}_{ij}])}{\partial \mathbf{v}_{r_s}} &= -[\mathcal{F}_{ij}]^{-1}\left([r_t(e_i, e_j)] - 1\right)\frac{\partial [r_s(e_i, e_j)]}{\partial \mathbf{v}_{r_s}} \\
\frac{\partial \mathcal{L}([\mathcal{F}_{ij}])}{\partial \mathbf{v}_{r_t}} &= -[\mathcal{F}_{ij}]^{-1}[r_s(e_i, e_j)]\frac{\partial [r_t(e_i, e_j)]}{\partial \mathbf{v}_{r_t}}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}([\mathcal{F}_{ij}])}{\partial \mathbf{v}_{e_i, e_j}} &= -[\mathcal{F}_{ij}]^{-1}\left([r_t(e_i, e_j)] - 1\right)\frac{\partial [r_s(e_i, e_j)]}{\partial \mathbf{v}_{e_i, e_j}} \\
&\quad - [\mathcal{F}_{ij}]^{-1}[r_s(e_i, e_j)]\frac{\partial [r_t(e_i, e_j)]}{\partial \mathbf{v}_{e_i, e_j}}.
\end{aligned} \tag{8}$$

Table of Contents

Table of Contents

Data

knowledge base completion of Freebase [Bollacker+ 08] with textual data from the NYTimes corpus [Sandhaus 08]

- $\mathcal{R}$
    - 151 Freebase relations
    - 3,960 textual (surface) patterns
- $\mathcal{P} \subseteq \mathcal{E} \times \mathcal{E}$
    - 41,913 entity-pairs of interest
- $\mathcal{R} \times \mathcal{P}$
    - 118,781 training facts
    - including 7,293 Freebase relations (alignments between entity-pairs and Freebase relations)

Table of Contents

Extract Background Knowledge

1. 全教師データで行列分解
2. すべての $(r_s, r_t) \in \mathcal{R} \times \mathcal{R}$（ただし r_t は Freebase relation）について $[\forall(e_i, e_j) \in \mathcal{P} r_s(e_i, e_j) \implies r_s(e_i, e_j)]$ を計算
3. 上位 100 件から手動で 36 件をフィルタリング

Formula	Score
$\forall x, y$: #2-unit-of-#1$(x, y) \Rightarrow$ org/parent/child(x, y)	0.97
$\forall x, y$: #2-city-of-#1$(x, y) \Rightarrow$ location/containedby(x, y)	0.97
$\forall x, y$: #2-minister-#1$(x, y) \Rightarrow$ person/nationality(x, y)	0.97
$\forall x, y$: #2-executive-#1$(x, y) \Rightarrow$ person/company(x, y)	0.96
$\forall x, y$: #2-co-founder-of-#1$(x, y) \Rightarrow$ company/founders(y, x)	0.96

Table 1: **Sample Extracted Formulae:** Top implica-

Metric

(weighted) mean average precision (***MAP***, ***wMAP***) on manually annotated predictions for Freebase relations

Table of Contents

1. Zero-shot Relation Learning

- Freebase relations $\subseteq \mathcal{R} \times \mathcal{E} \times \mathcal{E}$ をすべて隠して，当てる
- background knowledge は使う

Relation	#	MF	Inf	Post	Pre	Joint
person/company	102	0.07	0.03	0.15	0.31	**0.35**
location/containedby	72	0.03	0.06	0.14	0.22	**0.31**
author/works_written	27	0.02	0.05	0.18	**0.31**	0.27
person/nationality	25	0.01	*0.19*	0.09	0.15	*0.19*
parent/child	19	0.01	0.01	0.48	0.66	**0.75**
person/place_of_birth	18	0.01	0.43	0.40	0.56	**0.59**
person/place_of_death	18	0.01	0.24	0.23	**0.27**	0.23
neighborhood/neighborhood_of	11	0.00	0.00	0.60	0.63	**0.65**
person/parents	6	0.00	0.17	0.19	0.37	**0.65**
company/founders	4	0.00	0.25	0.13	0.37	**0.77**
film/directed_by	2	0.00	0.50	0.50	0.36	**0.51**
film/produced_by	1	0.00	*1.00*	*1.00*	*1.00*	*1.00*
MAP		0.01	0.23	0.34	0.43	**0.52**
Weighted MAP		0.03	0.10	0.21	0.33	**0.38**

Table 2: **Zero-shot Relation Learning:** Average and

2. Relations with Few Distant Labels

- Freebase relations $\subseteq \mathcal{R} \times \mathcal{E} \times \mathcal{E}$ を一部隠して，当てる
- トレーニングデータとして使う Freebase relations (distant labels) の割合を変化させる

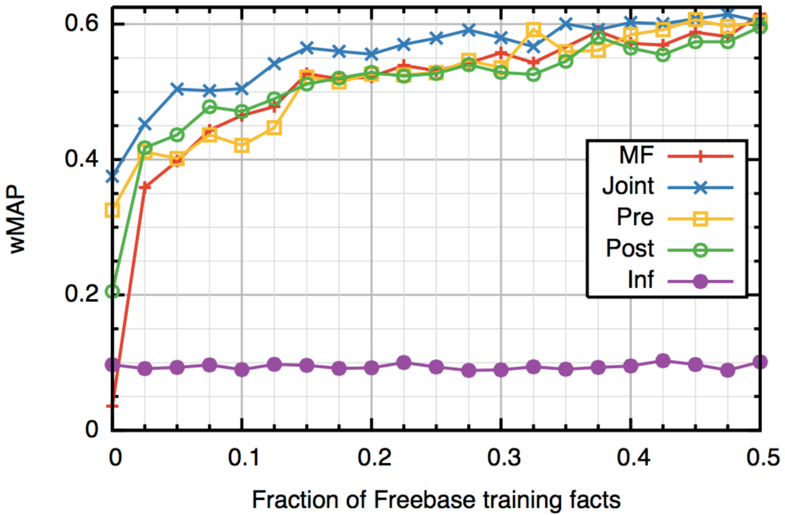


Figure 2: **Relations with Few Distant Labels:**

3. Comparison on Complete Data

- Freebase relations $\subseteq \mathcal{R} \times \mathcal{E} \times \mathcal{E}$ をすべて使う
- 先行研究 [Riedel+ 13] の matrix factorization model “F” との性能比較
- ただし [Riedel+ 13] 中の「もっとも良い」モデルである “NF” や “NFE” とは比較されていない

4. Analysis of Asymmetry in the Predictions

- 学習に利用した logical formulae はすべて $\forall x, y : r_s(x, y) \implies r_t(x, y)$ の形
- 学習されたモデルは " $\implies$ " の非対称性を捉えられているか？
- Joint は捉えられているように見える

	MF	Pre	Joint
$[\forall x, y : r_s(x, y) \implies r_t(x, y)]$	0.94	0.96	0.97
$[\forall x, y : r_t(x, y) \implies r_s(x, y)]$	0.81	0.83	***0.49***