CoAug: Combining Augmentation of Labels and Labelling Rules

Rakesh R Menon [1*] **Bingqing Wang** [2] **Jun Araki** [2]
Zhengyu Zhou [2] **Zhe Feng** [2] **Liu Ren** [2]
[1] UNC Chapel-Hill
[2] Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
rrmenon@cs.unc.edu
{bingqing.wang, jun.araki, zhengyu.zhou2, zhe.feng2, liu.ren}@us.bosch.com

Abstract

Collecting labeled data for Named Entity Recognition (NER) tasks is challenging due to the high cost of manual annotations. Instead, researchers have proposed few-shot self-training and rule-augmentation techniques to minimize the reliance on large datasets. However, inductive biases and restricted logical language lexicon, respectively, can limit the ability of these models to perform well. In this work, we propose **CoAug**, a co-augmentation framework that allows us to improve few-shot models and rule-augmentation models by bootstrapping predictions from each model. By leveraging rules and neural model predictions to train our models, we complement the benefits of each and achieve the best of both worlds. In our experiments, we show that our best **CoAug** model can outperform strong weak-supervision-based NER models at least by 6.5 F1 points on the BC5CDR, NCBI-Disease, WikiGold, and CoNLL-2003 datasets.[1]

1 Introduction

Named Entity Recognition (NER) is the task of identifying entity spans of specific types in a given document. While deep learning has led to the development of highly performant supervised NER models (Ma and Hovy, 2016; Lample et al., 2016; Devlin et al., 2019), their performance is contingent on the availability of high-quality large labeled datasets, which is often expensive to collect. Moreover, it is impractical to assume the availability of large datasets for all domains. Hence, learning from limited labeled data is a pressing challenge in named entity recognition research. The majority of research in this area can be broadly classified into two distinct paradigms: few-shot learning with pre-trained language models (LMs) and weak supervision methods that utilize heuristic rules for entity extraction.

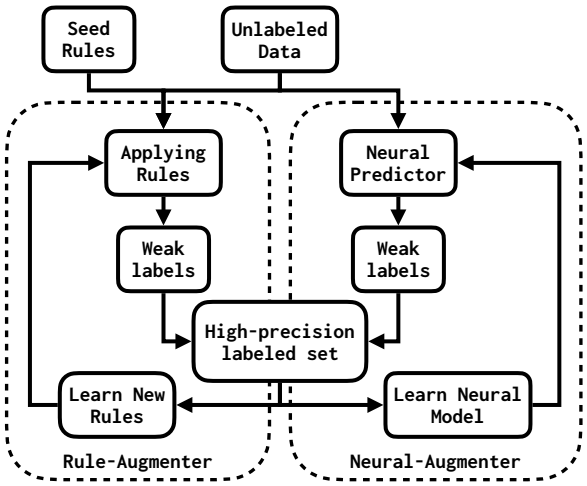

Figure 1: Illustration of the **CoAug** framework.

In few-shot learning, models are trained to identify novel entities given just a few labeled examples for each entity type. While pretrained LMs have been explored for this setting, their susceptibility to overfitting on small datasets results in poor performance. Consequently, recent works improve recognition using *prototypical networks* (**ProtoBERT**, Tänzer et al., 2022), improved representations from self-supervised pre-training of LMs (**QuIP**, Jia et al., 2022), and self-training (Huang et al., 2021). In the iterative learning process of self-training, many candidate entities are extracted and added into the training set for future iterations. However, premature models from initial iterations also add erroneous entities to the training set, resulting in models whose performance lags behind fully-supervised models that utilize large labeled datasets.

On the other hand, rule-based weak supervision methods utilize heuristic rules and manual lexicons (Shang et al., 2018; Peng et al., 2019) developed by domain experts to supervise entity recognition models. However, experts may find it challenging to enumerate all possible heuristics, which can limit the diversity of identified entities in docu-

[*] Work done during an internship at Bosch Research.
[1] Code: https://github.com/boschresearch/CoAug

Findings of the Association for Computational Linguistics: ACL 2023, pages 9062–9071

ments. In recent work, **TALLOR** (Li et al., 2021) overcomes this limitation by automatically learning rules given unlabeled data and an initial set of seed rules (tens of rules). Nonetheless, while rule-based methods offer high precision, their performance is constrained by the logical language specified by the developer, which limits the set of identifiable entities. Moreover, learning rules can fail to identify entities in new linguistic contexts that would otherwise be known.

