Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach

Liyuan Liu*, Xiang Ren*, Qi Zhu*, Huan Gui*, Shi Zhi*, Heng Ji* and Jiawei Han*

*University of Illinois at Urbana-Champaign, Urbana, IL, USA
○Computer Science Department, Rensselaer Polytechnic Institute, USA

Presented by: Mao-Chuang Yeh
The structure

• The Goal
• Previous works
• Heterogeneous supervision
  • Supervision conflicts
  • True Label Discovery

• REHESSION Framework
  • Relation Mention Representation
  • True Label Discovery component
  • Relation Extraction component

• Experiment
• Contribution
Relation Extraction

• Goal: find the entity relation from unstructured text

  • relation mention
  entity pair
  sentence / context
  
  \begin{tabular}{|c|c|}
  \hline
  \textbf{Hussein} & \textbf{Amman} \\
  \hline
  Hussein was born in Amman on 14 November 1935. \\
  \hline
  \end{tabular}

  \begin{itemize}
  \item multi-class classification
  \item relation types of interest
  \end{itemize}

  \begin{itemize}
  \item Born-in
  \item President-of
  \item Died-in
  \item Parents-of
  \item None
  \end{itemize}
Previous Work

• Supervised Learning:
  
  • Multi-class classification

Limited by human annotation

Limited, need domain experts
costly and time-consuming

......
Previous Work

• Bootstrap learning:

  • Start with a set of seed patterns / annotations, iteratively generate more

  • Suffers from semantic shift
Goal: conduct relation extractor learning annotations from Heterogeneous supervision at context level.
Heterogeneous supervision

(Using annotations from) heterogeneous information sources

1. knowledge base
2. domain specific pattern (domain heuristics).
Supervision conflicts

- context and labeling function:

\[ D \]

- Robert Newton "Bob" Ford was an American outlaw best known for killing his gang leader Jesse James (\( e_1 \)) in Missouri (\( e_2 \)).

- Gofraid (\( e_1 \)) died in 989, said to be killed in Dal Riata (\( e_2 \)).

- Hussein (\( e_1 \)) was born in Amman (\( e_2 \)) on 14 November 1935.

\[ \Lambda \]

- \( \lambda_1 \) return \textit{born in} for \(<e_1, e_2, s>\) if \text{BornIn}(e_1, e_2) in KB

- \( \lambda_2 \) return \textit{died in} for \(<e_1, e_2, s>\) if \text{DiedIn}(e_1, e_2) in KB

- \( \lambda_3 \) return \textit{born in} for \(<e_1, e_2, s>\) if match("* born in *", s)

- \( \lambda_4 \) return \textit{died in} for \(<e_1, e_2, s>\) if match("* killed in *", s)
Supervision conflicts

**Source consistency assumption:** a source is likely to provide true information with the same probability for all instances. (Ratner et al., 2016)

However:

labeling functions mistakes by certain “error routine”;
Different from human annotator.

**Proficient subset:** some subset is reliable than others for a labeling function (Varma et al., 2016)
True Label discovery

1. Identify and trust labeling function on proficient subsets

2. Context awareness (at sentence level): use context information to improve accuracy
A Representation Learning Approach

Heterogeneous Supervision generation

Relation Mention Representation

Relation Extraction

True Label Discovery

Relation Extraction View

interact through context representation

proficient subset

represention of proficient subset

represention of relation type died_in

representation of relation mention

represention of relation type born_in

Hussein (e1) was born in Amman (e2) on 14 November 1935.

Robert Newton "Bob" Ford was an American outlaw best known for killing his gang leader Jesse James (e1) in Missouri (e2).

Gofraid (e1) died in 989, said to be killed in Dal Riata (e2).

training Relation Extraction model

training True Label Discovery model

λ1, λ2, λ3, λ4, true

c1

c2

c3

proficient subset
Problem Definition

• For POS-tagged corpus D, we refer its relation mentions as
  \[ C = \{c_i = (e_{i,1}, e_{i,2}, d), \forall d \in D\} \]

• Goal: annotate entity mentions with relation types of interest
  \( \mathcal{R} = \{r_1, \ldots, r_K\} \) or None

• Labeling functions: \( \Lambda = \{\lambda_1, \ldots, \lambda_M\} \)

