Coarse2Fine: Fine-grained Text Classification on Coarsely-grained Annotated Data

Dheeraj Mekala[1] **Varun Gangal**[2] **Jingbo Shang**[1,3,*]

[1] Department of Computer Science and Engineering, University of California San Diego, CA, USA

[2] Language Technologies Institute, Carnegie Mellon University, PA, USA

[3] Halıcıoğlu Data Science Institute, University of California San Diego, CA, USA

[1,3] {dmekala, jshang}@ucsd.edu   [2] vgangal@cs.cmu.edu

Abstract

Existing text classification methods mainly focus on a fixed label set, whereas many real-world applications require extending to new fine-grained classes as the number of samples per label increases. To accommodate such requirements, we introduce a new problem called coarse-to-fine grained classification, which aims to perform fine-grained classification on coarsely annotated data. Instead of asking for new fine-grained human annotations, we opt to leverage label surface names as the only human guidance and weave in rich pre-trained generative language models into the iterative weak supervision strategy. Specifically, we first propose a label-conditioned fine-tuning formulation to attune these generators for our task. Furthermore, we devise a regularization objective based on the coarse-fine label constraints derived from our problem setting, giving us even further improvements over the prior formulation. Our framework uses the fine-tuned generative models to sample pseudo-training data for training the classifier, and bootstraps on real unlabeled data for model refinement. Extensive experiments and case studies on two real-world datasets demonstrate superior performance over SOTA zero-shot classification baselines.

1 Introduction

In traditional text classification problems, the label set is typically assumed to be fixed. However, in many real-world applications, new classes, especially more fine-grained ones will be introduced as the data volume increases. One commonly used method is to extend the existing label set to a label hierarchy by expanding every original coarse-grained class into a few new, fine-grained ones, and then assign a fine-grained label to each document. Using the directory structure for a set of files in computer as an example (see in Figure 1), people
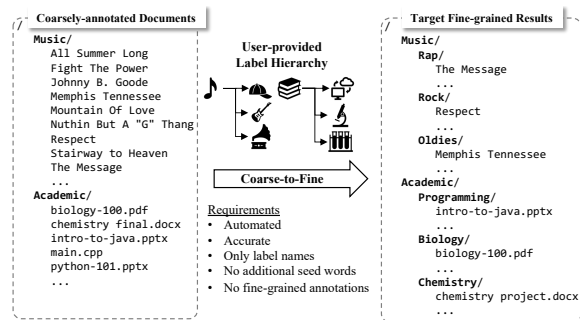

Figure 1: A visualization of our coarse-to-fine problem.

usually start organizing the files in a coarse-grained fashion like "Music" and "Academics". Once the number of files in each of these coarse-grained directories increases, the categorization serves little purpose. Therefore, we would like to create new fine-grained sub-directories inside coarse-grained directories like {"rap", "rock", "oldies"} for "music" and similarly for "academics". However, the process of assigning these files into fine-grained sub-directories typically begins with almost no supervision for fine-grained labels.

To accommodate such requirements, in this paper, we introduce a new, important problem called coarse-to-fine grained classification, which aims to perform fine-grained classification on coarsely annotated data without any fine-grained human annotations. There has been prior research on performing text classification using extremely weak supervision, i.e., only label surface names as source of supervision. For example, X-Class (Wang et al., 2021) learns class-aligned document representations to generate pseudo-labels and LOT-Class (Meng et al., 2020) assumes replacements of label surface names in a sentence are related to the classes and leverages pre-trained language models to extract those words. Note that, coarse-to-fine setting differs from generic zero-shot text classification in terms of having additional coarse supervision and a pre-conceived label hierarchy,

---

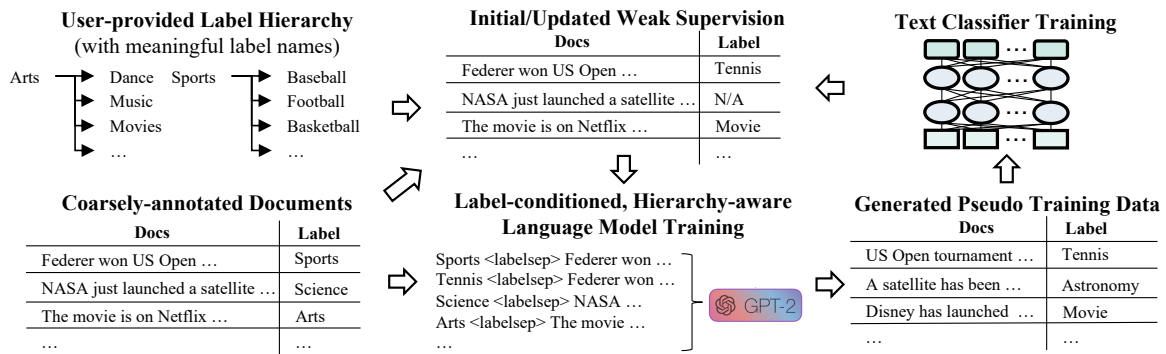* Jingbo Shang is the corresponding author.

