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

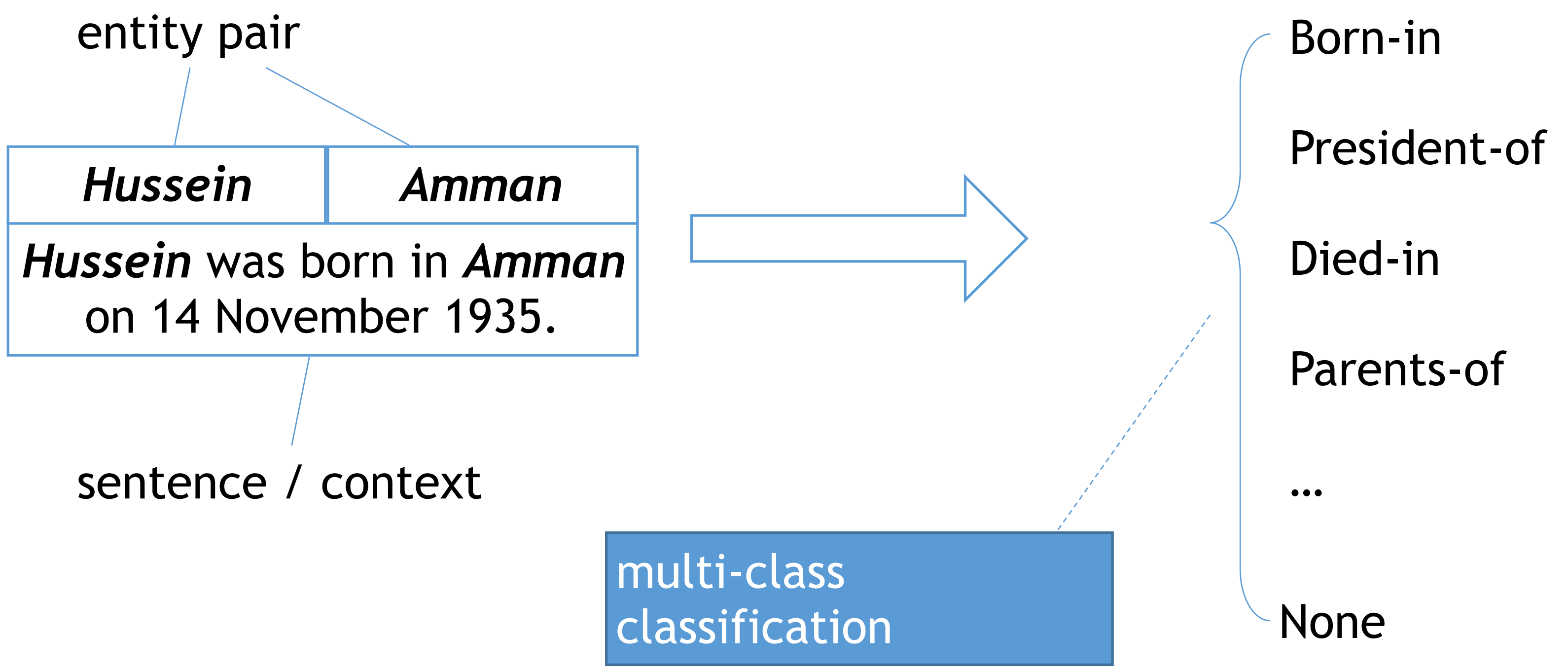
Presented by : Mao-Chuang Yeh

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
  - <u>relation mention</u>
  - <u>relation types of interest</u>



Previous Work

- Supervised Learning:
  - Multi-class classification

$\boxed{\text{Limited by human annotation}}$

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

Mintz et al. “Distant supervision for relation extraction without labeled data”, ACL 2009

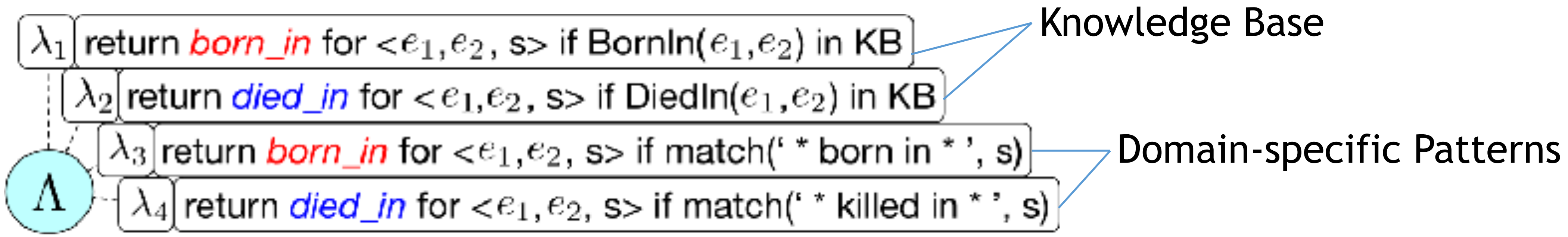
REHESSION

Goal : conduct relation extractor learning annotations from Heterogeneous supervision at context level.