• Annotation:
  \[ \mathcal{O} = \{o_{c,i}|\lambda_i \text{ generate annotation } o_{c,i} \text{ for } c \in C\} \]
REHESSION Framework (except Extraction and Representation of Text Features)
# Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_c$</td>
<td>$c$’s text features set, where $c \in \mathcal{C}$</td>
</tr>
<tr>
<td>$v_i$</td>
<td>text feature embedding for $f_i \in \mathcal{F}$</td>
</tr>
<tr>
<td>$z_c$</td>
<td>relation mention embedding for $c \in \mathcal{C}$</td>
</tr>
<tr>
<td>$l_i$</td>
<td>embedding for $\lambda_i$’s proficient subset, $\lambda_i \in \Lambda$</td>
</tr>
<tr>
<td>$o_{c,i}$</td>
<td>annotation for $c$, generated by labeling function $\lambda_i$</td>
</tr>
<tr>
<td>$o_{c}^{*}$</td>
<td>underlying true label for $c$</td>
</tr>
<tr>
<td>$\rho_{c,i}$</td>
<td>identify whether $o_{c,i}$ is correct</td>
</tr>
<tr>
<td>$S_i$</td>
<td>the proficient subset of labeling function $\lambda_i$</td>
</tr>
<tr>
<td>$s_{c,i}$</td>
<td>identify whether $c$ belongs to $\lambda_i$’s proficient subset</td>
</tr>
<tr>
<td>$t_i$</td>
<td>relation type embedding for $r_i \in \mathcal{R}$</td>
</tr>
</tbody>
</table>

**Table 1: Notation Table.**
REHESSION Framework

1. **Text Feature Representation:** After being extracted from context, text features are embedded in a low dimension space by Representation learning;

2. **Relation Mention Representation:** Text feature embeddings are utilized to calculate Relation Mention embeddings;

3. **True Label Discovery:** with relation mention embeddings, true labels are inferred by calculating labeling functions’ reliabilities in a context-aware manner;

4. **Modeling Relation Type:** Inferred true labels would ‘supervise’ all components to learn model parameters.
Text Feature Extraction

We adopted texture features, POS-tagging and brown clustering to extract features.

C3: Hussein was born in Amman on 14 November 1935.
Relation Mention Representation

• Text Feature Extraction

• Text Feature Representation

• Relation Mention Representation

\[ \text{Mapping from Text Embedding to Relation Mention Embedding:} \]
\[ \text{tanh}(W \cdot \frac{1}{|f|} \sum_{f_i \in f} v_i) \]

\[ z_c \in \mathbb{R}^{n_z} \]
Text Feature Representation

• Leverage features’ co-occurrence information to learn the representation, and help the model generalize better.

• Loss function of this part:

\[ J_E = \sum_{c \in C_{f_i, f_j} \in E_c} \left( \log \sigma(v_i^f v_j^f) - \sum_{k=1}^{V} \mathbb{E}_{f_{k'} \sim P} [\log \sigma(-v_i^f v_{k'}^f)] \right) \]

co-occurrence here refers to features occur in the same relation mention

Negative sampling
Relation Mention Representation

• Here, we adopted the bag-of-features average, then do linear mapping and nonlinear tanh on it to different semantic space.

\[ z_c = g(f_c) = \tanh\left( W \cdot \frac{1}{|f_c|} \sum_{f_i \in f_c} v_i \right) \]
Conflicts among Heterogeneous Supervision

• Truth Discovery:

  • Some sources (labeling functions) would be more reliable than others
  • Refer the reliability of different sources and the true label at the same time
  • Context awareness: A source is likely to provide true information with the same probability for instances **with similar context**.

  • Source Consistency Assumption: a source is likely to provide true information with the same probability for all instances.

Heterogeneous Supervision for Relation Extraction

• Relation Extraction:
  • Matching context with proper relation type

• Heterogeneous Supervision:
  • Refer true labels in a context-aware manner
True label discovery

Relation Mention Representation

Heterogeneous Supervision generation

Relation Extraction

True Label Discovery
True label discovery

- Describing the correctness of Heterogeneous Supervision

\[ \rho_{c,i} = \delta(o_{c,i} == o^*_c) \]

 observed annotation

correctness of annotation \( o_{c,i} \)

underlying true label

whether \( c \) belongs to the proficient subset of \( l_i \)

Representation of relation mention

Representation of labeling function

\( Z_c \)
True label discovery

• correctness of annotation: \( \rho_{c,i} = \delta(o_{c,i} = o_{c}^*) \).
• prob. in proficient subset: \( p(s_{c,i} = 1|z_{c}, l_i) = p(c \in S_i) = \sigma(z_{c}^T l_i) \).
• assume: \( p(\rho_{c,i} = 1|s_{c,i} = 1) = \phi_1 \) \( p(\rho_{c,i} = 1|s_{c,i} = 0) = \phi_0 \).

