On the use of Silver Standard Data for Zero-shot Classification Tasks in Information Extraction

Speaker：Jianwei Wang

Authors：Jianwei Wang, Tianyin Wang, Ziqian Zeng

Content

PART ONE

Introduction

Introduction

Information Extraction (IE)

- Supervised methods → Numerous labeled data
- Zero-shot methods

1. relation extraction (RE)
2. cross-lingual relation extraction
3. event argument classification (EAC)
...

Converted → Pretrained Language Model → textual entailment or summarization

Since pre-trained models can directly infer the categories of unlabeled data, they can serve as low-cost annotators, producing **large-scale silver standard data**. However, in the above works, silver standard data are not well-exploited.

1. Noise robust training
2. identification
3. Any better way + Pretrained Language Model

PART TWO

Related Work

Related Work

Zero-shot Relation Extraction

In classical zero-shot learning settings, the classes in the training and test phases are disjoint. In the training phase, it requires a large amount of annotated RE data from seen classes. In the test phase, zero example for each unseen relation type during the test phase is needed. Recent works formulated the zero-shot RE to **other NLP tasks**, such as reading comprehension and textual entailment.

However, the classical zero-shot setting still requires a large amount of annotated data in the training phase. Recent works push the zero-shot setting to an extreme case where annotated data is not available in the training phase. They obtained supervision from other available resources such as language models, relation descriptions, and off-the-shelf models from other NLP tasks.

QA4IE (Zhang et al., 2023) is the state-of-the-art method in the zero-shot RE task, primarily owing to the powerful capacity of Large Language Models.

Zero-shot Cross-lingual Information Extraction

Existing approaches to zero-shot cross-lingual In formation Extraction (IE) can be categorized into three main types: translation-based, feature-based, and distillation-based methods. However, all of these methods require significant manual effort to obtain labeled data for the source languages

Zero-shot Event Argument Classification.

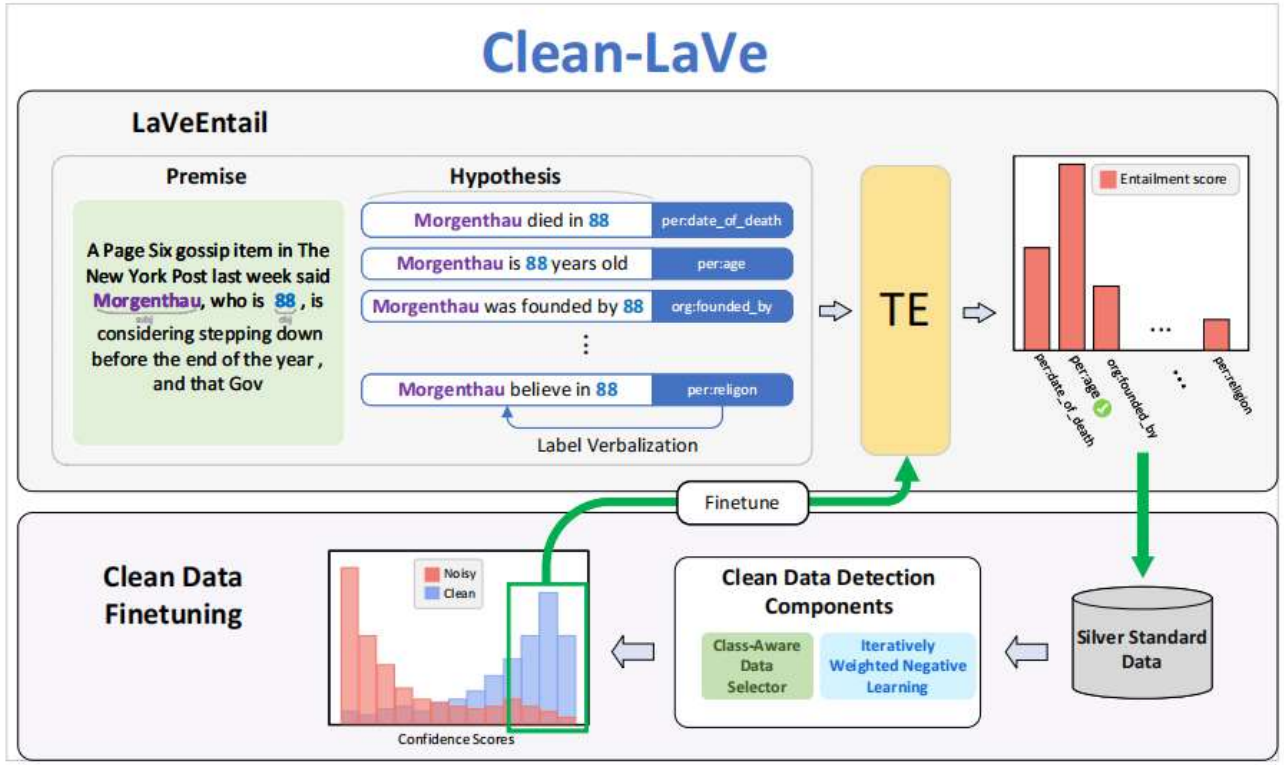
Exiting zero-shot event argument classification tasks are based on label representations , reading comprehension, and pretrained language models.