We hypothesize that the two paradigms of few-shot learning and rule-based weak supervision can effectively complement each other, as neural models are skilled at identifying candidates from different linguistic contexts but lack precision, while rule-based methods can identify accurate candidates with precision but lack the flexibility to identify entities in different contexts. Therefore, in this work, we propose *Co-Augmentation* (**CoAug**), as shown in Figure 1, an iterative bootstrapping framework that effectively combines neural models, rule-based weak supervision methods, and unlabeled data.

Our proposed framework draws inspiration from *co-training* (Blum and Mitchell, 1998), but it has its own unique approach. Like co-training, **CoAug** aims to combine two distinct inductive biases in limited labeled data settings. Unlike co-training, instead of improving two models that use different feature sets individually by bootstrapping labels from each other, **CoAug** accomplishes the same goal by using two models that use different forms of supervision to expand the same label set. Additionally, in each iteration of **CoAug**, both classifiers are trained with the predictions made by both models, rather than just one. Our choice allows the framework to function from really small initial training sets for the individual models.

We evaluate our approach on four named entity recognition datasets that span general and science domains. Our results indicate that (a) **CoAug** consistently improves performance over self-training rule-augmentation and few-shot models while being highly precise, (b) utilizing stronger pre-training for the neural models leads to improved performance of models in our framework.

In summary, our contributions are as follows:

- We present **CoAug**, a co-augmentation framework that leverages both rule-augmentation and label-augmentation approaches for NER.
- Experimental results show that **CoAug** can perform better than prior rule-based methods on four datasets in two domains.
- We provide a brief analysis of factors that contribute towards the success of **CoAug**.

2 CoAug

In this work, we consider a setting where we have access to an initial set of seed rules, $\mathcal{S}$, and a large unlabeled corpus, $\mathcal{U}$, to perform the named entity recognition task. Applying the rules, $\mathcal{S}$, on $\mathcal{U}$ provides the initial set of labeled examples, $\mathcal{L}$, to train models in our framework.

Our framework, **CoAug** (short for **Co-Aug**mentation), iteratively improves the performance of two models by leveraging the bootstrapped predictions on unlabeled data by each model. Given that prior work in low-resource NER focuses on two parallel tracks of rule-augmentation and few-shot learning methods that do not interact with each other, we instantiate **CoAug** with a rule-augmentation model and a few-shot model to leverage the best of both paradigms. We refer to these components of our framework as Rule Augmenter and Label Augmenter (Figure 1). In the subsections below, we describe the Rule Augmenter and Label Augmenter modules.

2.1 Rule Augmenter

Algorithm 1 TALLOR

Require: $\mathcal{U} = \{x_{1:N}\}$ unlabeled examples
Require: $\mathcal{R} = \{\mathcal{S}\}$ rules initialized with seed rules
Require: $\mathcal{C} = \{c_{1:M}\}$ candidate rules

Initialize: $\mathcal{L} = \{\}$
for t in $(1, \ldots, T)$ **do**
  // **Apply rules to get weak-label set**
  $\mathcal{W} = \text{RULEAPPLIER}(\mathcal{R}, \mathcal{U})$
  // **Filter accurate examples**
  $\mathcal{W} = \text{LABELSELECTOR}(\mathcal{W})$
  $\mathcal{L} = \mathcal{L} \cup \mathcal{W}$
  $\mathcal{U} = \mathcal{U} \setminus \mathcal{L}$
  // **Train NEURAL NER MODEL**
  $M \leftarrow \text{TRAIN}(M, \mathcal{L})$
  // **Label using NEURAL NER MODEL**
  $\mathcal{L}_M \leftarrow \text{PREDICT}(M, \mathcal{U})$
  // **Select High-precision Rules**
  $\mathcal{R}_S \leftarrow \text{RULESELECTOR}(\mathcal{L}_M, \mathcal{C})$
  $\mathcal{C} = \mathcal{C} \setminus \mathcal{R}_S$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_S$
end for

