Contrastive Bootstrapping for Label Refinement

Shudi Hou[†] and **Yu Xia**[†] and **Muhao Chen**[‡] and **Sujian Li**[†]
[†]Key Laboratory of Computational Linguistics, MOE, Peking University
[‡]University of Southern California
{housd, yuxia, lisujian}@pku.edu; muhaoche@usc.edu

Abstract

Traditional text classification typically categorizes texts into pre-defined coarse-grained classes, from which the produced models cannot handle the real-world scenario where finer categories emerge periodically for accurate services. In this work, we investigate the setting where fine-grained classification is done only using the annotation of coarse-grained categories and the coarse-to-fine mapping. We propose a lightweight contrastive clustering-based bootstrapping method to iteratively refine the labels of passages. During clustering, it pulls away negative passage-prototype pairs under the guidance of the mapping from both global and local perspectives. Experiments on NYT and 20News show that our method outperforms the state-of-the-art methods by a large margin.[1]

1 Introduction

Traditional text classification often categorize into a set of coarse-grained classes, which falls short in real-world scenarios where finer categories emerge. To this end, coarse-to-fine text classification is introduced (Mekala et al., 2021), which performs fine-grained classification given only annotation of coarse-grained categories and the coarse-to-fine mapping. Then, it finetunes a pre-trained language model for each coarse prototype.[2] However, this two-step method could be sub-optimal. For example, it is vulnerable to the noise which is propagated and accumulated through the pipeline. Besides, it requires finetuning and saving a pre-trained language model for each coarse prototype which is heavyweight.

To this end, we propose a lightweight bootstrapping method based on contrastive clustering to iteratively refine the labels of passages.[3] To be more specific, the method starts with an epoch of warm-up on the weakly-labeled dataset. During warm-up, it pulls away negative passage-prototype pairs under the guidance of the mapping from both global and local perspectives, *i.e.*, coarse inter-cluster and fine inter-cluster perspectives. After the warm-up, the distances between clusters are not significant which causes misclassification. Instead of continuing training on the weakly-labeled dataset which might greatly increase the noise (Figure 1(b)), we perform a bootstrapping process which finetunes the model on the selected dataset and updates the selected dataset by the finetuned model alternately. To mitigate the noise, we propose a selection strategy to identify high-quality pairs in terms of similarity and distinction. To further boost our method, we adopt a modified similarity metric from (Lample et al., 2018) and use the gloss knowledge to augment the prototype representation. As shown in (Figure 1(c)), the resulting clusters are well separated with less noise.

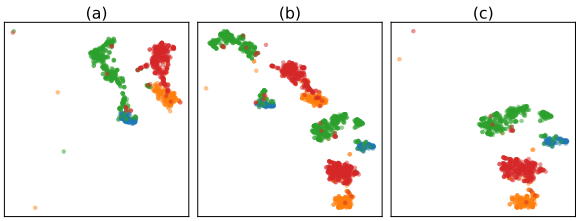

Figure 1: Passages with "Arts" coarse prototype on NYT dataset. Colors are used to denote different fine prototypes. (a) warm-up. (b) bootstrapping w/o selection strategy. (c) bootstrapping w/ selection strategy.

Our contributions are summarized as follows:

- We propose a lightweight bootstrapping method based on contrastive clustering to ad-

[1]Code is available at https://github.com/recorderhou/contrastive_bootstrapping_label_refinement

[2]We use prototype and category interchangeably.

[3]We focus on passage-level classification as it is consistent with prior studies (Mekala et al., 2021). Though, without loss of generality, the studied problem as well as the proposed method can be extended to classifying natural language text in other granularities.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 976–985
July 9-14, 2023

dress the problem of coarse-to-fine text classification.

- Our method outperforms the state-of-the-art methods on two widely-used datasets. Further analysis verifies the effectiveness of our proposed techniques.

2 Proposed Method

This section describes the technical details of the proposed method, starting with the task description.