Lin et al. (2023) is the state-of-the-art zero-shot EAC method, which prompts the pre-trained language models and regularizes the prediction by global constraints.

PART THREE

Clean-LaVe

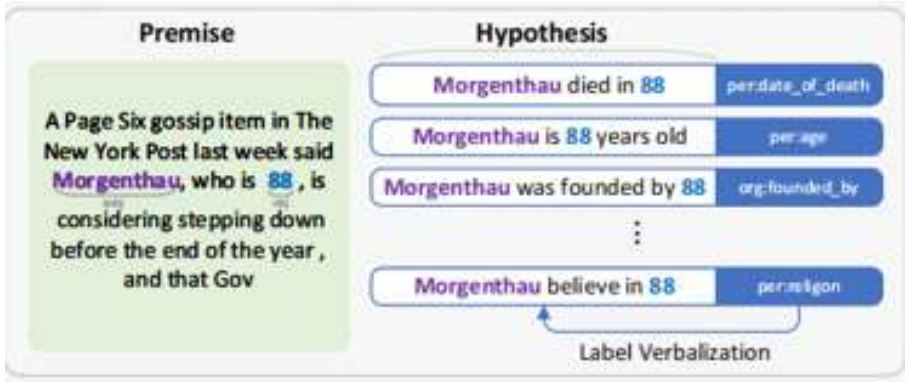
Clean_LaVe



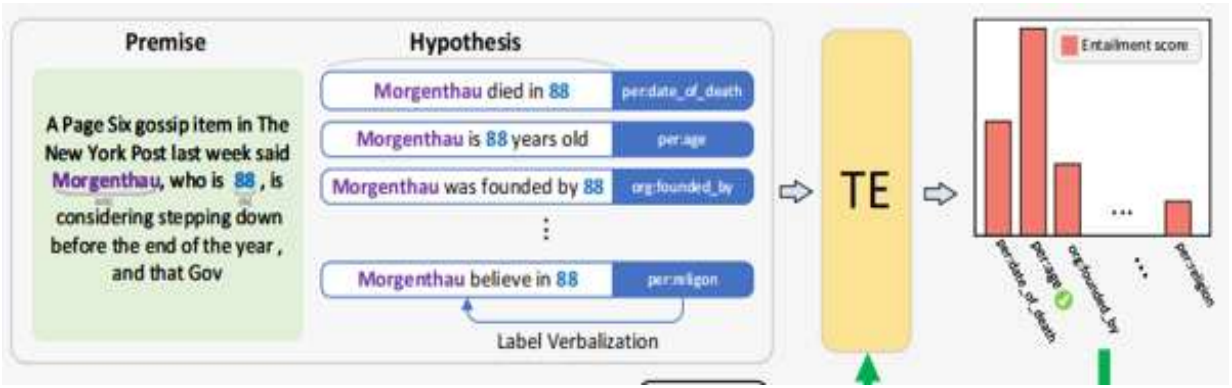
We propose a novel framework called Clean-LaVe. The framework involves two main steps:

1. Firstly, detecting a small subset of clean data from the silver standard data using the clean data detection module, and
2. Secondly, utilizing the selected clean data to finetune the pretrained model.