Figure 2: A visualization of our C2F framework.

though the final label set is available in either case. And also, coarse-to-fine setting is different from hierarchical classification. We have no supervision for fine-grained labels other than the label names whereas the few shot hierarchical setting can have a few samples for fine-grained labels. Therefore, we want to capture the coarse-grained supervision and label hierarchy available to perform fine-grained classification.

In this paper, we propose a novel framework C2F as illustrated in Figure 2. In the absence of fine-grained human annotations, it uses fine-grained label surface names as weak-supervision signals and leverages pre-trained language models as data generators. Similar to previous work, we first generate weak supervision from the whole corpus by assuming label surface names as their respective strong label-indicators. For two iterations, C2F fine-tunes a language model based on weak supervision and trains a classifier based on generated pseudo-training data to refine weak supervision. We observe that raw weak supervision usually has a highly-skewed label distribution, especially at the beginning, because the popularity of the label names varies. Since we have no prior knowledge about the underlying label distribution, to avoid significant deviations from that distribution, we opt to draw a balanced, weakly annotated subset through a stratified sampling before any model training. We propose to fine-tune language models in a label-conditioned, hierarchy-aware manner. Specifically, we inform the language models with label information by adding the label surface names at the beginning of each document. We further incorporate a regularization objective into the fine-tuning process that captures the constraints derived from the label hierarchy. Facilitated by this fine-tuning process, we then generate pseudo-training data for each fine-grained label and train a classifier. Next,

using this fine-grained classifier's predictions over the coarsely annotated data, we select the samples with a high predicted probability for each respective fine-grained label.

We conduct experiments on two real-world datasets containing both coarse and fine-grained labels. The results demonstrate the effectiveness of our framework in leveraging a label hierarchy and a rich pre-trained language model to perform fine-grained text classification with no supervision. Via thorough ablation, we isolate separate benefits accrued, initially just from using the label-conditioned, fine-tuned language model in the weak supervision pipeline, and the later incremental benefit once we incorporate our proposed regularization objective into the language model fine-tuning.

To the best of our knowledge, we are the first to work on the coarse-to-fine text grained classification, which aims to perform fine-grained classification on coarsely annotated data without any fine-grained annotations. It is also worth mentioning that C2F is compatible with almost any generative language model and text classifier. Our contributions are summarized as follows:

- We develop a label-conditioned fine-tuning formulation for language models to facilitate conditional corpus generation.
- We devise a regularization objective based on the coarse-fine label constraints derived from the label hierarchy to be consistent with the pre-conceived label hierarchy.
- We conduct extensive experiments to demonstrate the superiority of C2F.

Reproducibility. We will release the code and datasets on Github[1].

---

[1] https://github.com/dheeraj7596/C2F

584

1.1 Another Example Application

Another example task which motivates how our framework could be deployed (besides the aforementioned directory example) is Intent Classification in Task-based Dialog (Chen et al., 2019; Kumar et al., 2019; Schuster et al., 2019; Gangal et al., 2020). This is often seen as a hierarchical classification problem (Gupta et al., 2018), with domains (e.g. movie, airline, shopping search) and intents (e.g. book-movie vs check-reviews, book-flight vs add-meal-options) forming the higher and lower levels of the label hierarchy. For a real-world task based dialog system (e.g. Alexa), there's always a need over time to keep introducing both new domains (e.g. cruise, medical FAQ) and intents (order-popcorn, flight-entertainment) - as both data volume increases and the backend capabilities of the system expand.

2 Problem Formulation

The input of our problem contains: (1) A tree-structured label hierarchy $\mathcal{T}$ with coarse-grained labels $\mathcal{C}$ at the first level and fine-grained labels $\mathcal{F}$ as their children. The m coarse-grained classes are named $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$, and k fine-grained classes are named as $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k\}$. All these class names are in natural language (e.g., words or phrases) and assumed to be informative; and (2) a collection of n text documents $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ and their corresponding coarse-grained labels $\{c_1, c_2, \ldots, c_n\}$.