2.1 Task Description

We follow the task definition of coarse-to-fine text classification in previous work (Mekala et al., 2021). Given n passages $\{p_1, ..., p_n\}$ with their corresponding coarse-grained labels $\{c_1, ..., c_n\}$, along with the coarse-to-fine mapping $\mathcal{T}$, our goal is to assign a fine-grained label to each passage. The key notations used in our paper are defined as follows: (1) $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_m\}$ denotes the coarse prototypes. (2) $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k\}$ denotes the fine prototypes. (3) $\mathcal{T} : \mathcal{C} \rightarrow \mathcal{F}$ denotes the coarse-to-fine mapping, a surjective mapping which separates $\mathcal{F}$ into $|\mathcal{C}|$ non-overlapping partitions. (4) $\mathcal{S}_{pf} = \mathcal{T}(c_i)$ denotes the fine-grained candidate prototype of p_i, which is also dubbed as p for simplicity. (5) $\mathcal{S}_{nf} = \mathcal{F}/\mathcal{S}_{pf}$ denotes fine prototypes not belonging to $\mathcal{T}(c_i)$. (6) $\mathcal{S}_{nc} = \mathcal{C}/c_i$ denotes coarse prototypes in $\mathcal{C}$ other than c_i.

2.2 Our Method

Training Process As illustrated in Figure 2, we start with an epoch of warm-up, during which we optimize two contrastive losses $\mathcal{L}_{global}$, $\mathcal{L}_{local}$ on the weakly-labeled dataset and only the $\mathcal{L}_{global}$ on the unlabeled dataset. The two contrastive losses are detailed in the following paragraphs. Then, we conduct several epochs of bootstrapping with the above model. At each bootstrapping step, we first select a small set of passages on which labels are predicted with high confidence by the model. Then, we finetune the model on the selected dataset with the same losses as warm-up. We repeat the finetuning and the selection alternately.

Initial Weak Supervision Following previous work, we consider samples that exclusively contain the label surface name as their respective weak supervision. More details can be referred to the prior study.

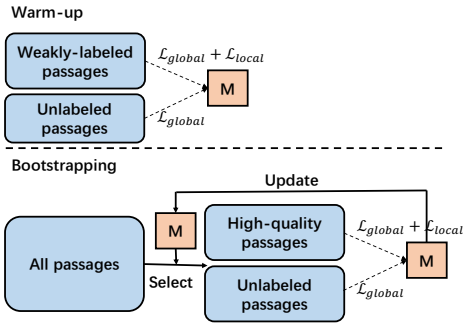

Figure 2: Illustration of our training process.

Passage and Prototype Representation We encode passages $\{p_1, ..., p_n\}$ and all prototypes $\mathcal{C} \cup \mathcal{F}$ into the same embedding space with a pretrained language model. The resulting passage representation and prototype representation are denoted as $\boldsymbol{p}$ and $\boldsymbol{l}$ respectively. During the training process, the prototype representations are dynamically updated to fit the current passage representations. Specifically, we use the last hidden representation of [CLS] as their representations.

Similarity Metric Cosine similarity is often used to measure semantic similarity of embedding representations. However, in high-dimensional spaces, some "hub" vectors may be close to many other vectors while some other vectors are instead being isolated. For example, a passage's representation $\boldsymbol{p}$ may get high cosine with a large number of labels in S_{pf} due to such hubness issues. In this case, a high similarity score does not necessarily lead to a high discrepancy among labels. Selecting a highly-scored label from the hub as the seed is potentially detrimental to our pairing-based method. Inspired by cross-domain similarity local scaling (Lample et al., 2018), we adopt a modified similarity metric $c(\boldsymbol{p}, \boldsymbol{l})$ to prevent passage vectors from becoming hubs:

$$c(\boldsymbol{p}, \boldsymbol{l}) = \cos(\boldsymbol{p}, \boldsymbol{l}) - KNN(\boldsymbol{p}) \quad (1)$$

$$KNN(\boldsymbol{p}) = \frac{1}{K} \sum \max_{\boldsymbol{l} \in \mathcal{F}} K\{\cos(\boldsymbol{p}, \boldsymbol{l})\} \quad (2)$$

where $KNN(.)$ denotes K nearest neighbors.