Clean_LaVe



The label verbalization process creates templates of classes and then uses them to generate hypotheses.
For example, the relation **per:schools_attended** can be verbalized as {subj} studied in {obj}, where {subj} and {obj} are placeholders for subject and objective entities.

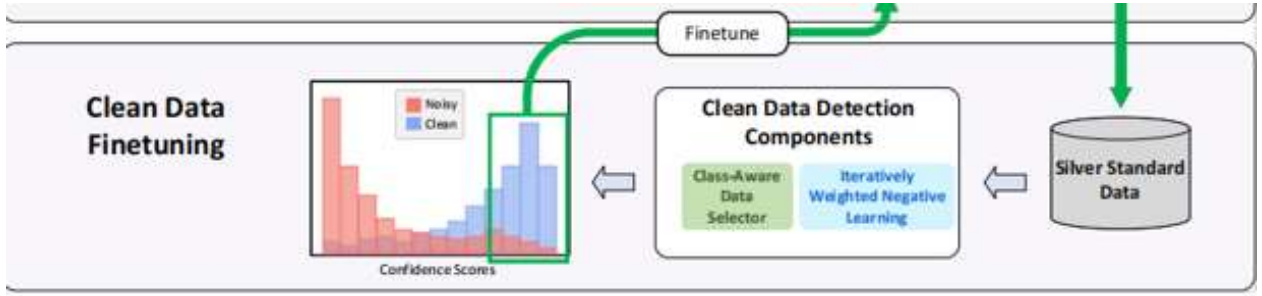


For each input sentence, Clean_LaVe constructed hypotheses that are generated by verbalization templates of all relation types (or argument roles), and fed them to a TE model, and obtained entail ment scores of all hypotheses.
Clean_LaVe inferred that the predicted relation (or role) type of the input sentence is the relation (or role) type whose hypothesis **yields the highest entailment score.**

Entity type information is helpful to infer relation types. In the inference stage, when the entity type information is given, we could **rule out** some relation types that are impossible to be ground truth.

In the case where there is no relation between two entities, a **threshold-based** approach is used to detect no_relation.

Clean_LaVe



$$\mathcal{L}_{neg} = -\sum_{d \in D}\sum_{i=1}^{|\mathcal{Y}|} \hat{\mathbf{y}}_i^d \log(1-\mathbf{p}_i^d), \qquad (1)$$

Within the clean data detection module, we introduce a **iteratively weighted negative learning algorithm** to obtain confidence scores that allow us to distinguish clean data from noisy data.
The original negative learning algorithm (Kim et al., 2019) only performs well when the dataset is balanced. However, in real-world scenarios, this assumption may not hold.
To address this issue, we introduce an iterative weighting strategy to allow the algorithm to handle an imbalance dataset.

Negative Learning loss is robust to noise, as shown in (1).
The original NL loss in equation (1) treats each class equally, which may not be appropriate when dealing with real-world datasets that exhibit severe class imbalance. Consequently, the model encounters much fewer samples in the minority classes, leading to underfitting (i.e., high loss values) during the training process.
It poses a challenge to distinguish between clean and noisy samples in minority classes as they both have high loss values.

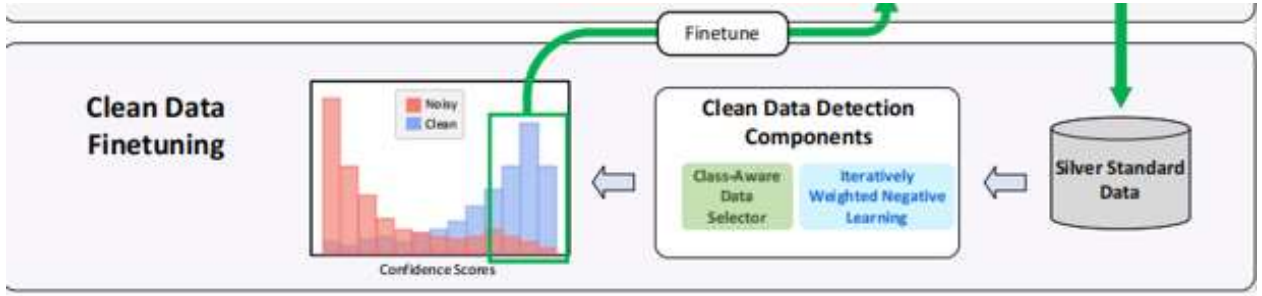
$$\mathcal{L}_{neg}^{j} = -\sum_{d \in D}\sum_{i=1}^{|\mathcal{Y}|} w_i^j \cdot \hat{\mathbf{y}}_i^d \log(1-\mathbf{p}_i^d) \qquad (2)$$