Heterogeneous supervision

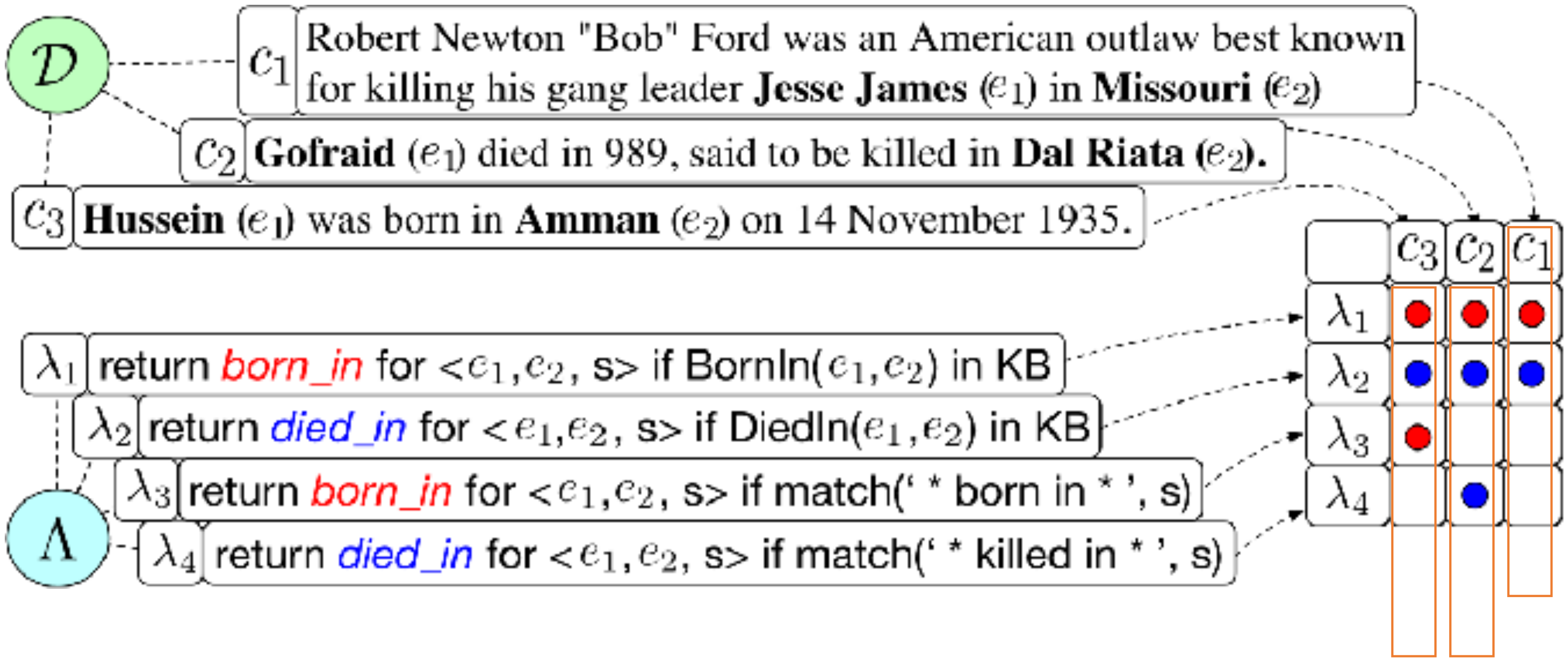
(Using annotations from)
heterogeneous information sources

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



Supervision conflicts

- context and labeling function:



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.

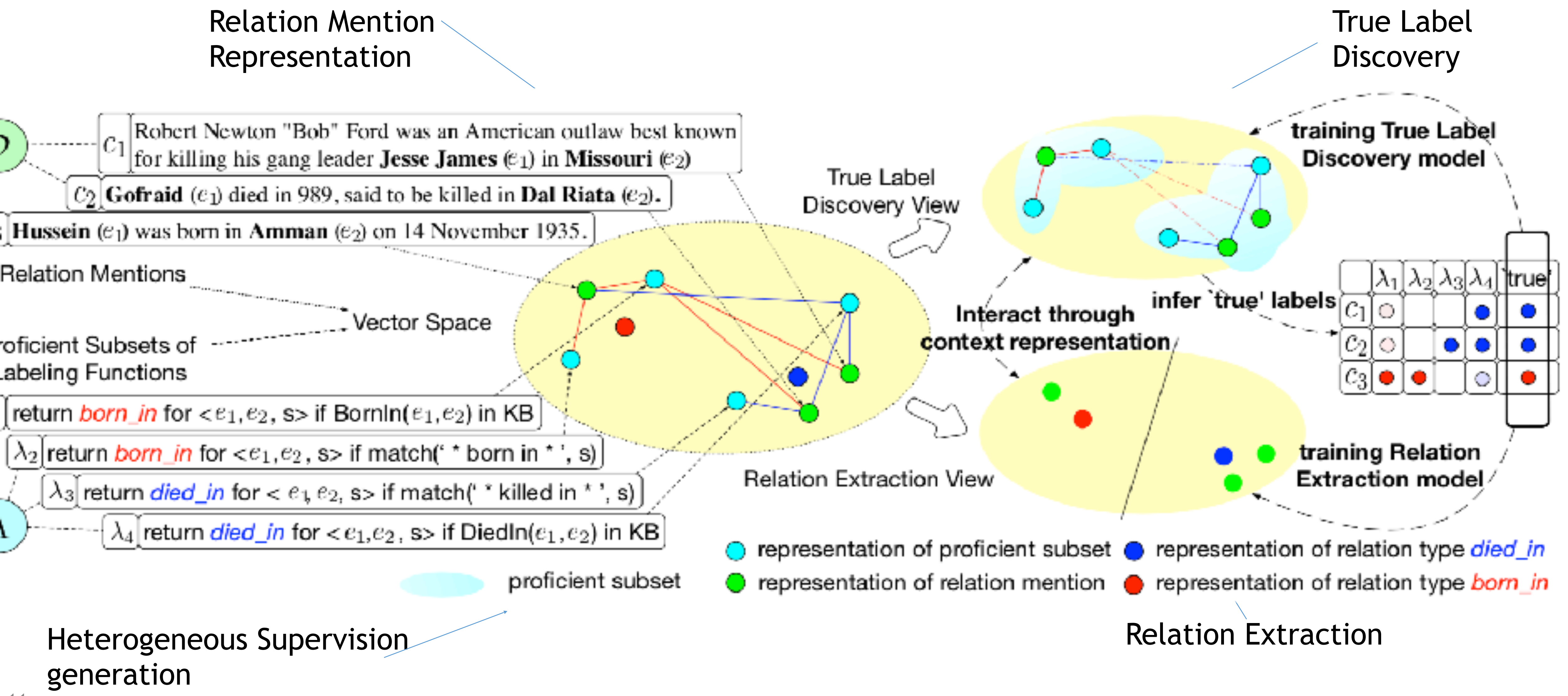
Proficient subset: some subset is reliable than others for a labeling function