• Prob of correct annotation:

\[
p(\rho_{c,i} = 1) = p(\rho_{c,i} = 1|s_{c,i} = 1) \cdot p(s_{c,i} = 1) + p(\rho_{c,i} = 1|s_{c,i} = 0) \cdot p(s_{c,i} = 0)
\]

• The true label loss:

\[
J_T = \sum_{o_{c,i} \in O} \log(\sigma(z_{c}^T l_i))\phi_1^\delta(o_{c,i} = o_{c}^*) (1 - \phi_1)^\delta(o_{c,i} \neq o_{c}^*) \\
+ (1 - \sigma(z_{c}^T l_i))\phi_0^\delta(o_{c,i} = o_{c}^*) (1 - \phi_0)^\delta(o_{c,i} \neq o_{c}^*)
\]
Relation Extraction

Relation Mention
Representation

Heterogeneous Supervision generation

Relation Extraction View

training Truth Discovery model

λ]: return born in for <e1, e2, s> if BornIn(e1, e2) in KB
λ2] return died in for <e1, e2, s> if DiedIn(e1, e2) in KB
λ3] return born in for <e1, e2, s> if match(" born in ", s)
λ4] return died in for <e1, e2, s> if match(" killed in ", s)

proficient subset

representation of proficient subset
representation of relation type died in
representation of relation mention
representation of relation type born in

inter 'true' label

 Relation Extraction View

training Relation Extraction model

Truth Discovery View

λ]: [λ2] [λ3] [λ4]
Relation Extraction (context aware)

• Adopts soft-max as the relation extractor:

\[ p(r_i | z_c) = \frac{\exp(z_c^T t_i)}{\sum_{r_j \in R \cup \{\text{None}\}} \exp(z_c^T t_j)} \]

• Loss function: KL-Divergence:

\[ J_R = - \sum_{c \in C_l} KL(p(., z_c) || p(., o_c^*)) \]

• true label distribution

\[ p(r_i | o_c^*) = \begin{cases} 1 & r_i = o_c^* \\ 0 & r_i \neq o_c^* \end{cases} \]
A Representation Learning Approach

Heterogeneous Supervision generation

Relation Mention Representation

True Label Discovery

Relation Extraction
Model Learning

• Joint optimize three components

\[
\begin{align*}
\min_{W,v,v^*,l,t,o^*} \mathcal{J} &= -\mathcal{J}_R - \lambda_1 \mathcal{J}_E - \lambda_2 \mathcal{J}_T \\
\text{s.t. } \forall c \in \mathcal{C}_l, o_c^* &= \arg\max_{o_c^*} \mathcal{J}_T, z_c = g(f_c)
\end{align*}
\]
Two Data Sets

**NYT** (Riedel et al., 2010):

*a news corpus sampled from 294k 1989-2007 New York Times news articles. 1.18M sentences, 395 of them are annotated by authors of (Hoffmann et al., 2011) and used as test data*

**Wiki-KBP**:

1.5M sentences sampled from 780k Wikipedia articles as training corpus (Ling and Weld, 2012),

the 2k sentences in test set manually annotated in 2013 KBP slot filling assessment results (Ellis et al., 2012)
Number of relation types

Table 4: Number of labeling functions and the relation types they can annotated w.r.t. two kinds of information
Number of None type

<table>
<thead>
<tr>
<th>Datasets</th>
<th>NYT</th>
<th>Wiki-KBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of None in Training</td>
<td>0.6717</td>
<td>0.5552</td>
</tr>
<tr>
<td>% of None in Test</td>
<td>0.8972</td>
<td>0.8532</td>
</tr>
</tbody>
</table>