$$w_i^j = w_i^{j-1} \cdot e^{1-\frac{c_i^{j-1}}{c_{\mathcal{A}}^{j-1}}} \qquad (3)$$

$$w_i^0 = \frac{\sum_{k=1}^{|\mathcal{Y}|} c_k^0}{c_i^0} \qquad (4)$$

We propose a iteratively weighted NL loss to alleviate this issue, giving more weight on minority classes, as shown in (2),(3),(4).
Iterative weight can fix the problem that initial weight may be wrong.
Our silver dataset is noisy and is likely to give wrong weight in initial. Wrong weight can degrade the performance of Clean-LaVe.
Iterative weight can make some **correction** to wrong proportions in initial.

Clean_LaVe



$$D_{clean} = \arg\max_{D_s:|D_s|=\eta\cdot|D_{silver}|} \mathcal{S}(D_s). \quad (5)$$

To select clean data, confidence scores serve as a straightforward metric. However, data from certain classes possibly have high confidence scores while others yield low scores.
In such cases, selecting data solely based on confidence scores may lead to a narrow range of classes being selected, potentially harming overall performance.
To mitigate this issue, we develop a **class-aware data selector** that enables the selection of data from a broader range of classes.

A straightforward approach to selecting clean data is sorting all samples according to their confidence scores and then selecting a fixed proportion η of whole data as the clean data set, as (5).
However, it does not consider class diversity. Samples in some classes can yield very high confidence scores while some classes have very low confidence scores.
Large quantities of samples in those classes are selected while some classes even do not have any clean data selected, which harms performance badly.

Algorithm 1 Class-Aware Data Selector

Input: silver standard data set D_{silver}, proportion η, diversity number m, the set of classes $\mathcal{C}$, the total confidence scores function $\mathcal{S}(\cdot)$.

1: $D_{clean} = \varnothing$.
2: Obtain D_{clean} using Eq. 5 by setting the proportion to η.
3: $D_{rest} = D_{silver} - D_{clean}$, divide D_{rest} into $|\mathcal{C}|$ subsets according to class predictions. The subset for class c is denoted as D^c.
4: **for** c in $\mathcal{C}$ **do**
5: $D^c_{clean} = \arg\max_{D_s:|D_s|=\frac{|D^c|}{|D_{rest}|}\cdot m} \mathcal{S}(D_s)$
6: $D_{clean} = D_{clean} \cup D^c_{clean}$
7: **end for**

Output: clean data set D_{clean}.

We propose a class-aware data selection algorithm that **considers confidence scores as well as class diversity**.
First, we select a proportion η of data with high confidence scores. This step can ensure that samples with low noise levels are selected.
Next, we select m more samples to encourage diversity.
η is used to select high confidence data and m is used to make supplement according to label distribution.

