DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction

Shun Zheng[1] & Xu Han[2] & Yankai Lin[2] & Peilin Yu[3] & Lu Chen[1]
Ling Huang[1,4] & Zhiyuan Liu[2] & Wei Xu[1]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3] Department of Computer Sciences, University of Wisconsin-Madison, Madison, USA
[4] AHI Fintech Inc., Beijing, China

{zhengs14,hanxu17,linyk14,lchen17}@mails.tsinghua.edu.cn; peilin@cs.wisc.edu;
linghuang@fintec.ai; {liuzy,weixu}@tsinghua.edu.cn;

交叉信息研究院
Institute for Interdisciplinary Information Sciences

Paper Code

Motivation

Distant supervision (DS) [3] can generate training data for relation extraction automatically, but it may also introduce intolerable labeling noises, as Figure 1 shows.

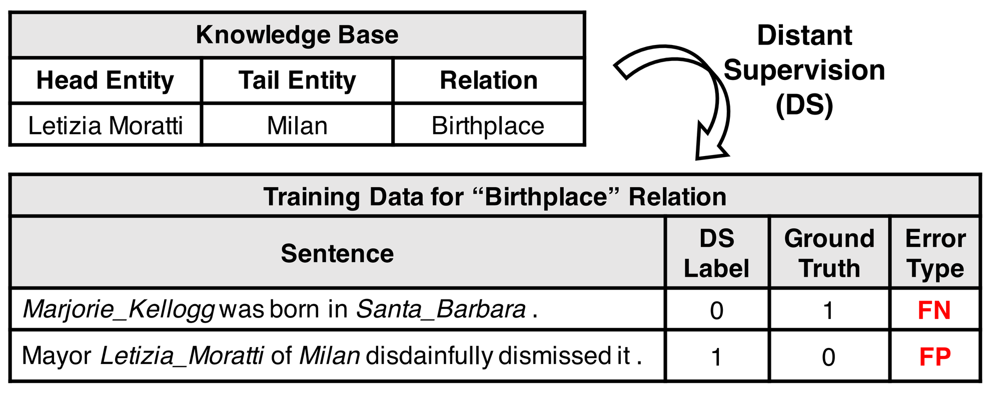


Figure 1: Two types of error labels, false negatives (FN) and false positives (FP), caused by DS.

Although the weak label fusion (WLF) paradigm [4] can leverage both DS and pattern-based labeling to produce denoised training labels, it requires human experts to write relation-specific patterns, which is both a high-skill and labor-intensive task.

Based on DS and WLF, we propose DIAG-NRE for distantly supervised neural relation extraction (NRE), which includes the following advantages:

- denoising noise labels with reduced human skill requirements by generating patterns automatically;
- enabling quick generalization to new relation types by only requiring a few human annotations;
- interpreting which patterns NRE models have learned;
- interpreting from what kinds of noises the target relation type suffers.

DIAG-NRE

As Figure 2 shows, DIAG-NRE contains two key stages: *pattern extraction* and *pattern refinement*.