Warm-up Viewing a passage as an anchor, we expect that its semantic similarity to the correct fine-grained prototype should be closer than any other fine-grained candidate prototypes. We regard the distance in the representation space as the similarity. Specifically, we optimize the following

margin ranking loss:

$$\mathcal{L}_{global} = \frac{1}{|S_{pf}|} \sum_{\substack{l \in S_{pf} \\ l' \in S_{nf}}} \max\{c(\boldsymbol{p}, \boldsymbol{l}) - c(\boldsymbol{p}, \boldsymbol{l'}) + \gamma, 0\} \tag{3}$$

where γ is a hyper-parameter denoting the margin. We use all fine candidate prototypes in S_{pf} as positive examples and randomly sample the same number of prototypes from S_{nf} as negative examples. We view this loss as a global loss to cluster samples according to their coarse labels (Figure 3).

For instances labeled in the initial weak supervision stage, we adopt another margin ranking loss:

$$\mathcal{L}_{local} = \max\{sec\_max - c(\boldsymbol{p}, \boldsymbol{l}) + \sigma, 0\} \tag{4}$$

$$sec\_max = \max_{l' \in S_{pf}, l' != l} c(\boldsymbol{p}, \boldsymbol{l'}) \tag{5}$$

We regard this loss as a local loss to cluster samples according to their fine-grained labels (Figure 1 (a)).

Bootstrapping After the warm-up, representations show an inclination to form clusters. Yet, the distances between them are not significant enough to separate the classes. To further get compact clusters, we perform bootstrapping which finetunes the model on the selected dataset and updates the selected dataset by the finetuned model alternately. Instead of using the initial weak supervision which might greatly increase the noise as observed, we propose a selection strategy to select high-quality passage-prototype pairs. Specifically, we assign a pseudo label to each passage by their similarity (Eq.(6)). Apart from **similarity**, we assume high-quality pairs should also be **discriminative** (Eq.(7)):

$$l = arg \max_{l \in S_{pf}} c(\boldsymbol{p}, \boldsymbol{l}) \tag{6}$$

$$c(\boldsymbol{p}, \boldsymbol{l}) - \max_{l' \in S_{pf}, l' != l} c(\boldsymbol{p}, \boldsymbol{l'}) > \beta \tag{7}$$

where β is a threshold updated at each epoch. We construct a confident set CS with top $r\%$ pairs satisfying these two conditions. We update β with the lowest similarity in CS. Then, we optimize Eq.(4) and Eq.(3) on CS and the rest passages accordingly.

Gloss Knowledge Since the surface names alone can not well represent the semantics of labels, we enrich them with external semantic knowledge. To be more specific, we select the first two sentences in each surface name's first Wikipedia webpage to augment the original surface name with a predefined template (Table 3). We adopt the format of "template, surface name, gloss" and use the last hidden representation of [CLS] as their representation.

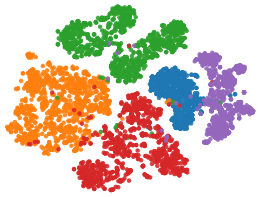

Figure 3: Passage representations after warm-up on 20News dataset. Colors are used to denote different coarse prototypes.

Prediction It is worth noticing that applying our similarity metric $c(\boldsymbol{p}, \boldsymbol{l})$ do not change the relative ranking among labels in $\mathcal{S}_{pf}$ compared with the cosine similarity. For simplicity, we use cosine similarity for prediction.

$$l = arg \max_{l \in S_{pf}} \cos(\boldsymbol{p}, \boldsymbol{l}) \tag{8}$$

3 Experiments

In this section, we describe the experimental evaluation for the proposed method.

3.1 Datasets and Metrics

For a fair comparison with prior work, we use the same hierarchical datasets used by We report both Macro-F1 and Micro-F1 for evaluation on the following two datasets.

The 20 Newsgroups (20News) The passages in 20News was organized into 5 coarse-grained newsgroups and 20 fine-grained newsgroups corresponding to different topics (Table 2). Passages in 20News were partitioned evenly across the 20 different fine-grained newsgroups.[4] Following (Mekala et al., 2021), we omitted the 3 miscellaneous newsgroups ("misc.forsale," "talk.politics.misc" and "talk.religion.misc") and expanded the abbreviation to full words.

The New York Times (NYT) This dataset contains 5 coarse-grained topics and 25 subtopics (Table 2). The NYT dataset is highly skewed with the coarse-grained topic "sports" containing more than 80% passages.