Table 3: Proportion of None in Training/Test Set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Wiki-KBP</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of RM</td>
<td>225977</td>
<td>530767</td>
</tr>
<tr>
<td>RM annotated as None</td>
<td>100521</td>
<td>356497</td>
</tr>
<tr>
<td>RM with conflicts</td>
<td>32008</td>
<td>58198</td>
</tr>
<tr>
<td>Conflicts involving None</td>
<td>30559</td>
<td>38756</td>
</tr>
</tbody>
</table>

Table 6: Number of relation mentions (RM), relation mentions annotated as None, relation mentions with conflicting annotations and conflicts involving None
## Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Relation Extraction</th>
<th>Relation Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NYT</td>
<td>Wiki-KBP</td>
</tr>
<tr>
<td><strong>Prec</strong></td>
<td><strong>Rec</strong></td>
<td><strong>F1</strong></td>
</tr>
<tr>
<td>NL+FIGER</td>
<td>0.2364</td>
<td>0.2914</td>
</tr>
<tr>
<td>NL+BFK</td>
<td>0.1520</td>
<td>0.0508</td>
</tr>
<tr>
<td>NL+DSL</td>
<td>0.4150</td>
<td>0.5414</td>
</tr>
<tr>
<td>NL+MultiR</td>
<td>0.5196</td>
<td>0.2755</td>
</tr>
<tr>
<td>NL+FCM</td>
<td>0.4170</td>
<td>0.2890</td>
</tr>
<tr>
<td>NL+CoType-RM</td>
<td>0.3967</td>
<td>0.4049</td>
</tr>
<tr>
<td>TD+FIGER</td>
<td>0.3664</td>
<td>0.3350</td>
</tr>
<tr>
<td>TD+BFK</td>
<td>0.1011</td>
<td>0.0504</td>
</tr>
<tr>
<td>TD+DSL</td>
<td>0.3704</td>
<td>0.5025</td>
</tr>
<tr>
<td>TD+MultiR</td>
<td><strong>0.5232</strong></td>
<td>0.2736</td>
</tr>
<tr>
<td>TD+FCM</td>
<td>0.3394</td>
<td>0.3325</td>
</tr>
<tr>
<td>TD+CoType-RM</td>
<td>0.4516</td>
<td>0.3499</td>
</tr>
<tr>
<td>REHESsion</td>
<td>0.4122</td>
<td><strong>0.5726</strong></td>
</tr>
</tbody>
</table>

Table 6: Performance comparison of relation extraction and relation classification
Experiments

- Effectiveness of proposed true label discovery component:
  - Ori: with proposed context-aware true label discovery component
  - LD: with Investment (compared true label discovery model)

<table>
<thead>
<tr>
<th>Dataset &amp; Method</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki-KBP</td>
<td>Ori</td>
<td>0.3677</td>
<td>0.4933</td>
<td>0.4208</td>
</tr>
<tr>
<td></td>
<td>TD</td>
<td>0.3032</td>
<td>0.5279</td>
<td>0.3850</td>
</tr>
<tr>
<td>NYT</td>
<td>Ori</td>
<td>0.4122</td>
<td>0.5726</td>
<td>0.4792</td>
</tr>
<tr>
<td></td>
<td>TD</td>
<td>0.3758</td>
<td>0.4887</td>
<td>0.4239</td>
</tr>
</tbody>
</table>

Table 7: Comparison between REHESSION (Ori) and REHESSION-TD (TD) on relation extraction and relation classification
Case Study

<table>
<thead>
<tr>
<th>Relation Mention</th>
<th>REHESION</th>
<th>Investment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann Demeulemeester (born 1959, Waregem, Belgium) is a ...</td>
<td>born-in</td>
<td>None</td>
</tr>
<tr>
<td>Raila Odinga was born at ..., in Maseno, Kisumu District, ...</td>
<td>born-in</td>
<td>None</td>
</tr>
<tr>
<td>Ann Demeulemeester (elected 1959, Waregem, Belgium) is a ...</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Raila Odinga was examined at ..., in Maseno, Kisumu District, ...</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 8: Example output of true label discovery. The first two relation mentions come from Wiki-KBP, and their annotations are \{born-in, None\}. The last two are created by replacing key words of the first two. Key words are marked as bold and entity mentions are marked as Italicics.
Summary

• Deal with heterogeneous supervisions

• Go beyond the “source consistency assumption” in prior works and leverage context-aware embeddings to induce proficient subsets

• bridges true label discovery and relation extraction with context representation