The primary function of the Rule Augmenter is to automatically learn labeling rules from unlabeled data and use them to generate weak labels for training a neural model. In this work, we instantiate the rule augmenter module using the **TALLOR** framework. Accordingly, our rule augmenter has

the following subcomponents: (a) RULE APPLIER that applies rules over unlabeled data to generate weak labels, (b) LABEL SELECTOR that filters the most accurate examples based on the similarity of averaged token-level BERT (Devlin et al., 2019) representations of proposed entities to the representations of previously identified entities of the same label in the training set, (c) NEURAL NER MODEL that is trained on the accurate instances and proposes new entities in the unlabeled data that can be used to develop new rules, and (d) RULE SELECTOR that scores candidate labeling rules and selects high-precision rules that satisfy the predictions from the NEURAL NER MODEL. We summarize the iterative process of automatic rule identification by **TALLOR** in Algorithm 1.

2.2 Label Augmenter

The Label Augmenter module consists of a NEURAL MODEL that learns to perform entity recognition with minimal supervision and LABEL SELECTOR that selectively adds the weak labels proposed by the NEURAL MODEL into the training set for the next iteration.

Algorithm 2 Label Augmenter

Require: $\mathcal{U} = \{x_{1:N}\}$ unlabeled examples
Require: $\mathcal{L} = \{\mathcal{S}\}$ rules initialized with seed rules
Require: $\beta_0, \beta 1$ ▷ initial threshold and increment

Initialize: $\mathcal{L} = \mathcal{R}(\mathcal{U})$
for t in $(1, \ldots, T)$ **do**
  // Train NEURAL MODEL
  $M \leftarrow \text{TRAIN}(M, \mathcal{L})$
  // Label using NEURAL MODEL
  $\mathcal{L}_M \leftarrow \text{PREDICT}(M, \mathcal{U})$
  // Select Examples Using Adaptive Threshold
  $\mathcal{L}_M \leftarrow \text{LABELSELECTOR}(\mathcal{L}_M, \beta_0 + t \times \beta_1)$
  $\mathcal{L} = \mathcal{L} \cup \mathcal{L}_M$
end for

In this work, we experiment with two instantiations of the NEURAL MODEL using recent few-shot NER models, namely, **ProtoBERT** and **QuIP**. We use an adaptive threshold for the Label Selector to filter out low-quality, weakly labeled instances. Initially, we add 20% of the proposed instances from the Neural Model to the training set. Then, as the model becomes more confident in its predictions over iterations, we gradually increase the proportion of instances incorporated, with a 5% increase per iteration. We summarize the label augmenter algorithm in Algorithm 2.

We provide an outline for the **CoAug** algorithm in Algorithm 3. In each training iteration, we alternatively train the Rule Augmenter and Label Augmenter models. Different from co-training (Blum and Mitchell, 1998), in **CoAug**, the Rule-Augmenter (Label-Augmenter) utilizes the examples that have been labeled by the Rule-Augmenter (Label-Augmenter) and the Label-Augmenter (Rule-Augmenter) to improve its entity recognition performance over iterations.