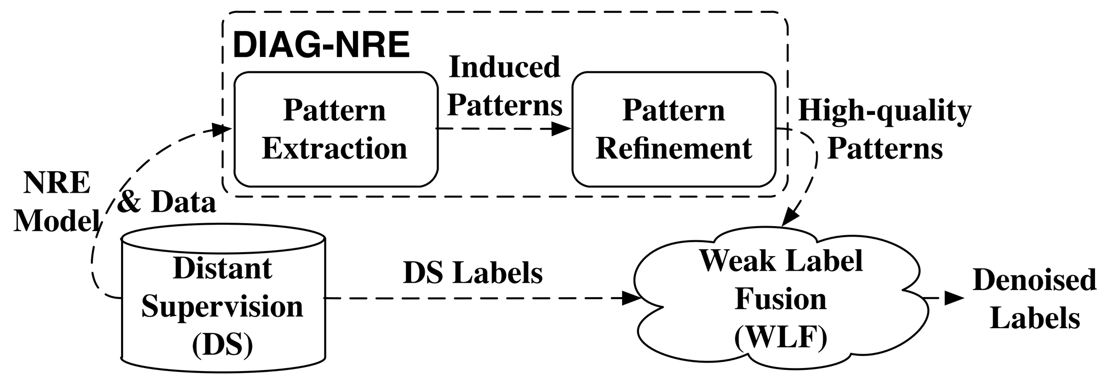
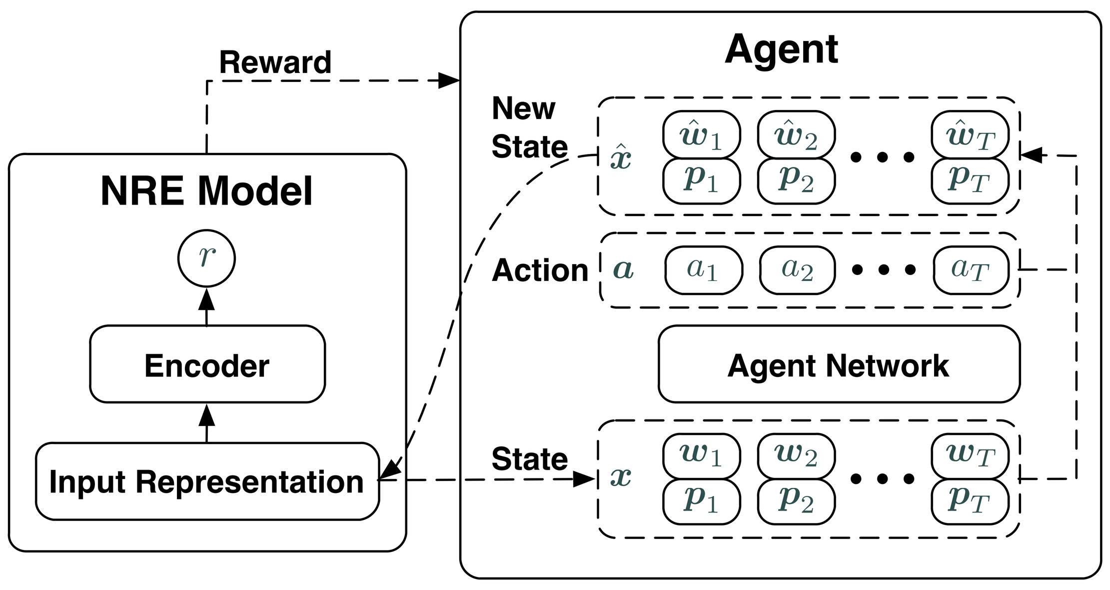


Figure 2: An overview of DIAG-NRE.

Pattern Extraction. We build an agent to distill relation-specific patterns from pretrained NRE models by reinforcement learning (RL), where the reward design encourages to erase irrelevant tokens and preserve the raw target prediction simultaneously.



Pattern-induction Example						
Entities	PER				CITY	
Tokens	*Joachim_Fest*	was	born	in	*Berlin*	.
Actions	0	1	0	0	*0*	1
Pattern	ENTITY1:PER PAD{1,3} born in ENTITY2:CITY					

Figure 3: The RL-based pattern-extraction workflow and a typical pattern-induction example, where we induce a pattern for the `Birthplace` relation via a series of actions (0: *retaining*, 1: *erasing*).

Pattern Refinement. We build a pattern hierarchy to remove redundant ones and ask human experts to annotate a certain number of actively selected instances, which are matched by those most representative patterns. Based on human annotations, we can refine previously induced patterns and get high-quality ones for the WLF stage.