We record the mapping from each fine-grained class to its corresponding coarse-grained parent class as $f^{\uparrow}: \mathcal{F} \rightarrow \mathcal{C}$. The fine-grained classes in a coarse-grained label are represented by the coarse-to-fine mapping $f_{\downarrow}: \mathcal{C} \rightarrow \mathcal{P}(\mathcal{F})$, where $\mathcal{P}(\cdot)$ is the powerset operator, which generates the set of all subsets. In this problem, each coarse class maps to a non-empty subset of fine classes, and all these subsets of fine-classes taken together are mutually non-overlapping and exhaustive.

We aim to build a high-quality document classifier from these inputs, assigning a fine-grained class label $\mathcal{F}_j \in f_{\downarrow}(c_i)$ to each document $\mathcal{D}_i \in \mathcal{D}$.

3 Our C2F Framework

As visualized in Figure 2, C2F aims to build a text classifier that can assign fine-grained labels to a set of coarsely-annotated documents based on only the label surface names and their hierarchical relations. In the absence of fine-grained human annotations, it uses fine-grained label surface names as weak-supervision signals and leverages a pre-trained language model as data generators. Following an iterative process, C2F fine-tunes a language model based on weak supervision. This fine-tuned language model is used to generate pseudo training data to train a fine-grained text classifier. Based on the classifier's predictions, we select highly probable samples for each fine-grained class and repeat this process for one more iteration by replacing weak supervision with these samples. This bootstrapping increases the quality of weak supervision by eliminating the mislabeled samples and improves the performance of text classifier as we show later in our case studies.

Our major contributions lie in how to better incorporate the label names and their hierarchical relations into the language model and therefore generate more high-quality psuedo training data. Our framework is compatible with any generative language model and we choose GPT-2 (Radford et al., 2019) in our implementation. We feed label names to the language model through a label-conditioned formulation. We further incorporate a regularization objective into the fine-tuning process that captures the constraints derived from the label hierarchy. The key components of C2F are discussed in detail in the following sections.

3.1 Initial Fine-grained Weak Supervision

We assume that user-provided label surface names are of high quality and are strong indicators for their respective classes, following the state-of-the-art weakly supervised text classification methods that only rely on label surface names (Wang et al., 2021; Meng et al., 2020). This assumption is intuitive and valid because there is no guidance other than class names from user and we expect them to be of high quality and indicative of the categories.

Ideally, the posterior probability of a document belonging to a class after observing the presence of strong indicators should be close to 1. Therefore, we consider samples that *exclusively* contain the label surface name as its respective weak supervision. Mathematically, let $W(\mathcal{F}_j)$ denote weak supervision of fine-grained class $\mathcal{F}_j$:

$$W(\mathcal{F}_j) = \{\mathcal{D}_i | \mathcal{D}_i \cap f_{\downarrow}(c_i) = \{\mathcal{F}_j\}\}$$

where $\mathcal{D}_i \cap f_{\downarrow}(c_i)$ returns a set of fine-grained label names under the coarse-grained class c_i that appear

in the document $\mathcal{D}_i$. When this set only contains $\mathcal{F}_j$, it means "exclusive" to other fine-grained labels. This "exclusiveness" could help us improve the precision of the initial weak supervision. Note that, it is implied that $\mathcal{F}_j \in f_\downarrow(c_i)$.

We observe that the initial weak supervision obtained usually has a highly-skewed label distribution, because the popularity of the label names varies. This difference in distribution could bias the generative language model towards the majority label and might affect the quality of generated samples, which in turn, would affect the performance of the text classifier. To address this problem, as there is no other prior knowledge, we opt to draw a balanced, weakly annotated subset through a stratified sampling before any model training. In other words, we make the size of weak supervision uniform for all labels, equal to the size for the minority label.

3.2 Tailored Language Model Training

In this section, we describe our label-conditioned, hierarchy-aware language model training formulations that facilitates conditional corpus generation. Specifically, we continuously train a pre-trained language model to capture the distribution $P(D|l)$, where D is a document and l is a (coarse or fine) label surface name. Thus, this model can generate pseudo-training documents for fine-grained labels, when we plug in fine-grained label surface names.

3.2.1 Label-Conditioned Generation

Before we describe our formulation, we briefly introduce GPT-2 and its pre-training objective.