3.2 Main Results

We compare our model with the previous work (Mekala et al., 2021), as well as several zero-shot weakly supervised text classification methods

[4] http://qwone.com/~jason/20Newsgroups/

	NYT		20News	
	Mi-F1(%)	Ma-F1(%)	Mi-F1(%)	Ma-F1(%)
LOT-Class	79.26	63.16	56.38	54.80
X-Class	58.15	60.50	52.95	53.47
C2F	89.23	84.36	75.77	75.24
C2F w/ our select ⋆	89.64	82.72	77.20	76.41
Ours	**92.64**	**89.90**	77.64	77.22
w/o fine	91.15 (↓ 1.49)	84.90 (↓ 5.00)	74.34 (↓ 3.30)	73.78 (↓ 3.44)
w/o bootstrap	89.49 (↓ 3.15)	82.50 (↓ 7.40)	76.01 (↓ 1.63)	75.46 (↓ 3.30)
w/o gloss	89.91 (↓ 2.73)	80.48 (↓ 9.42)	72.68 (↓ 4.86)	70.31 (↓ 6.91)
w/o select	87.56 (↓ 5.08)	81.98 (↓ 8.02)	**79.74** (↑ 2.10)	**79.21** (↑ 1.99)
w/o similarity	89.25 (↓ 3.39)	82.44 (↓ 7.46)	61.21 (↓ 16.43)	54.76 (↓ 22.46)
w/ Manhattan similarity †	33.45 (↓ 59.19)	39.47 (↓ 50.43)	41.83 (↓ 35.81)	36.50 (↓ 40.72)
w/ Euclidean similarity ‡	92.46 (↓ 0.18)	89.17 (↓ 0.73)	72.11 (↓ 5.53)	70.65 (↓ 6.57)

Table 1: Results on NYT and 20News. "⋆" equips C2F with our selection strategy. "†" replaces our similarity metric with Manhattan distance. "‡" replaces our similarity metric with Euclidean distance.

(Wang et al., 2021b; Meng et al., 2020a) following previous works. We reproduce them using their implementation.[5][6][7]

As shown in Table 1, our method outperforms the baselines by 5.67% in Micro-F1 and 5.54% in Macro-F1 on the NYT dataset, as well as 3.97% in Micro-F1 and 3.04% in Macro-F1 on 20News dataset.

3.3 Analysis

To verify the effectiveness of different model components , we conduct ablation studies to test each of those.

Effect of Bootstrapping The "w/o bootstrap" results in Table 1 report the performance with warm-up only. These results are consistently lower than those with bootstrapping. Specifically, bootstrapping improves the warm-up by 3.15% Micro-F1, 7.40% Macro-F1 and 1.63% Micro-F1, 3.30% Macro-F1 on NYT and 20News respectively. Figure 1(a)(c) shows passage representations are more separated from each other.

Effect of Selection Strategy We replace the selection strategy in bootstrapping with the initial weakly-labeled samples. From the "w/o bootstrap" results in Table 1, we can see that, our selection strategy brings an improvement of 4.26% Micro-F1, 7.46% Macro-F1 on NYT. It is better to use the seed dataset on 20News. We hypothesize that this observation is because the seed dataset has a more balanced label distribution than our selected high-quality samples on 20News. We also incorporate our selection strategy to the C2F baseline in the bootstrapping stage. As shown in Table 1 row "C2F w/ our select," this strategy improves the performance of C2F by 1.43% Micro-F1, 1.17% Macro-F1 on 20News and 0.41% Micro-F1 on NYT, exhibiting the effectiveness of our strategy.

Effect of Similarity Metric We replace our similarity metric with the cosine similarity. From Table 1 "w/o similarity" we can see that, our similarity metric brings along an improvement of 3.39% in Micro-F1, 7.46% in Macro-F1 on NYT, and 16.43% in Micro-F1 and 22.46% in Macro-F1 on 20News. From Figure 4, we can see that 63% of samples belonging to the "Law Enforcement" prototype are misclassified using the cosine similarity. However, 18% are misclassified using our similarity metric, verifying its effectiveness. Besides, results for "w/ Manhattan similarity" and "w/ Euclidean similarity" show that alternating cosine similarity in $c(\boldsymbol{p}, \boldsymbol{l})$ causes performance drops of 35.81% (5.53%) in Micro-F1, 40.72% (6.57%) in Macro-F1 and 50.19% (0.18%) in Micro-F1, 50.43% (0.73%) in Macro-F1 on 20News and NYT data, further proving the effectiveness of our similarity metric.