(Varma et al., 2016)

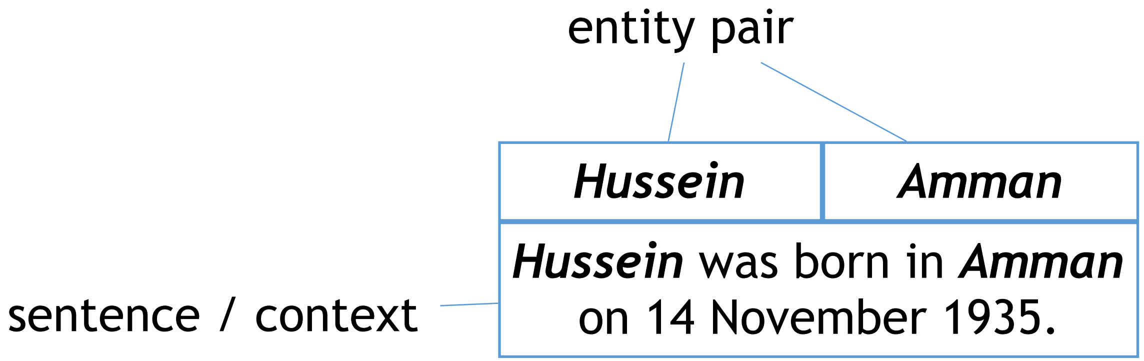
True Label discovery

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



Problem Definition



- For POS-tagged corpus D, we refer its relation mentions as

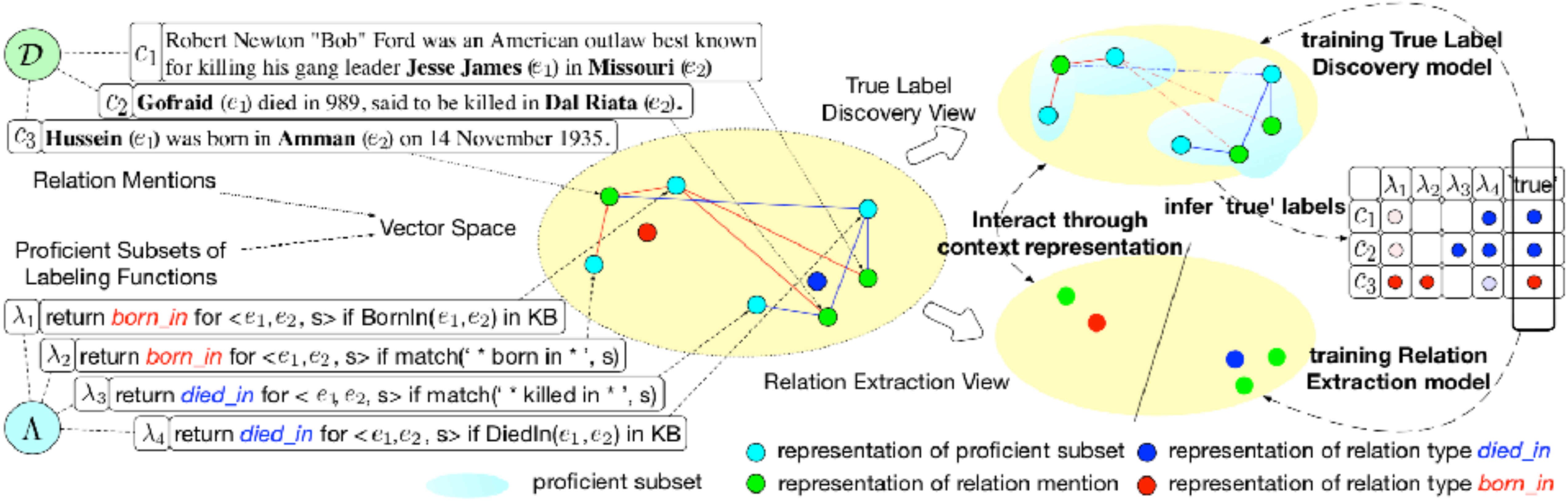
$$\mathcal{C} = \{c_i = (e_{i,1}, e_{i,2}, d), \forall d \in \mathcal{D}\}$$