Algorithm 3 **CoAug** algorithm

Require: $\mathcal{U} = \{x_{1:N}\}$ unlabeled examples
Require: $\mathcal{R} = \{\mathcal{S}\}$ rules initialized with seed rules
Require: `RuleAugmenter` M_1, `LabelAugmenter` M_2
  $\mathcal{L} = \mathcal{R}(\mathcal{U})$
  for t in $(1, \ldots, T)$ **do**
    $\mathcal{U} = \mathcal{U} \setminus \mathcal{L}$
    // Training the Rule Augmenter section
    $M_1 \leftarrow \text{TRAIN}(M_1, \mathcal{L})$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \text{UPDATERULES}(M_1)$
      ▷ Select high-precision rules
    $\mathcal{L} = \mathcal{L} \cup \mathcal{R}(\mathcal{U})$ ▷ Add examples after applying rules
    // Training the Label Augmenter section
    $M_2 \leftarrow \text{TRAIN}(M_2, \mathcal{L})$
    $\mathcal{W} \leftarrow \text{HIGHCONFWEAKLABEL}(M_2, \mathcal{U})$
      ▷ Select high-confident weak-labels
    $\mathcal{L} = \mathcal{L} \cup \mathcal{W}$
  end for

3 Experiments

3.1 Experimental Settings

We evaluate our framework on four popular datasets that are composed of two science-domain and two general-domain datasets. Following Li et al. (2021), we utilize the training data without labels as our unlabeled data. Further, for all experiments, we use a set of 20 initial seed rules. These rules specify highly frequent entities for each category within a dataset.

BC5CDR (Li et al., 2016) contains 1,500 PubMed abstracts with manual annotations for disease and chemical entity mentions. The abstracts are split equally among train, dev, and test sets (500/500/500).

NCBI-Disease (Doğan et al., 2014) contains 793 PubMed abstracts with manual annotations for disease entity mentions. The abstracts are split as 593/100/100 for train, dev, and test sets.

CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) contains about 20,744 sentences from Reuters news articles. We split the data into 14,987/3,469/3,685 sentences for the train, dev, and test set. Additionally, for our experiments, we only

Method	BC5CDR	CoNLL-2003	NCBI-Disease	WikiGold
TaLLOR (Li et al., 2021)	$59.4_{(3.2)}$	$50.3_{(9.6)}$	$39.3_{(1.5)}$	$23.7_{(4.3)}$
ProtoBERT (Tänzer et al., 2022)	$33.1_{(3.5)}$	$47.3_{(2.9)}$	$25.5_{(4.4)}$	$37.3_{(3.8)}$
CoAug (**TaLLOR** + **ProtoBERT**)	$\mathbf{64.4_{(1.5)}}$	$\mathbf{65.0_{(0.8)}}$	$\mathbf{46.8_{(3.5)}}$	$\mathbf{50.6_{(2.1)}}$
QuIP (Jia et al., 2022)	$64.9_{(1.7)}$	$70.6_{(3.7)}$	$\mathbf{75.3_{(0.7)}}$	$43.6_{(2.3)}$
CoAug (**TaLLOR** + **QuIP**)	$\mathbf{65.9_{(1.5)}}$	$\mathbf{76.8_{(2.0)}}$	$50.5_{(4.9)}$	$\mathbf{51.8_{(2.8)}}$

Table 1: Test set F1 scores of models on BC5CDR, CoNLL-2003, NCBI-Disease, and WikiGold datasets. Numbers reported in each cell correspond to the mean and standard deviation of five runs. Bold numbers indicate the best numbers relative to component models. From the results, we can see that utilizing **CoAug** can help models outperform their single-model counterparts by a large margin in most cases. Our strongest model, **CoAug** with **QuIP**, outperforms the weak-supervision baseline, **TaLLOR**, by a large margin and **QuIP** on three out of four datasets.

consider the Person, Location, and Organization entities[2] following Li et al. (2021).

WikiGold (Balasuriya et al., 2009) contains 1,696 sentences from Wikipedia articles with annotations for Person, Location, and Organization entity categories similar to CoNLL2003. We split the dataset into 1,142/280/274 sentences for the train, dev, and test sets.