Effect of Gloss Knowledge We remove the gloss knowledge and use the label surface name only. Comparing the "w/o gloss" results in Table 1 with the full-setting ones, we observe that the gloss knowledge brings an improvement of 2.73% in Micro-F1, 9.42% in Macro-F1 on NYT and 4.86% in Micro-F1, 6.91% in Macro-F1 on 20News. Figure 5 further shows the effect of gloss knowledge on different prototypes.

[5] https://github.com/yumeng5/LOTClass
[6] https://github.com/ZihanWangKi/XClass
[7] https://github.com/dheeraj7596/C2F

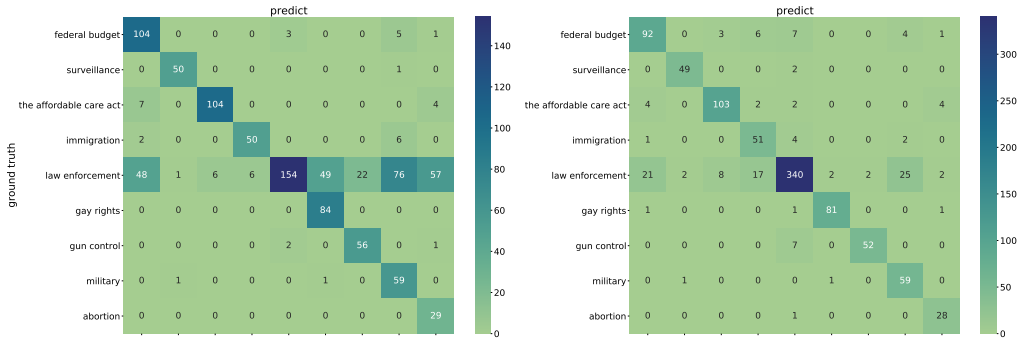

Figure 4: Confusion matrix on "Politics" coarse prototype. Our similarity metric (right) outperforms cosine similarity (left) by 12.25% Macro-F1 and 16.68% Micro-F1 under "Politics."

Extending to the setting without coarse-to-fine mapping We extend our method to the setting without the coarse-to-fine mapping. In other words, the only supervision is the gold coarse labels. We modify $\mathcal{L}_{global}$ as follows:

$$\mathcal{L}_{c\_global} = \max\{c(\boldsymbol{p}, \boldsymbol{l_c}) - c(\boldsymbol{p}, \boldsymbol{l'_c}) + \gamma, 0\} \quad (9)$$

where we use the golden coarse label $\boldsymbol{l_c}$ as the positive example and randomly sample one coarse label $\boldsymbol{l'_c}$ from $\mathcal{S}_{nc}$ as the negative example. The "w/o fine" results in Table 1 show that the performance does not degrade much when the association between coarse and fine-grained labels does not exist, showing the feasibility of our method in a more general setting.

4 Related Work

Previous works in weakly supervised text classification have explored different kinds of weak supervision. (1) a set of related keywords. (Mekala and Shang, 2020) augment and disambiguate the initial seed words with contextualized and highly label-indicative keywords. (Meng et al., 2020b) identify keywords for classes by querying replacements for class names using BERT and pseudo-labels the documents by heuristics with the selected keywords. (2) a few labeled documents. (Tang et al., 2015) represent the labeled documents and different levels of word co-occurrence information as a large-scale text network. (Meng et al., 2018) propose a pseudo-document generator that leverages the seed labeld documents to generate pseudo-labeled documents for model pre-training. (3) label surface names. (Wang et al., 2021b) propose an adaptive representation learning method to obtain label and document embedding, and cluster them to pseudo-label the corpus. Our setting is different from theirs in that we use coarse-grained annotation to improve the fine-grained text classification.