GPT-2. GPT-2 is a large, pre-trained left-to-right language model which exhibits strong performance with minimal in-task fine-tuning on many generation tasks, such as dialog (Zhang et al., 2020) and story generation (See et al., 2019). Its strong zero-shot ability across tasks stems from its pre-training on the vast and diverse WebText corpus ($\approx$8M documents), besides the good inductive bias of its transformer-based architecture. GPT-2 is trained on standard language modeling objective to maximize the likelihood of a document D as follows:

$$\mathcal{L}(D) = \sum_i \log P(w_i|w_{i-1}, \ldots, w_1; \Theta)$$

where $P(\cdot)$ is modeled with a transformer-based architecture with parameters Θ.

To continuously train GPT-2 in a label-conditioned way, one has to maximize $P(D|l)$ in-

stead of $P(D)$. We designate the label surface names as the special token sequences and append them in the beginning to their respective documents with another special token `<labelsep>` separating the label sequence and document. For example, a sample document "*Messi plays for FC Barcelona*" belonging to "*soccer*" is modified to "*soccer* `<labelsep>` *Messi plays for FC Barcelona*". Therefore, our objective is to maximize $\mathcal{L}(D|l)$ defined below:

$$\mathcal{L}(D|l) = \sum_i \log P(w_i|w_{i-1}, \ldots, w_1; l; \Theta)$$

Note that, the l here could be the label surface name of a coarse-grained or fine-grained class. One can view our formulation as asking the label token sequence to play the role of prompt and the document D to be the continuation, thus facilitating conditional corpus generation.

During the continuous training process, we have access to both the gold, coarse-grained labels and weak, fine-grained labels. Examples included in weak supervision give rise to two label-conditioned documents — one by prefixing with the coarse-grained, gold label and the other with the weak, fine-grained one (due to the "exclusiveness" in the initial weak supervision). Those not in the weakly supervised set only give rise to the first kind. Since there is no conflict between these two labels, we simply treat a document as belonging independently to either of them.

3.2.2 Hierarchy-Aware Regularization

Our label-conditioned generation treats both fine- and coarse-grained labels as prompts and does not use any information from the hierarchy. Therefore, we propose to add a regularization to the language model with constraints derived from hierarchy.

Intuitively, fine-grained labels are more specific to coarse-grained labels, and therefore, when generating the same document conditioned on its gold fine-grained label, it should have a higher probability than that conditioned on its coarse-grained label. We believe the same intuition is applicable to the high-quality weak supervision. Therefore, we seek to enforce the constraint while continuously training on weak supervision. Specifically, a document should be more likely given its fine-grained (weak) label rather than its coarse-grained label. Mathematically,

$$P(\mathcal{D}_i|\mathcal{F}_j) > P(\mathcal{D}_i|c_i), \forall \mathcal{D}_i \in W(\mathcal{F}_j)$$

where $W(\mathcal{F}_j)$ is the weak supervision for fine-grained label $\mathcal{F}_j$. Note that, it is implied from $W(\mathcal{F}_j)$ that $\mathcal{F}_j \in f_\downarrow(c_i)$.

This inequality can be expressed in the form of a margin between $P(\mathcal{D}_i|\mathcal{F}_j)$ and $P(\mathcal{D}_i|c_i)$, which can be implemented in practice through an additional Hinge loss term:

$$HL(\mathcal{D}_i, \mathcal{F}_j) = \max(0, \log P(\mathcal{D}_i|c_i) - \log P(\mathcal{D}_i|\mathcal{F}_j) + \epsilon)$$

where ϵ is a positive constant.

We incorporate this hierarchy-aware regularization into the final objective function as follows:

$$\mathcal{O} = \sum_{\mathcal{D}_i \in \mathcal{D}} \mathcal{L}(D_i|c_i)$$
$$+ \sum_{\mathcal{F}_j} \sum_{\mathcal{D}_i \in W(\mathcal{F}_j)} \mathcal{L}(\mathcal{D}_i|\mathcal{F}_j) - \lambda HL(\mathcal{D}_i, \mathcal{F}_j)$$

The final optimization aims to maximize $\mathcal{O}$.

3.3 Pseudo Training Data Generation, Text Classifier, & Weak Supervision Update

After continuously training the language model in a label-conditioned way, we generate the data for each fine-grained category. Specifically, we send the corresponding label surface name as the prompt to our language model, and it then generates samples for that respective class. Since we don't know the label distribution beforehand, we assume it's a balanced distribution and thus, avoiding inducing potential bias in the classifier. We generate twice the required documents divided equally among fine-grained labels. Specifically, for a fine-grained label $\mathcal{F}_j \in f_\downarrow(c)$, we generate $2\frac{N_c}{|f_\downarrow(c)|}$ documents, where N_c is the number of documents that belong to coarse-grained label c.