PART FOUR

Experiments

Experiments

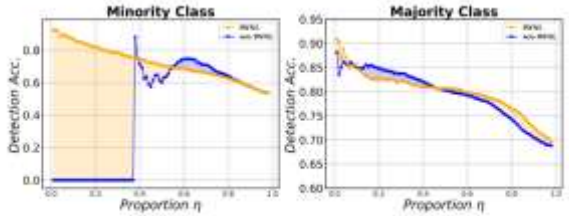
	RE		Cross-lingual RE			EAC
	TACRED	Wiki80	Smiler-It	Smiler-Po	Smiler-Kr	ACE05-E+
CE	45.35±0.58	40.76±0.29	40.79±0.12	41.56±0.19	49.75±0.50	71.79±0.96
GCE (Zhang and Sabuncu, 2018)	45.93±0.67	41.28±0.61	47.27±0.21	45.99±0.60	53.35±0.49	71.61±0.79
SCE (Wang et al., 2019)	45.82±0.92	41.12±0.24	40.97±0.70	40.41±0.31	47.79±0.09	71.88±0.26
Co-Regularization (Zhou and Chen, 2021)	48.86±0.34	28.48±0.42	42.17±0.61	41.86±0.19	50.16±0.70	72.93±0.17
O2U (Huang et al., 2019)	47.52±0.81	42.62±0.03	41.12±0.23	44.47±0.66	49.67±0.49	69.83±0.06
DivideMix (Li et al., 2020)	49.78±0.80	45.52±0.26	41.94±0.78	43.79±0.69	52.48±0.80	69.13±0.64
Global_Constraints (Lin et al., 2023) -	-	-	-	-	-	66.1*
QA4RE (Zhang et al., 2023)	58.55±0.05	43.93±0.09	**56.42**±0.84	38.09±0.19	56.08±0.73	64.74±0.84
LaVeEntail[1] (Sainz et al., 2021)	52.18	41.16	39.96	37.84	44.30	71.60
Labeled Data Finetune (1%)	56.61±1.29	47.39±0.33	51.85±0.96	46.44±0.66	47.33±0.91	76.21±1.50
Labeled Data Finetune (5%)	63.72±1.03	53.89±0.46	52.56±0.57	49.56±0.54	55.30±0.14	78.87±0.17
Silver-LaVe	54.67±0.58	44.57±0.31	48.91±0.55	50.60±0.38	54.64±0.81	80.18±0.08
Clean-LaVe	**63.36**±1.03	51.53±0.53	55.09±0.05	**52.99**±0.88	**59.41**±0.84	**81.22**±0.38
– Iteratively Weighted Negative Learning	58.66±0.93	48.44±0.44	54.20±0.97	48.09±0.59	57.18±0.95	78.07±0.82
– Class-Aware Data Selector	59.55±0.98	**52.52**±0.21	54.97±0.33	50.14±0.64	57.34±0.26	78.14±0.61
– Above Both	56.41±1.82	52.34±0.38	54.28±0.65	45.99±0.41	54.97±0.74	77.37±0.67

Table 2: Results of zero-shot classification tasks. We report the average of micro F1 scores in 3 runs. The best F1 scores are marked in **bold**. SOTA baselines are highlighted with underline. Results marked with * are retrieved from the original paper.

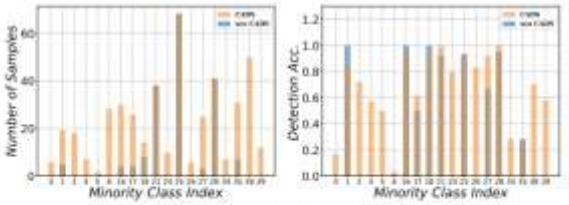
We use the common **zero-shot setting**, all data used for training are unlabeled, and only 1% development set are available for adjusting hyperparameters.
For the zero-shot RE task, we evaluate our method on the **TACRED** and **Wiki80** dataset.
For the zero-shot cross-lingual RE task, we evaluate our method on the **Smiler** dataset which contains 14 languages. We evaluate Clean-LaVe in three languages, i.e., **Italian, Polish, and Korean**.
For the zero-shot EAC task, we evaluate our method on **ACE05-E+**.
The TE model we used for RE and EAC tasks is **Deberta-v2**.

As shown in the first and second block of Table 2, Clean-LaVe outperforms **noise-robust loss based methods** and semi-supervised based noisy labels learning methods across all datasets.
As shown in the third block, Clean-LaVe outperforms the **SOTA methods** by 3% ~15% on all datasets except on the Smiler-It. On Smiler-It, QA4RE outperforms Clean-LaVe by 1%. Despite facing a stronger competitor based on ChatGPT, Clean-LaVe delivers commendable overall performance.
As shown in the fourth block, Clean-LaVe can gain significant improvement compared to **LaVeEntail** by 10% ~16%. Additionally, our method is comparable to or even outperforms the **supervised LaVeEntail with 5% labeled data**.
As shown in the last block, Clean-LaVe outperforms **Silver-LaVe**, indicating the effectiveness of the clean data detection module.
We also provide results after **removing** Iteratively Weighted Negative Learning, Class-Aware Data Selector, and both of them respectively. After removing, we observe decreases in performance across all datasets, which validates the effectiveness of these component.