Contrastive learning (He et al., 2020; Chen et al., 2020; Khosla et al., 2020) aims at learning representations by contrasting the positive pairs and negative pairs. In NLP, existing works can be primarily categorized into two distinct streams. Unsupervised contrastive learning seeks to contrast grouped or perturbed instances to generate more robust representation of unlabeled textual data (Gao et al., 2021; Wei et al., 2021; Kim et al., 2021; Wang et al., 2021a). On the contrary, supervised contrastive learning (Suresh and Ong, 2021; Zhou et al., 2021; Yu et al., 2021; Huang et al., 2022) is label-aware and seeks to create representations for differently labeled data with more discrepancy. Our work has shown that supervised contrastive learning incorporating label names, with minimal external knowledge, improves the model's performance in label refinement.

5 Conclusion

In this paper, we study the task of coarse-to-fine text classification. We propose a novel contrastive clustering-based bootstrapping method to refine the label in an iterative manner. Experiments on two real-world datasets for coarse-to-fine text classification verify the effectiveness of our method. Future work could consider extending this method to other fine-grained decision-making tasks that could potentially benefit from coarse-grained labels, such as various kinds of lexical semantic typing tasks (Huang et al., 2022). Another meaningful direction is to consider incorporating other partial-label learning techniques (Zhang et al., 2016) that are relevant to coarse-to-fine prediction tasks.

Limitations

Our paper has the following limitations: (1) In real-world applications, the label hierarchy may be more than two levels. It is worth extending our method to such a setting and empirically verifying it. (2) Our selection strategy simply takes top $r\%$ confident samples, which might result in class imbalance problem. Alleviating the imbalance problem may further improve our performance. We leave them as future work.

Acknowledgement

We appreciate the reviewers for their insightful comments and suggestions. We would like to express our gratitude to the authors of the C2F paper (Mekala et al., 2021) for their collective effort in open-sourcing the dataset and code. Their released materials played a vital role in our research.

Shudi Hou, Yu Xia and Sujian Li were supported by National Key R&D Program of China (No. 2020AAA0109703). Muhao Chen was supported by the National Science Foundation of United States Grant IIS 2105329, a subaward of the INFER Program through UMD ARLIS, an Amazon Research Award and a Cisco Research Award.

References

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

James Y. Huang, Bangzheng Li, Jiashu Xu, and Muhao Chen. 2022. Unified semantic typing with meaningful label inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Seattle, United States. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. 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 2528–2540, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 583–594, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *proceedings of the 27th ACM International Conference on information and knowledge management*, pages 983–992.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020a. Text classification using label names only: A language

model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020b. Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1165–1174.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021a. CLINE: Contrastive learning with semantic negative examples for natural language understanding. 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 2332–2342, Online. Association for Computational Linguistics.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021b. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077, Online. Association for Computational Linguistics.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1335–1344, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Dataset Statistics

We list the statistics of the datasets in Table 2.

B Templates

We list the templates used in Table 3.

C Effect of gloss knowledge on different prototypes

We show the confusion matrix over all fine prototypes in Figure 5.

D Implementation Details

We use RoBERETa-base (Liu et al., 2019) as the encoder. The models are trained on one GeForce RTX 3090 GPU. We set the batch size as 8. We do one epoch of warmup and four epochs of bootstrapping. We use the predictions from the last epoch as the final predictions. We use AdamW (Loshchilov and Hutter, 2017) as the optimizer. r is set as 15 for NYT and 1 for 20News. γ and σ are set as 0.05 for both NYT and 20News. We run our model 3 times using different random seeds. We used t-SNE (Pedregosa et al., 2011; Buitinck et al., 2013) for the visualization in this paper.

E Selection of r

We select the value of r from set {1, 5, 10, 15, 20}. For each coarse prototype $\mathcal{C}_i$, we calculate the ratio of initial weak supervision $W_{\mathcal{C}_i}$ in category $\mathcal{C}_i$ to the total number of instance $I_{\mathcal{C}_i}$ in $\mathcal{C}_i$, we denote the ratio as $R_{\mathcal{C}_i} = W_{\mathcal{C}_i}/I_{\mathcal{C}_i}$. After that, we select the r closest to $\min_{\mathcal{C}_i \in \mathcal{C}}\{R_{\mathcal{C}_i}\}$. As shown in Table 4a and Table 4b, the minimal $R_{\mathcal{C}_i}$ in NYT dataset is 13.43%, closest to 15, while the minimal $R_{\mathcal{C}_i}$ in 20News dataset is 2.05%, closest to 1.