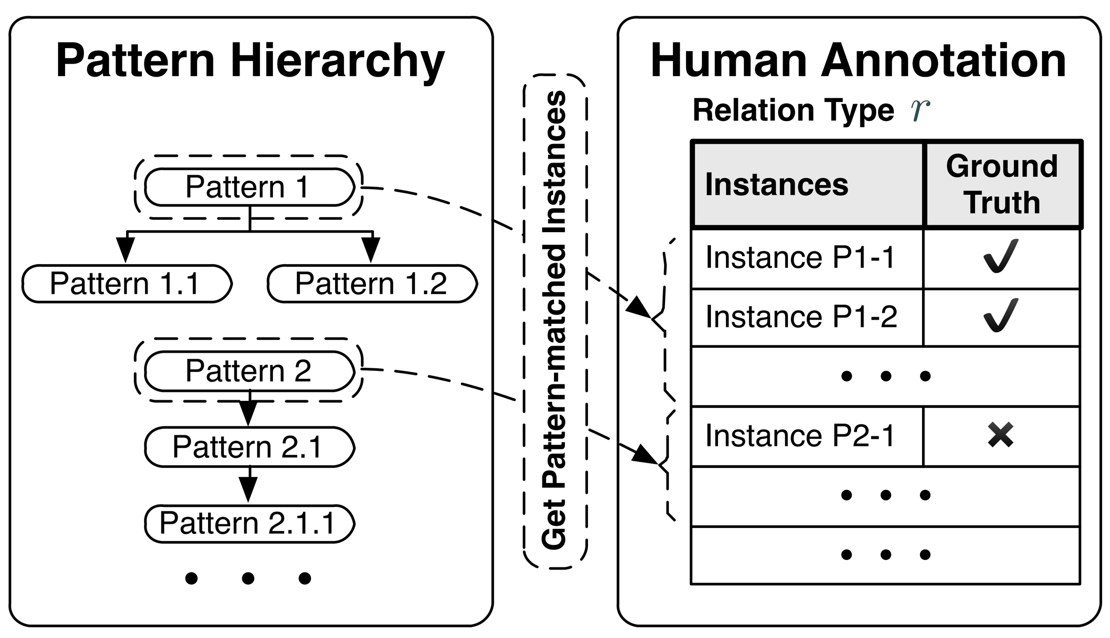


Figure 4: The human-in-the-loop pattern refinement workflow.

Experiments

To clearly show different noise behaviors for various relation types, we

- create an independent binary classification task for each relation type;
- measure the quality of different weak training labels by the testing performance of NRE models trained on them;
- utilize human-annotated labels for testing.

Based on the above setup, we compare DIAG-NRE with three baselines:

- *DS*, the vanilla distant supervision strategy;
- *Gold Label Mix* [2], mixing DS-generated noise labels with high-quality human labels;
- *RLRE* [1], a latest algorithm that automatically adjust DS-generated labels by RL.

Specifically, we compare them on 14 relation types of two public datasets, NYT and UW, whose statistics are summarized in Table 1.

	TID	Relation Abbreviation	Train	Test
NYT	R_0	`Bus./Company`	5.3k	186
	R_1	`Loc./Admin._Div.`	4.9k	180
	R_2	`Loc./Capital`	5.3k	20
	R_3	`Loc./Contains`	44.6k	263
	R_4	`Loc./Country`	4.9k	89
	R_5	`Loc./Neighbor.`	5.6k	55
	R_6	`Peo./National.`	7.5k	84
	R_7	`Peo./Place_Lived`	6.7k	230
	R_8	`Peo./Birthplace`	3.1k	16
	R_9	`Peo./Deathplace`	1.9k	19
UW	R_6^u	`Peo./National.`	107k	1.8k
	R_7^u	`Peo./Place_Lived`	20.9k	3.8k
	R_8^u	`Peo./Birthplace`	15.3k	458
	R_9^u	`Peo./Deathplace`	5.7k	1.3k

Table 1: Dataset statistics.

Main Results

From Table 2, we can observe that DIAG-NRE achieves considerable improvements in most cases.

