PLM-Augmented Rule-Based Classifiers: A Lightweight Method for Improving the Generalizability of Expert Knowledge in Novel Information Extraction Tasks

Grace LeFevre, Liam Frölund, Lori Beaman, Rob Voigt

Northwestern University

Introduction

The growing use of NLP for social impact applications [1,2] has resulted in an increasing number of novel information extraction (IE) tasks that share several common characteristics:

- Low-resource: Existing task-specific labels are unavailable [3]
- Need for expert knowledge: Extracting features of interest often relies on domain-specific expert knowledge, e.g. genre or content
- Sensitive data: Privacy concerns place constraints on usable methods

These tasks call for IE methods that can capture expert knowledge with both high precision and high recall in a way that is *lightweight*, *low-cost*, and *accessible* to researchers across many domains. We propose a PLM-Augmented Rule-Based Classifier: a high-precision rule-based classifier augmented with the results of a PLM finetuned on its output.

Related Work

- *Rule-based classifiers* can explicitly incorporate expert knowledge with high precision in an interpretable way [4,5], but generating fully comprehensive rules is time-consuming and limits recall
- *Finetuned PLMs* often perform well on domain-specific tasks but face the data scarcity bottleneck. PLMs finetuned on a small quantity of gold labels or a large quantity of labels generated via weak/distant supervision [3] are susceptible to issues like overfitting [6]
- *In-context Learning (ICL)* has enabled LLMs to perform well on many unseen tasks but even very large LLMs like ChatGPT lag behind SOTA performance on standard IE tasks [7,8]

Our Method

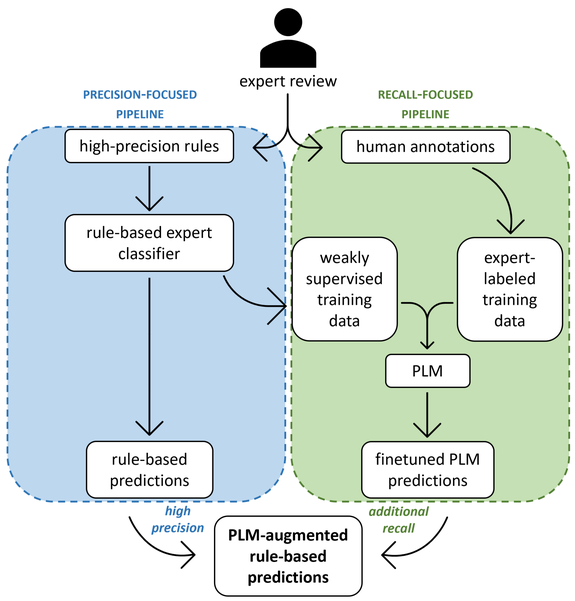

Figure 1: Method Overview

Test Case: a novel linguistic IE task

Task

(1) I strongly recommend him
(2) I recommend her without hesitation
(3) I am delighted to recommend him
(4) my highest recommendation
(5) my full and enthusiastic endorsement
(6) this letter of strong support

Figure 2: Examples of feature

- Feature: "evaluative expressions," a novel linguistic feature we identify in recommendation letters
- Dataset: ~120k recommendation letters written in support of graduate school applications, anonymized using NER
- Performance metric: relaxed span match (2+ overlapping tokens) [9]

Method implementation

1. Perform expert review of small data subset
   - 326 letters randomly sampled for expert annotation by 3 annotators
   - High IAA, with average pairwise positive specific agreement of 0.921
2. Develop a high-precision rule-based classifier
   - Subset of HA train set used to identify patterns representing most common evaluative expression structures in the dataset
   - RB algorithm written to perform regular-expression string-matching
   - High precision design constrained prediction noise to false negatives.
3. Use weakly supervised training data to finetune a PLM
   - Generated weakly supervised labels with RB classifier
     - If RB classifier predicted at least one EE in a randomly selected document, one positive example (sentence containing EE) and one negative example (sentence not containing an EE) were sampled from the document
     - Tokens labeled using IO format

	Human-annotated (HA)		Rule-based (RB)	
	train set	*test set*	*train set*	*test set*
documents	225	101	7500	1000
examples	692	326	15,000	2000

Table 1: Data distribution

   - Finetuned DistilBERT [10] models for token classification task
     - Three models trained on different datasets: HA only (gold labels), RB only (weak supervision), and both
     - Performed hyperparameter tuning with random search, selecting model with lowest validation loss
4. Augment rule-based predictions with PLM predictions
   - Used union operation as simple combination approach, including the longer prediction when RB and PLM predictions were non-identical

Results

	Model:	*Training approach:*	*Pr*	*Rec*	*F1*
(1)	rule-based classifier		**1**	0.687	0.815
(2)	finetuned PLM	gold labels (GL)	0.718	0.969	0.825
(3)		weak supervision (WS)	0.933	0.515	0.664
(4)		GL + WS	0.927	0.546	0.687
(5)	generative LLM	few-shot prompting	0.586	0.791	0.674
(6)	PLM-augmented rule-based classifier	gold labels (GL)	0.721	**0.982**	0.831
(7)		weak supervision (WS)	0.959	0.853	0.903
(8)		GL + WS	0.954	0.883	**0.917**

Table 2: Summary of results

1. Rule-based classifier (1) achieves high precision but moderate recall, since it is impossible to create completely comprehensive rules
2. PLM finetuned on small gold label set (2) achieves high recall but moderate precision, while the PLMs finetuned on large weakly-supervised sets (3 & 4) overfit to the training data
3. Few-shot prompted Mistral-7B (5) achieves only moderate precision and moderate recall even after significant post-processing
4. Our approach, the PLM-augmented rule-based classifier (7 & 8), achieves both high precision and high recall for the highest overall performance of all models tested

Contributions & Limitations

- We combine the advantages of rule-based and PLM-based approaches to achieve high precision and high recall on a novel IE task requiring domain-specific knowledge
- Our method is computationally lightweight, low-cost, and relatively accessible to researchers in a variety of domains
- This approach is applicable only to the subset of low-resource IE tasks relying on expert knowledge that can be formalized into high-precision rules

References

[1] Jin, Z., Chauhan, G., Tse, B., Sachan, M., & Mihalcea, R. 2021. How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* 2021 (pp. 3099-3113).
[2] Adauto, F., Jin, Z., Schölkopf, B., Hope, T., Sachan, M., & Mihalcea, R. 2023. Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good. *Findings of the Association for Computational Linguistics: EMNLP 2023.*
[3] Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545-2568).
[4] Chiticariu, L., Li, Y., & Reiss, F. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems!. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 827-832).
[5] Waltl, B., Bonczek, G., & Matthes, F. 2018. Rule-based information extraction: Advantages, limitations, and perspectives. *Jusletter IT* (02 2018), 4.
[6] Mahabadi, R. K., Belinkov, Y., & Henderson, J. 2021. Variational information bottleneck for effective low-resource fine-tuning. *arXiv preprint arXiv:2106.05469*.
[7] Han, R., Peng, T., Yang, C., Wang, B., Liu, L., & Wan, X. (2023). Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.
[8] Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. 2023. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. *arXiv preprint arXiv:2304.11633*.
[9] Wang, K., Stevens, R., Alachram, H., Li, Y., Soldatova, L., King, R., ... & Rzhetsky, A. 2021. NERO: a biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. *NPJ systems biology and applications*, 7(1), 38.
[10] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.