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

Shun Zheng1 & Xu Han2 & Yankai Lin2 & Peilin Yu3 & Lu Chen4 & Ling Huang2,4 & Zhiyuan Liu2 & Wei Xu1
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 Fin-tech Inc, Beijing, China

{zhengs14,hanxu17,linyk34,lichen7}@mails.tsinghua.edu.cn; peilin@cs.wisc.edu
linghuang@fintecc.ai; liuyu.weixiu@tsinghua.edu.cn

**Motivation**

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

Although the weak label fusion (WLF) paradigm 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-skilled 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.

**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 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.

**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: mixing DS-generated noise labels with high-quality human labels;
- ERLE: a latest algorithm that automatically adjust DS-generated labels by RL.

Specifically, we compare on 11 relation types of two public datasets, NYT and UWE, whose statistics are summarized in Table 1.

**Main Results**

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

**Case Studies**

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

For DP 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-inherent but frequently occurred patterns.

**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**