TID	DS			Gold Label Mix			RLRE			DIAG-NRE				
	P.	R.	F1	P.	R.	F1	P.	R.	F1	P.	R.	F1	Inc-DS	Inc-Best
R_0	95.1	41.5	57.8	95.7	40.8	57.2	97.7	32.4	48.6	95.7	42.8	**59.1**	**+1.4**	**+1.4**
R_1	91.9	9.1	16.4	90.2	11.7	20.2	92.6	4.2	8.0	94.5	44.8	**60.7**	**+44.3**	**+40.4**
R_2	37.0	83.0	50.8	40.0	85.0	54.0	64.8	68.0	**66.1**	42.4	85.0	56.0	**+5.2**	-10.1
R_3	87.5	79.2	83.2	87.1	80.2	**83.5**	87.5	79.2	83.2	87.0	79.8	83.2	+0.0	-0.3
R_4	95.3	50.1	64.7	94.1	49.0	63.9	98.2	47.9	64.0	94.5	57.5	**71.5**	**+6.7**	**+6.7**
R_5	82.7	29.1	42.9	84.7	29.5	43.6	82.7	29.1	42.9	84.5	37.5	**51.8**	**+8.9**	**+8.3**
R_6	82.0	83.8	**82.8**	81.6	84.0	82.7	82.0	83.8	82.8	81.5	83.3	82.3	-0.5	-0.5
R_7	82.3	22.3	35.1	82.0	22.6	35.4	83.5	21.8	34.5	82.0	25.6	**39.0**	**+3.8**	**+3.6**
R_8	66.2	32.5	39.8	70.5	47.5	55.8	66.2	32.5	39.8	73.4	61.3	**65.5**	**+25.7**	**+9.7**
R_9	85.4	73.7	77.9	85.9	80.0	81.5	85.4	73.7	77.9	89.0	87.4	**87.1**	**+9.2**	**+5.6**
Avg.	80.5	50.4	55.1	81.2	53.0	57.8	84.1	47.3	54.8	82.5	60.5	**65.6**	**+10.5**	**+6.5**
R_6^u	35.9	75.7	48.7	35.8	75.0	48.5	36.0	75.3	**48.7**	36.2	74.5	48.7	+0.0	-0.0
R_7^u	57.8	18.5	28.0	59.3	19.1	28.8	57.8	18.5	28.0	56.3	23.5	**33.1**	**+5.1**	**+4.3**
R_8^u	37.3	64.0	46.9	40.0	64.9	49.1	37.3	64.0	46.9	48.1	71.9	**57.5**	**+10.6**	**+8.3**
R_9^u	77.1	71.3	74.0	77.5	70.3	73.5	77.1	71.3	74.0	80.7	71.1	**75.4**	**+1.5**	**+1.5**
Avg.	52.0	57.4	49.4	53.1	57.3	50.0	52.0	57.3	49.4	55.3	60.2	**53.7**	**+4.3**	**+3.5**

Table 2: Main experimental results

Case Studies

To intuitively explain how DIAG-NRE works, we show some typical cases in Table 3.

- For FN error labels, positive patterns can help to remedy the incompleteness of the knowledge base and encourage the learning of valuable patterns.
- For FP error labels, negative patterns can prevent the model from remembering such relation-irrelevant but frequently occurred patterns.

TID	Patterns & Matched Examples	DS	RLRE	Ours
R_1	**Pos:** `in ENTITY2:CITY PAD{1,3} ENTITY1:COUNTRY` **#DS/#P:** 382 / 2072			
	Example: He will , however , perform this month in *Rotterdam* , the *Netherlands* , and Prague .	0	`None`	0.81
R_8	**Pos** `ENTITY1:PER PAD{1,3} born PAD{1,3} ENTITY2:CITY` **#DS/#P:** 44 / 82			
	Example: *Marjorie_Kellogg* was born in *Santa_Barbara* .	0	0	1.0
	Neg: `mayor ENTITY1:PER PAD{1,3} ENTITY2:CITY` **#DS/#P:** 21 / 62			
	Example: Mayor *Letizia_Moratti* of *Milan* disdainfully dismissed it .	1	1	0.0
R_9^u	**Pos:** `ENTITY1:PER died PAD{4,9} ENTITY2:CITY` **#DS/#P:** 66 / 108			
	Example: *Dahm* died Thursday at an assisted living center in *Huntsville* ...	0	0	1.0
	Neg: `ENTITY1:PER PAD{4,9} rally PAD{1,3} ENTITY2:CITY` **#DS/#P:** 40 / 87			
	Example: *Bhutto* vowed to hold a rally in *Rawalpindi* on Friday ...	1	1	0.0

Table 3: We show five cases with positive (Pos) or negative (Neg) patterns, the number of DS-generated positive labels over the number of pattern-matched instances (#DS/#P), one pattern-matched example, and associated training labels produced by various methods.

Conclusion & Future Work

In summary, DIAG-NRE introduces a novel strategy to efficiently utilize human efforts for DS-based NRE. Therefore, it will be interesting to extend DIAG-NRE to other DS-related applications, such as event extraction and question answering.

References

[1] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. Reinforcement learning for relation classification from noisy data. In *AAAI*, 2018.

[2] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. Effective crowd annotation for relation extraction. In *NAACL-HLT*, 2016.

[3] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, 2009.

[4] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: creating large training sets, quickly. In *NIPS*, 2016.