- 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:

$$\mathcal{O} = \{o_{c,i} | \lambda_i \text{ generate annotation } o_{c,i} \text{ for } c \in \mathcal{C}\}$$

REHESSION Framework
(except Extraction and Representation of Text Features)



Notations

$\mathbf{f}_c$	c's text features set, where $c \in \mathcal{C}$
$\mathbf{v}_i$	text feature embedding for $f_i \in \mathcal{F}$
$\mathbf{z}_c$	relation mention embedding for $c \in \mathcal{C}$
$\mathbf{l}_i$	embedding for λ_i's proficient subset, $\lambda_i \in \Lambda$
$o_{c,i}$	annotation for c, generated by labeling function λ_i
o_c^*	underlying true label for c
$\rho_{c,i}$	identify whether $o_{c,i}$ is correct
$\mathcal{S}_i$	the proficient subset of labeling function λ_i
$s_{c,i}$	identify whether c belongs to λ_i's proficient subset
$\mathbf{t}_i$	relation type embedding for $r_i \in \mathcal{R}$

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;

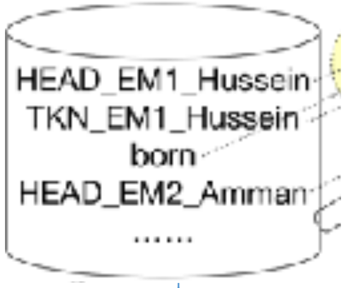
3. True Label Discovery: with relation mention embeddings, true labels are inferred by calculating labeling functions' reliabilities in a context-aware manner;

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 **<u>brown clustering</u>** to extract features

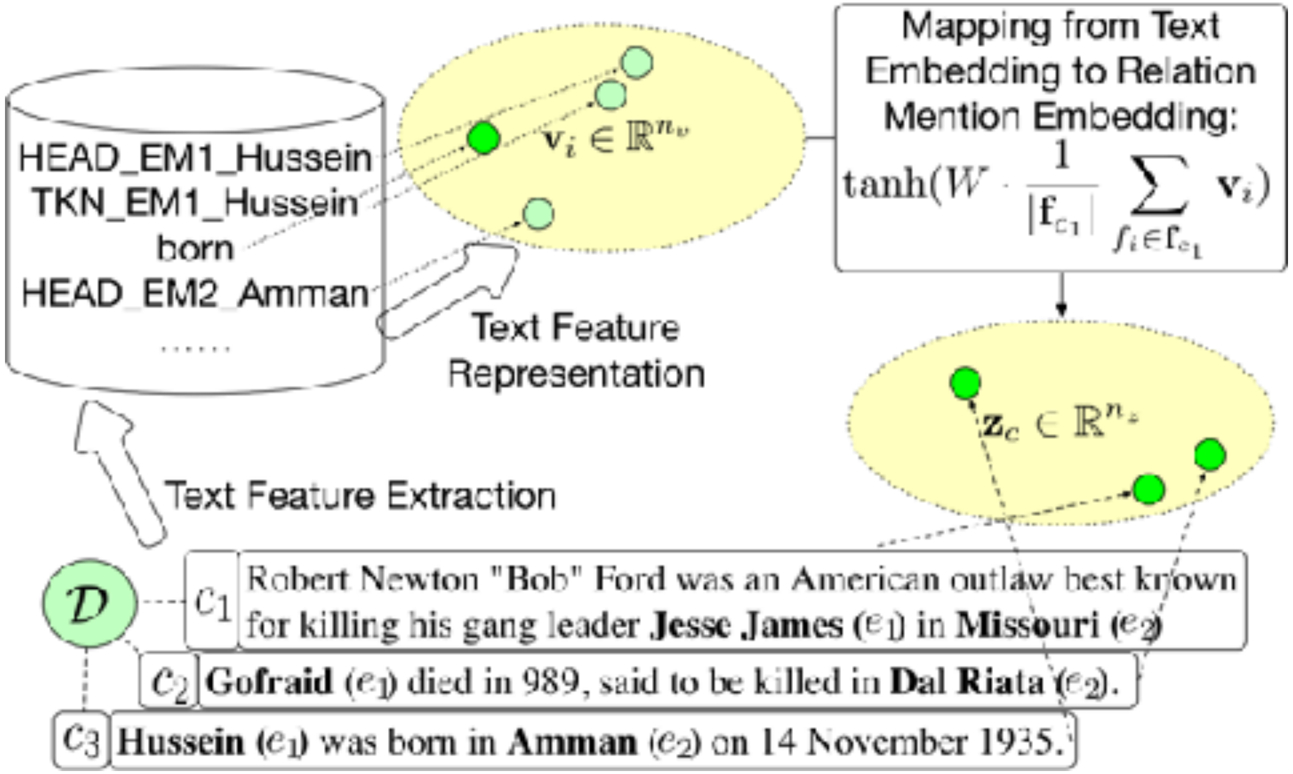
C3: ***Hussein*** was born in ***Amman*** on 14 November 1935