Experiments



(a) Analysis of IWNL on minority and majority classes.



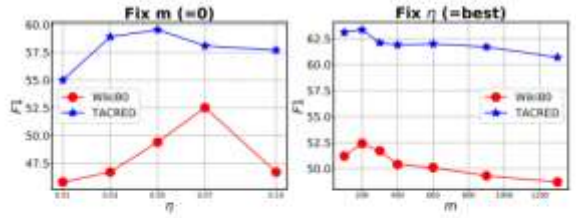
(b) Analysis of CADS.

Figure 2: Analysis of IWNL and CADS on TACRED.

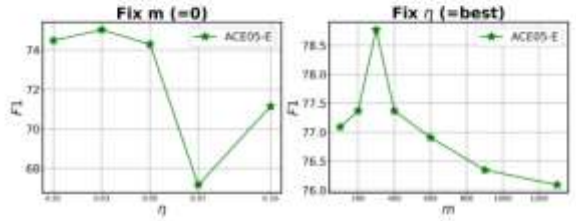
Iteratively Weighted Negativing Learning (IWNL) can alleviate the effect of underfitting and improve the clean data detection accuracy of minority classes.

As depicted in Figure 2(a) , IWNL yields consistently higher detection accuracy scores than w/o IWNL on minority classes, while the performance of IWNL on majority classes is comparable to w/o IWNL.

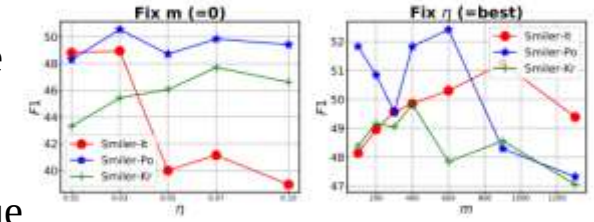
Class-Aware Data Selector (CADS) can encourage clean samples from a broader range of classes. As depicted in Figure 2(b), there are more orange bars than blue bars, indicating CADS selects samples from more classes, especially from minority classes.



(a) RE task



(b) EAC task



(c) Cross-lingual RE task

Figure 3: Results of different η and m.

Class-Aware Data Selector introduces two hyper-parameters to control the selection.

η controls the number of clean data and m controls the number of samples from diverse classes.

We analyse these hyperparameters on1% development set on each dataset.

And we perform manual quality checks on the filtered data as shown in table5.

Fix m (=0)					
η	0.01	**0.05**	0.1	0.5	1
Data Accuracy	92.10	**86.02**	80.16	68.51	64.03
Class Count	4	**18**	26	32	40
Fix η at best (=0.05)					
m	100	**200**	500	700	1000
Data Accuracy	79.48	**76.40**	75.15	74.70	74.20
Class Count	26	[illegible]	39	40	40

Table 5: The reliability of the data filtered by IWNL and CADS on TACRED. The results under the best hyper-parameter are marked in bold.

PART FIVE

Conclusion

Conclusion

Our contributions are summarized as follows:

1. We propose **Clean-LaVe** to first detect a small amount of clean data which are later used to finetune the pre-trained model. We then use the finetuned model to infer the categories on the test data.

2. We propose a clean data detection module that enhances the selection process through **Iteratively Weighted Negative Learning** and **Class-Aware Data Selector**.

3. The experimental results demonstrate that our method can **outperform** the baseline by a large margin on various zero-shot classification tasks. The code is shared in https://github.com/ZeroNLP/Clean_LaVe

Future Work:

1. For further improvement, our work can be easily combined with LLM, using **LLM** as pretrained models.

2. Clean-Lave has great potential to be extend to **other tasks and settings** which is worthwhile to be explored.

Thanks!

- THANK YOU FOR YOUR LISTENING -

Email: wjwfyu@gmail.com