We evaluate two instantiations of the **CoAug** framework where the Rule Augmenter uses **TaLLOR**, and the Label Augmenter uses either **ProtoBERT**/**QuIP**. For baselines, our main experiments compare **CoAug** against **TaLLOR**, self-trained **ProtoBERT**, and self-trained **QuIP**. Our code is implemented in Pytorch (Paszke et al., 2019) using the Huggingface library (Wolf et al., 2020). For the Rule Augmenter section, all experimental hyperparameters follow that from Li et al. (2021). Notably, we use the same hyperparameters for the NCBI-Disease, and WikiGold datasets as Li et al. (2021) did for BC5CDR and CoNLL2003. For science-domain datasets, we utilize `SciBERT-base` (Beltagy et al., 2019) as the base for the **ProtoBERT** model and `BERT-base` (Devlin et al., 2019) otherwise. We do not make any such distinctions for **QuIP** as it is a specially fine-tuned `RoBERTa-large` (Liu et al., 2019) model designed to perform well on extraction-based tasks (Jia et al., 2022). We report the hyperparameters used for all experiments in more detail in Appendix C.

3.2 Results and Analysis

3.2.1 Main Results

Table 1 reports the test set F1 scores for all models on each of the four datasets. We observe that **CoAug** with **QuIP**/**ProtoBERT** outperforms **TaLLOR** on all 4 datasets substantially (average F1 on WikiGold for **CoAug** is more than 2× **TaLLOR**). Further, we also observe that utilizing the co-augmentation framework as opposed to self-training also aids models to produce similar results more reliably, as indicated by the variance of the results (in 3 out of 4 datasets). Further, we also observe that utilizing larger few-shot models, such as **QuIP** (which has a `RoBERTa-large` base), is complementary to our framework and continues to push the NER performance further. On comparing with **QuIP**, we observe that **CoAug** with **QuIP** performs better on 3 out of 4 datasets.

However, on the NCBI-Disease dataset, we observe that **QuIP** outperforms **CoAug** by a considerable margin. On analysis, we identify that **QuIP** adds too many incorrect instances during the initial few iterations for this dataset. Consequently, the rule augmenter selects rules that lose precision, and the overall quality of examples in **CoAug** deteriorates. Nonetheless, since entity recognition for this dataset is hard for **TaLLOR** as well, we observe some improvement from using **CoAug**. Future work should look to address the issue of controlling candidates from neural models in order to maintain the reliability of the high-precision set.

In Figure 2, we identify that the success of **CoAug** over high-precision rule-augmentation approaches, such as **TaLLOR**, lies in its ability to identify more instances in the unlabeled that improve precision as well as recall over **TaLLOR**.

3.2.2 Effect of Task-aligned Pre-training

In this subsection, we analyze the contribution of pre-training strategies towards the performance of **CoAug**. Specifically, we ablate the effect of changing the pre-training initialization from **QuIP** to that of `RoBERTa-large`, the base model for **QuIP**. As shown in Table 2, the performance of **CoAug** with `RoBERTa-large` lags far behind the performance

[2] skipping entities from the Miscellaneous category.

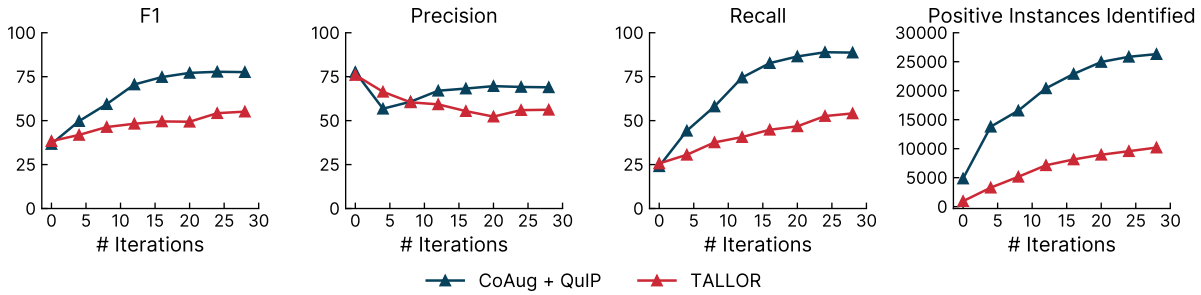

Figure 2: Validation statistics for **CoAug** +**QuIP** and **TaLLOR** over iterations for one run of the CoNLL2003 dataset. **CoAug** +**QuIP** identifies more high-precision positive instances from the unlabeled data than **TaLLOR** while also maintaining high precision.