Feature	Description	Example
Entity mention (EM) head	Syntactic head token of each entity mention	"*HEAD_EM1_Hussein*", ...
Entity Mention Token	Tokens in each entity mention	"*TKN_EM1_Hussein*", ...
Tokens between two EMs	Tokens between two EMs	"*was*", "*born*", "*in*"
Part-of-speech (POS) tag	POS tags of tokens between two EMs	"*VBD*", "*VBN*", "*IN*"
Collocations	Bigrams in left/right 3-word window of each EM	"*Hussein was*", "*in Amman*"
Entity mention order	Whether EM 1 is before EM 2	"*EM1_BEFORE_EM2*"
Entity mention distance	Number of tokens between the two EMs	"*EM_DISTANCE_3*"
Body entity mentions numbers	Number of EMs between the two EMs	"*EM_NUMBER_0*"
Entity mention context	Unigrams before and after each EM	"*EM_AFTER_was*", ...
Brown cluster (learned on $\mathcal{D}$)	Brown cluster ID for each token	"*BROWN_010011001*", ...

Relation Mention Representation

- Text Feature Extraction
- Text Feature Representation
- Relation Mention Representatio



Text Feature Representation

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

- Loss function of this part:

$$\mathcal{J}_E = \sum_{\substack{c \in \mathcal{C}_l \\ f_i, f_j \in \mathbf{f}_c}} \left(\log \sigma(\boxed{\mathbf{v}_i^T} \mathbf{v}_j^*) - \boxed{\sum_{k=1}^{V} \mathbb{E}_{f_{k'} \sim \hat{P}} [\log \sigma(-\mathbf{v}_i^T \mathbf{v}_{k'}^*)]} \right)$$

Feature embedding for feature f_i

Feature set of c

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.

$$\mathbf{z}_c = g(\mathbf{f}_c) = \tanh\Big(W \cdot \frac{1}{|\mathbf{f}_c|} \sum_{f_i \in \mathbf{f}_c} \mathbf{v}_i\Big)$$

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.

Li, Yaliang, et al. "A survey on truth discovery." Acm Sigkdd Explorations Newsletter 17.2 (2016): 1-16.

Heterogeneous Supervision for Relation Extraction

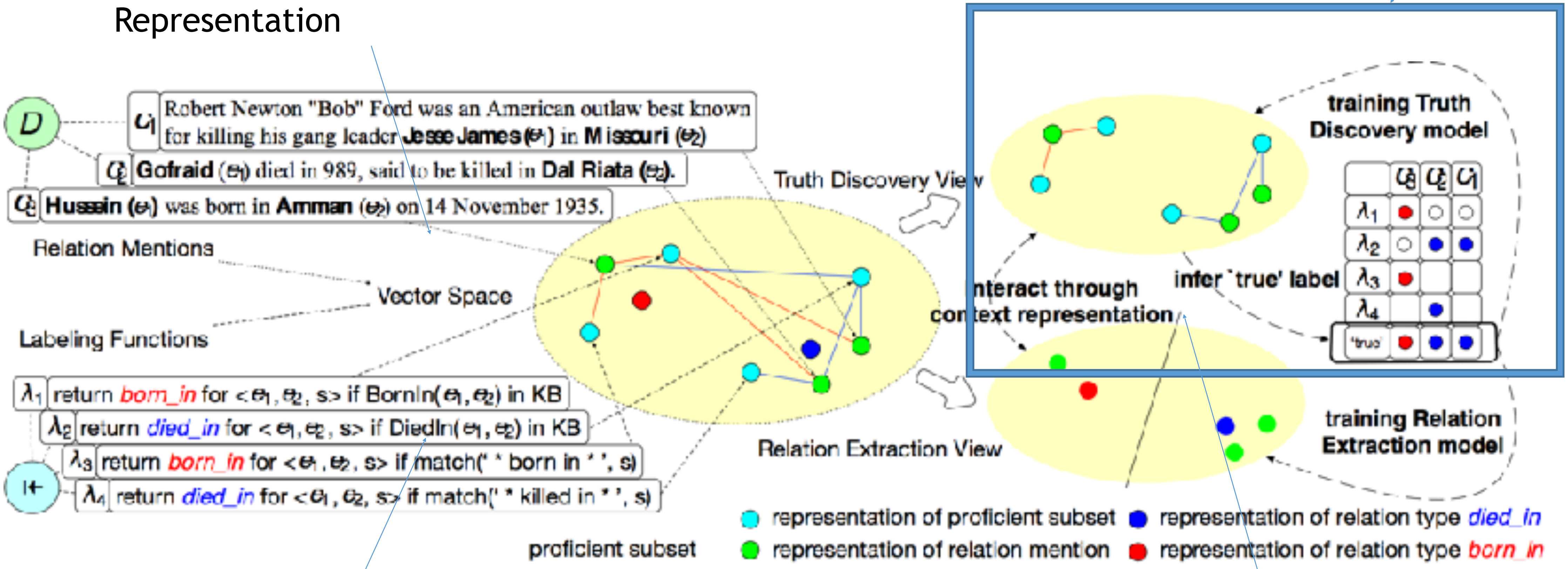
- Relation Extraction:
  - Matching ***context*** with proper relation type
- Heterogeneous Supervision:
  - Refer true labels in a ***context***-aware manner