Dataset	Passage	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{F}\rvert$	Coarse Prototype	Fine Prototype
20News	16468	5	17	computer, politics, recreation, religion, science	graphics, windows, ibm, mac, x window, mideast, guns, autos, motorcycles, baseball, hockey, christian, atheism, encryption, electronics, medicine, space
NYT	11744	5	26	arts, business, politics, science, sports	dance, music, movies, television, economy, energy companies, international business, stocks and bonds, abortion, federal budget, gay rights, gun control, immigration, law enforcement, military, surveillance, the affordable care act, cosmos, environment, baseball, basketball, football, golf, hockey, soccer, tennis

Table 2: Dataset Statistics.

Dataset	Template
NYT	1 : The news is about, 2 : The news is related to, 3 : The topic of this passage is
20News	1 : The topic of this post is , 2 : They are discussing , 3 : This post mainly talks about

Table 3: Three variants of templates used to concatenate the gloss knowledge and the surface name. The first template is best for NYT and the third template is best for 20News.

$\mathcal{C}_i$	$W_{\mathcal{C}_i}$	$I_{\mathcal{C}_i}$	$R_{\mathcal{C}_i}$ (%)
arts	184	1043	17.64
business	132	983	13.43
politics	216	989	21.84
science	42	90	46.67
sports	1890	8639	21.88

(a) Ratio of the initial weak supervision in NYT

$\mathcal{C}_i$	$W_{\mathcal{C}_i}$	$I_{\mathcal{C}_i}$	$R_{\mathcal{C}_i}$ (%)
computer	100	4880	2.05
politics	56	1850	3.03
recreation	924	3976	23.24
religion	150	1976	8.35
science	100	3951	2.53

(b) Ratio of the initial weak supervision in 20News

Table 4: Ratio of the initial weak supervision

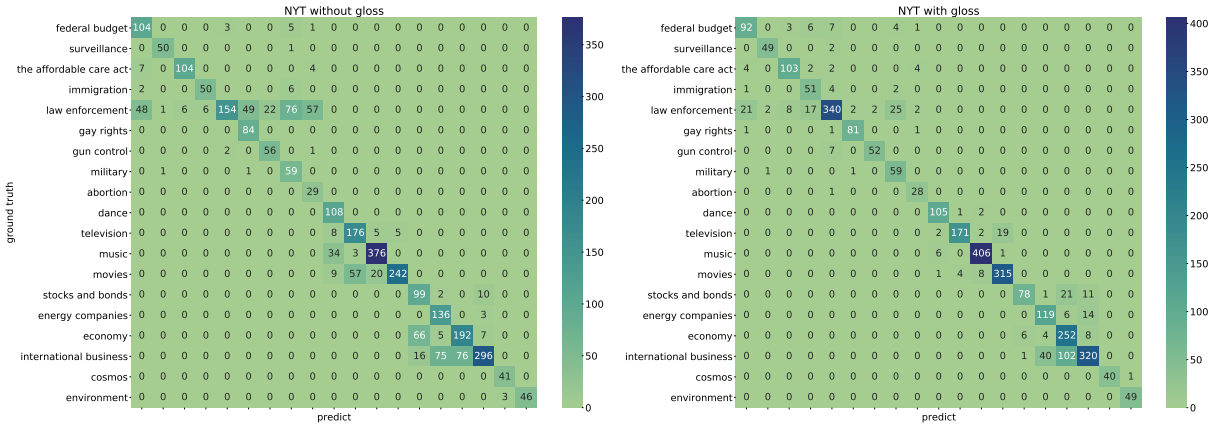

Figure 5: Confusion matrix over all fine prototypes without (left) and with (right) the gloss knowledge.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Conclusion

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☑ Did you use or create scientific artifacts?

3.1

☑ B1. Did you cite the creators of artifacts you used?
3.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We obtain the license and will not distribute it .

☒ 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)?
We use the dataset following their intended use.

☐ 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?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. 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.
Appendix A

C ☑ Did you run computational experiments?

3

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

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?
Appendix D and E

☑ 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?
Appendix D

☑ 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.)?
Appendix D

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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.