Model	BC5CDR	CoNLL2003
CoAug (**TaLLOR** + **RoBERTa**)	$45.6_{(0.1)}$	$64.4_{(0.2)}$
CoAug (**TaLLOR** + **QuIP**)	$\mathbf{65.9}_{(1.5)}$	$\mathbf{76.8}_{(2.0)}$

Table 2: Test set performance of **CoAug**-**QuIP** and **CoAug**-**RoBERTa** models on the BC5CDR and CoNLL 2003 datasets. Results reported are the mean and standard deviation of five runs. **QuIP** initialization provides a boost to the performance of **CoAug**.

of **CoAug** with **QuIP**. On BC5CDR, we observe that the **CoAug** with `RoBERTa-large` performs poorly in comparison to **TaLLOR** as well. This indicates that any form of task-aligned pre-training, such as **QuIP**, can help design NER models for a diverse domain of tasks which corroborates some of the earlier work in task-adaptive pre-training (Gururangan et al., 2020).

4 Conclusion

In this work, we introduce **CoAug**, a co-augmentation framework that utilizes unlabeled data to train rule-augmentation and neural-augmentation models to become better NER taggers. Our results on datasets from two domains demonstrate the effectiveness of **CoAug** for low-resource domains. Our analysis reveals that **CoAug** is able to perform better than weak-supervision methods like **TaLLOR** because of an ability to find more positive instances while maintaining high precision. Further analysis shows the importance of factors such as the strength of pre-training that can contribute towards the success of models in domain-specific datasets.

Limitations

We observe that although **CoAug** outperforms baselines on multiple datasets, it is still prone to errors that emerge from the bootstrapping process. Specifically, our framework utilizes models to augment weak labels to the training set, and if the proposals are extremely noisy, training on noisy examples in future iterations will further exacerbate the ability of the framework to identify entities with high precision. Incorporating constraints to preserve the quality of the pseudo-labeled data (Shrivastava et al., 2012) is an exciting direction for future work in low-resource named-entity recognition.

Ethics Statement

All our experiments are performed over publicly available datasets. We do not use any identifiable information about crowd workers who provide annotations for these datasets. Neither do we perform any additional annotations or human evaluations in this work. We do not foresee any risks using **CoAug** if the inputs to our model are designed as per our procedure. However, our models may exhibit unwanted biases that are inherent in pre-trained language models. This aspect is beyond the scope of the current work.

Acknowledgement

We would like to express our appreciation to Haibo Ding for the insightful discussions and helpful suggestions during the initial phases of this project.

References

Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. 2017. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pages 273–282. PMLR.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the*

2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web), pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2004. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 89–96.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Avrim Blum and Tom. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Sally A. Goldman and Yan Zhou. 2000. Enhancing supervised learning with unlabeled data. In *ICML'00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 327–334.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2022. Question answering infused pre-training of general-purpose contextualized representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 711–728, Dublin, Ireland. Association for Computational Linguistics.

Hyunjae Kim, Jaehyo Yoo, Seunghyun Yoon, Jinhyuk Lee, and Jaewoo Kang. 2021. Simple questions generate named entity recognition datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6220–6236.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Hunter Lang, Monica Agrawal, Yoon Kim, and David Sontag. 2022. Co-training improves prompt-based learning for large language models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11985–12003.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6634–6647, Online. Association for Computational Linguistics.

Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, and Zhe Feng. 2021. Weakly supervised named entity tagging with learnable logical rules. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4568–4581, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. In *Database (Oxford)*, volume 2016, pages 1–10.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR 2019*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32. Curran Associates, Inc.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5570–5578.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2012. Constrained semi-supervised learning using attributes and comparative attributes. In *European Conference on Computer Vision*, pages 369–383. Springer.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tao Yu and Shafiq Joty. 2021. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8501, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Related Work