context

True label discovery

True Label Discovery

Relation Mention Representation

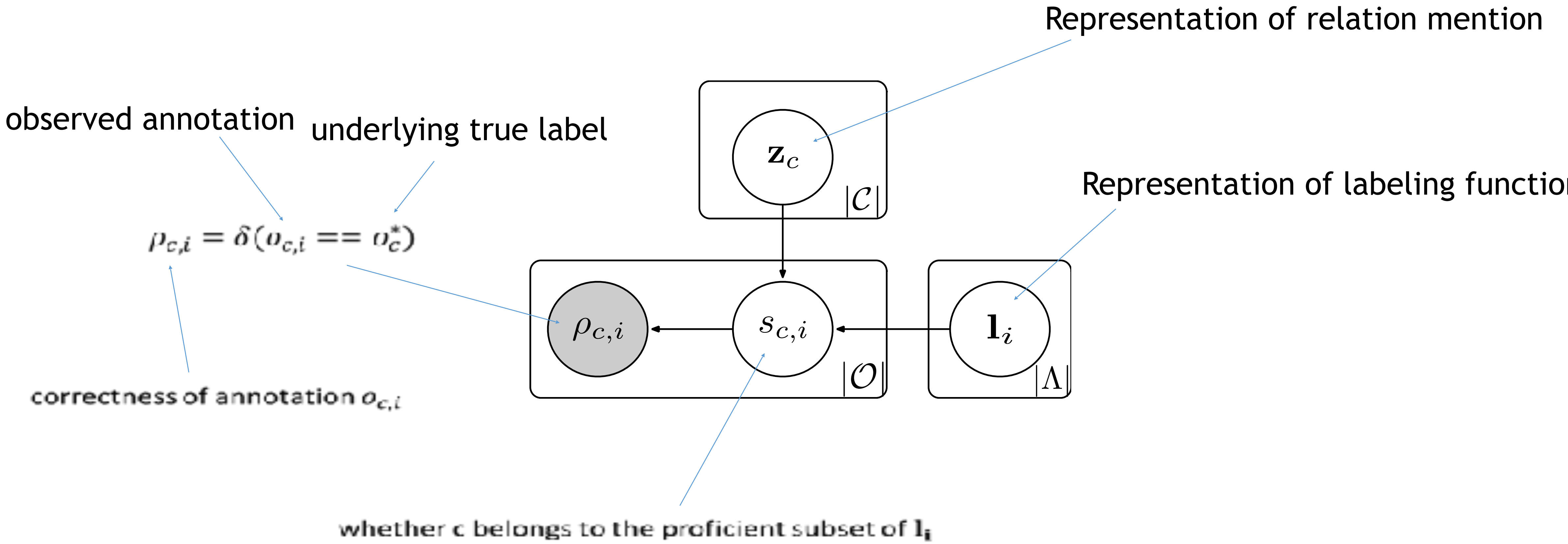


Heterogeneous Supervision generation

Relation Extraction

True label discovery

- Describing the correctness of Heterogeneous Supervision



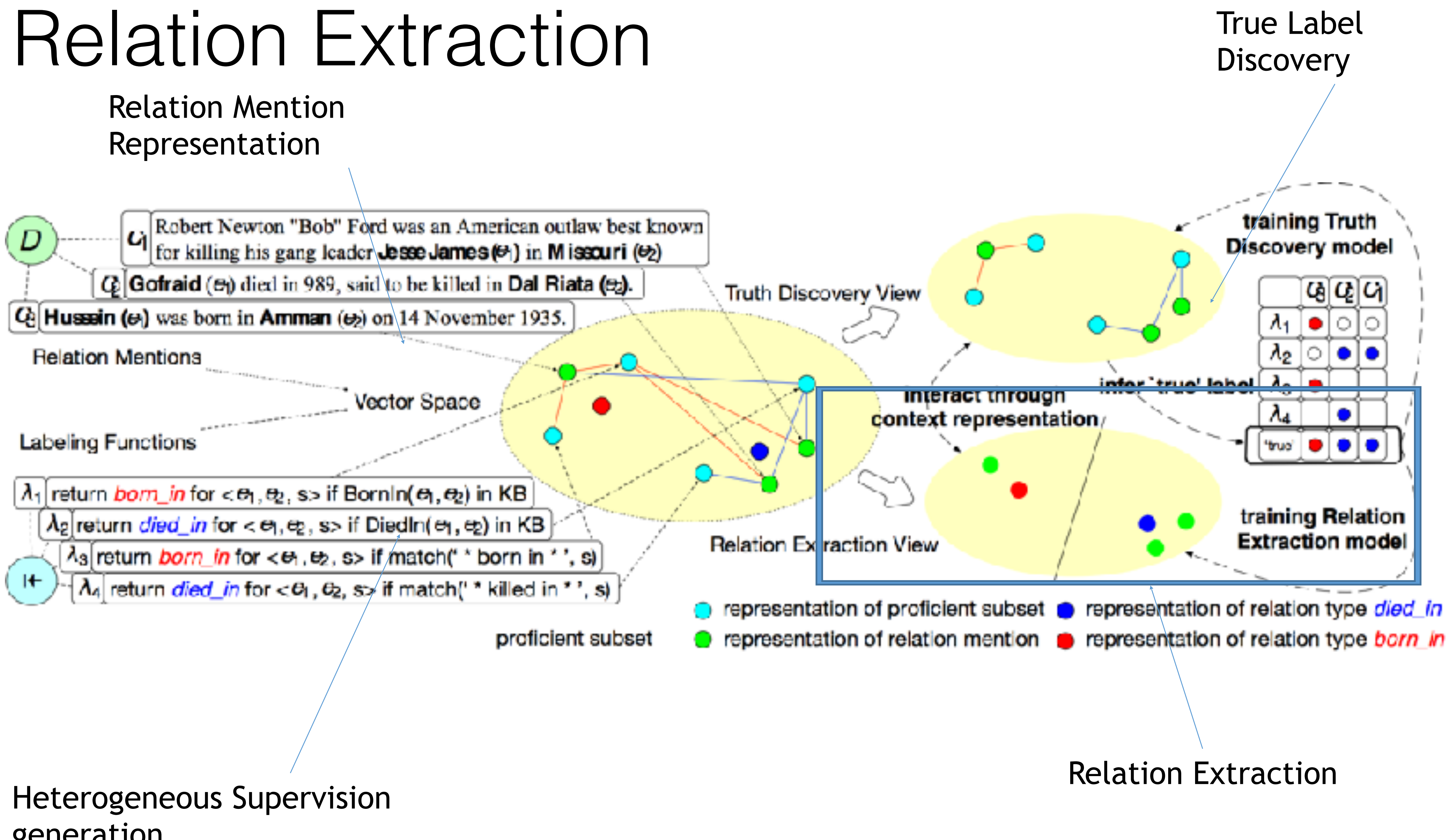
True label discovery

- correctness of annotation: $\rho_{c,i} = \delta(o_{c,i} = o_c^*)$
- prob.in proficient subset: $p(s_{c,i} = 1|\mathbf{z}_c, \mathbf{l}_i) = p(c \in \mathcal{S}_i) = \sigma(\mathbf{z}_c^T \mathbf{l}_i)$
- assume: $p(\rho_{c,i} = 1|s_{c,i} = 1) = \phi_1 \quad p(\rho_{c,i} = 1|s_{c,i} = 0) = \phi_0$
- Prob of correct annotation:

$$p(\rho_{c,i} = 1) = p(\rho_{c,i} = 1|s_{c,i} = 1) * p(s_{c,i} = 1) + p(\rho_{c,i} = 1|s_{c,i} = 0) * p(s_{c,i} = 0)$$

- The true label loss: $$\mathcal{J}_T = \sum_{o_{c,i} \in \mathcal{O}} \log(\sigma(\mathbf{z}_c^T \mathbf{l}_i)\phi_1^{\delta(o_{c,i}=o_c^*)}(1-\phi_1)^{\delta(o_{c,i}\neq o_c^*)} + (1-\sigma(\mathbf{z}_c^T \mathbf{l}_i))\phi_0^{\delta(o_{c,i}=o_c^*)}(1-\phi_0)^{\delta(o_{c,i}\neq o_c^*)})$$

Relation Extraction



Relation Extraction (context aware)

- Adopts soft-max as the relation extractor:

$$p(r_i|\mathbf{z}_c) = \frac{\exp(\mathbf{z}_c^T \mathbf{t}_i)}{\sum_{r_j \in \mathcal{R} \cup \{\texttt{None}\}} \exp(\mathbf{z}_c^T \mathbf{t}_j)}$$

- Loss function: KL-Divergence:

$$\mathcal{J}_R = -\sum_{c \in \mathcal{C}_l} KL(p(.|\mathbf{z}_c)||p(.|o_c^*))$$

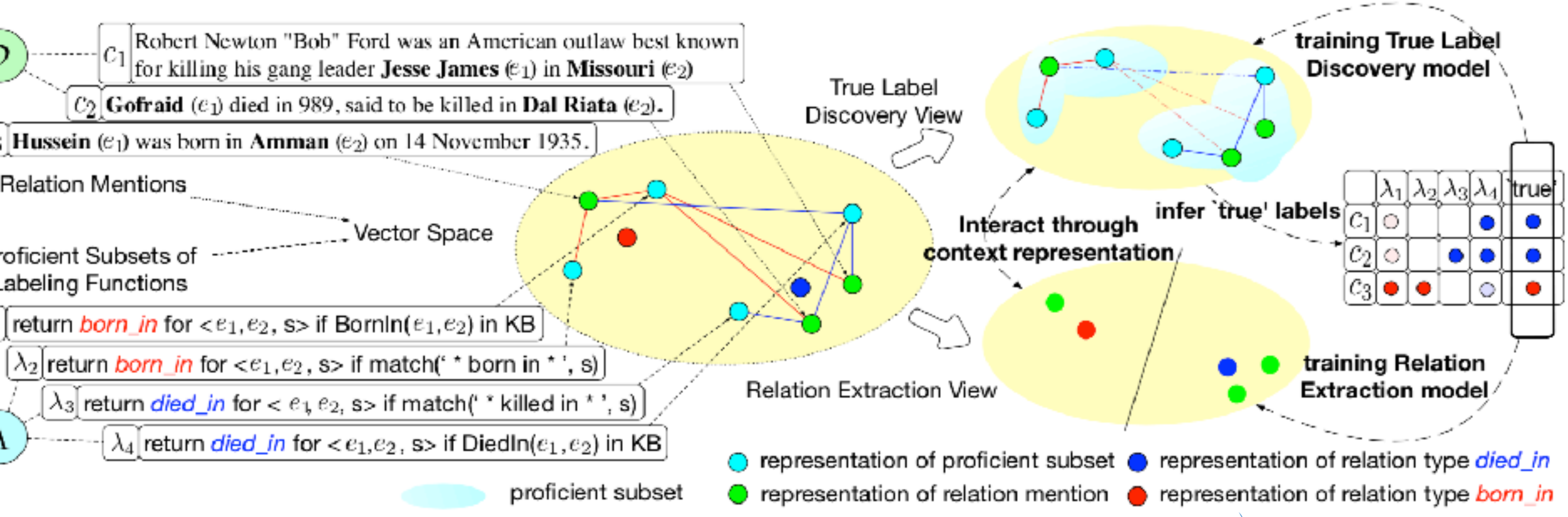
- true label distribution

$$p(r_i|o_c^*) = \begin{cases} 1 & r_i = o_c^* \\ 0 & r_i \neq o_c^* \end{cases}$$

A Representation Learning Approach

Relation Mention Representation

True Label Discovery



Heterogeneous Supervision generation

Relation Extraction

Model Learning

- Joint optimize three components

$$\min_{W,\mathbf{v},\mathbf{v}^*,\mathbf{l},\mathbf{t},o^*} \mathcal{J} = -\mathcal{J}_R - \lambda_1 \mathcal{J}_E - \lambda_2 \mathcal{J}_T$$

$$\text{s.t. } \forall c \in \mathcal{C}_l, o_c^* = \operatorname*{argmax}_{o_c^*} \mathcal{J}_T, \mathbf{z}_c = g(\mathbf{f}_c)$$

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

Kind	Wiki-KBP		NYT	
	#Types	#LF	#Types	#LF
Pattern	13	147	16	115
KB	7	7	25	26

Table 4: Number of labeling functions and the relation types they can annotated w.r.t. two kinds of information

Number of None type

Datasets	NYT	Wiki-KBP
% of None in Training	0.6717	0.5552
% of None in Test	0.8972	0.8532

Table 3: Proportion of None in Training/Test Set

Dataset	Wiki-KBP	NYT
Total Number of RM	225977	530767
RM annotated as None	100521	356497
RM with conflicts	32008	58198
Conflicts involving None	30559	38756

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

Experiments

Method	Relation Extraction						Relation Classification	
	NYT			Wiki-KBP			NYT	Wiki-KBP
	Prec	Rec	F1	Prec	Rec	F1	Accuracy	Accuracy
NL+FIGER	0.2364	0.2914	0.2606	0.2048	0.4489	0.2810	0.6598	0.6226
NL+BFK	0.1520	0.0508	0.0749	0.1504	0.3543	0.2101	0.6905	0.5000
NL+DSL	0.4150	0.5414	0.4690	0.3301	0.5446	0.4067	0.7954	0.6355
NL+MultiR	0.5196	0.2755	0.3594	0.3012	0.5296	0.3804	0.7059	0.6484
NL+FCM	0.4170	0.2890	0.3414	0.2523	0.5258	0.3410	0.7033	0.5419
NL+CoType-RM	0.3967	0.4049	0.3977	**0.3701**	0.4767	0.4122	0.6485	0.6935
TD+FIGER	0.3664	0.3350	0.3495	0.2650	**0.5666**	0.3582	0.7059	0.6355
TD+BFK	0.1011	0.0504	0.0670	0.1432	0.1935	0.1646	0.6292	0.5032
TD+DSL	0.3704	0.5025	0.4257	0.2950	0.5757	0.3849	0.7570	0.6452
TD+MultiR	**0.5232**	0.2736	0.3586	0.3045	0.5277	0.3810	0.6061	0.6613
TD+FCM	0.3394	0.3325	0.3360	0.1964	0.5645	0.2914	0.6803	0.5645
TD+CoType-RM	0.4516	0.3499	0.3923	0.3107	0.5368	0.3879	0.6409	0.6890
REHESSION	0.4122	**0.5726**	**0.4792**	0.3677	0.4933	**0.4208**	**0.8381**	**0.7277**

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)

Dataset & Method		Prec	Rec	F1	Acc
Wiki-KBP	Ori	**0.3677**	0.4933	**0.4208**	**0.7277**
	TD	0.3032	**0.5279**	0.3850	0.7271
NYT	Ori	**0.4122**	**0.5726**	**0.4792**	**0.8381**
	TD	0.3758	0.4887	0.4239	0.7387

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

Case Study

Relation Mention	REHESSION	Investment
Ann Demeulemeester (**born** 1959 , Waregem , *Belgium*) is a ...	born-in	None
Raila Odinga was **born** at ..., in *Maseno*, Kisumu District, ...	born-in	None
Ann Demeulemeester (**elected** 1959 , Waregem , *Belgium*) is a ...	None	None
Raila Odinga was **examined** at ..., in *Maseno*, Kisumu District, ...	None	None

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 Italics.

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