Weakly-supervised NER. Utilizing distant supervision in the form of knowledge bases or typed lexicons dates back to the work of Mintz et al. (2009). However, obtaining pre-defined lexicons for all domains is challenging. Therefore, more recent work proposes to use manually-defined labeling functions to provide weak labels for documents at scale (Bach et al., 2017). Safranchik et al. (2020); Lison et al. (2020) have proposed improved techniques for leveraging such labeling functions to derive weak labels for entities. However, exhaustively defining labeling functions to identify entities can be a cumbersome task, even for domain experts. Hence, **TALLOR** (Li et al., 2021) introduces an automatic technique for learning rules through an iterative approach of proposing weak labels and new rules for entity extraction. More recently, GeNER (Kim et al., 2021) utilizes DensePhrases (Lee et al., 2021) to query Wikipedia for documents that contain entities from desired categories. However, Wikipedia may not contain enough information for new emerging domains or even new

languages. In contrast, **CoAug** can be applied in such situations as well using few rules and some unlabeled datasets.

Co-training. In co-training (Blum and Mitchell, 1998), given two views of an input that are conditionally independent of each other given the true label, classifiers learned over both views can be improved by bootstrapping the performance of each view iteratively with unlabeled data. Some recent studies suggest, however, that the conditional independence assumption of the views can be relaxed when the models are "different enough" (Balcan et al., 2004; Goldman and Zhou, 2000). Within language processing methods, co-training has been used for cross-lingual adaptation (Yu and Joty, 2021) and improving prompt-based techniques (Lang et al., 2022). In our work, we improve named entity recognition with a combination of rule-augmentation and neural-augmentation techniques.

B Background

Re-iterating, we consider a setting where we have access to an initial set of seed rules, $\mathcal{S}$, and a large unlabeled corpus, $\mathcal{U}$, to perform the named entity recognition tasks. Applying the initial set of rules on $\mathcal{U}$ provides an initial set of labeled examples, $\mathcal{L}$, to train models in our framework.

B.1 TALLOR

Our work primarily builds on top of the **TALLOR** framework introduced in Li et al. (2021). In **TALLOR**, a neural model is trained on $\mathcal{L}$ to provide weak labels for the potential entities present in $\mathcal{U}$. Based on the computed weak labels, a Rule Selector module proposes new labeling rules that align well with the weak labels while maintaining high precision for entity recognition. Finally, the newly proposed rules are used to label more examples in $\mathcal{U}$, and the process is repeated over many iterations. At the end of training, **TALLOR** is evaluated by the neural model's performance on the test set of the corresponding task. For more details on **TALLOR**, we refer the reader to (Li et al., 2021).

C Experiment Hyperparameters

Across all datasets, we limit the span of the entities to 5 tokens.

Following Li et al. (2021), the neural model in the Rule-Augmentation model is initialized with a `BERT-base`/ `SciBERT-base` model depending on the domain of the dataset. During training, we use a minibatch size of 32 with the Adam optimizer (Kingma and Ba, 2015), a learning rate of $2e{-}5$, and perform gradient clipping (clipped at norm of 5.0) to stabilize training.

Dataset	Category	Question Prompt
BC5CDR	Chemical	*What is a chemical compound?*
	Disease	*What is a disease?*
CoNLL2003/ WikiGold	Person	*Who is a person?*
	Location	*What is a location?*
	Organization	*What is an organization?*
NCBI-Disease	Disease	*What is a disease?*

Table 3: Question prompts used for token classification head initialization in **QuIP**.

For the Label-Augmenter model, we utilize two models: **ProtoBERT** and **QuIP**. Since these models have different characteristics, we utilize a different set of hyperparameters to fine-tune each model for our task. Specifically, for the **ProtoBERT** model, we use the AdamW (Loshchilov and Hutter, 2019) optimizer, a learning rate of $1e{-}4$, and apply weight decay of $1e{-}2$ for all parameters except the layer-norm weights. For **QuIP**, we follow the recommendations from Jia et al. (2022) and adopt a learning rate of $2e{-}5$ with the AdamW optimizer for fine-tuning. Further, we initialize the token prediction head for the NER task using question prompt embeddings from this model. The set of questions we use for the different datasets has been summarized in Table 3.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.