We train a text classifier over these generated documents and their corresponding fine-grained labels. Our framework is compatible with any text classifier and we use BERT(`bert-base-uncased`) (Devlin et al., 2019) classifier in our experiments.

After training the text classifier, we obtain fine-grained predictions and probability scores for all coarsely annotated documents $\mathcal{D}$. Finally, we bootstrap it on unlabelled data by *replacing* weak supervision $W(\mathcal{F}_j)$ by top-k predictions where $k = |W(\mathcal{F}_j)|$ in every fine-grained label $\mathcal{F}_j$ and repeat this process *one more time*. In our experiments, we observe that these top-$|W(\mathcal{F}_j)|$ predictions are of significantly higher quality than the initial weak supervision, thus improving the text classifier.

4 Experiments

In this section, we start with introducing datasets, compared methods, and experimental settings. Next, we present quantitative evaluation results of C2F together with all compared methods. In the end, we show qualitative studies to analyze different aspects of our C2F framework.

4.1 Datasets

We evaluate our framework on two hierarchical datasets where each document has one coarse-grained label and one fine-grained label. The dataset statistics are provided in Table 1. The details of these datasets are as follows:

- **The New York Times (NYT):** Following the previous work (Meng et al., 2018; Mekala and Shang, 2020; Wang et al., 2021) we experiment on the NYT dataset. It is a collection of news articles written and published by The New York Times. Each news article is classified into one of 5 coarse-grained genres (e.g., arts, sports) and 25 fine-grained categories (e.g., movies, baseball).
- **The 20 Newsgroups (20News):** The 20News dataset[2] is a collection of newsgroup documents partitioned widely into 6 groups (e.g., recreation, computers) and 20 fine-grained classes (e.g., graphics, windows, baseball, hockey). There are three miscellaneous labels (i.e., "misc.forsale", "talk.politics.misc", "talk.religion.misc"). As one can notice, their label names are about 'miscellaneous' and contain information of various types. Since these labels and label surface names have no focused meaning, we drop the documents annotated as these labels in our experiments.

4.2 Compared Methods

Since we aim to perform fine-grained classification with no fine-grained supervision, we compare our framework with a wide range of zero-shot and weakly supervised text classification methods described below:

- **Word2Vec** learns word vector representations (Mikolov et al., 2013) for all words in the corpus and consider the word vectors of label surface name vectors as their respective label representations. In the case of multi-word label descriptors, the embeddings of individual words are averaged. Each document is labeled with the most similar label based on cosine similarity.

---

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

587

Table 1: Dataset statistics

Dataset	$|\mathcal{D}|$	$|\mathcal{C}|$	$|\mathcal{F}|$	Coarse labels	Fine labels
NYT	11,744	5	26	arts, business, politics, science, sports	dance, music, movies, television, economy, energy companies, international business, stocks & 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
20News	16,468	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

- **WeSTClass** (Meng et al., 2018) assumes words and documents share a joint semantic space and model each class as a high-dimensional spherical distribution. Words are sampled from this learned distribution to create pseudo-training data over which a classification model is trained. This model is refined through self-training on unlabeled documents. This is the same as applying its hierarchical counterpart WeSHClass (Meng et al., 2019) individually on each coarse-grained class.
- **ConWea** (Mekala and Shang, 2020) is a seed-driven contextualized weak supervision framework. They leverage pre-trained language models to resolve interpretation of seed words and make the weak supervision contextualized.
- **LOTClass** (Meng et al., 2020) uses pre-trained language model like BERT (Devlin et al., 2019) to query replacements for class names and constructs a category vocabulary for each class. This is further used to fine-tune the language model on a word-level category prediction task and identifies potential classes for documents via string matching. A classifier is trained on this pseudo-labeled data with further self-training.
- **X-Class** (Wang et al., 2021) learns class-oriented document representations that make it adaptive to the user-specified classes. These document representations are aligned to classes through PCA + GMM, harvesting pseudo labels for a supervised classifier training.

We also compare **C2F** with its ablated variants. **C2F-NoHier** uses label-conditioned generation alone without the hierarchy-aware regularization. **C2F-Ind** and **C2F-Ind-NoHier** are run individually on each coarse-grained label c to assign a fine-grained label $\mathcal{F}_j \in f_\downarrow(c)$ and the predictions are accumulated at the end to compute aggregated results. However, **C2F-Ind** uses both label-conditioned generation with the hierarchy-aware regularization whereas **C2F-Ind-NoHier** uses label-conditioned generation alone. **C2F-1IT** is a BERT classifier

trained on initial fine-grained weak supervision. We also consider C2F with different generative LMs and classifiers. **C2F-GPT-BERT**, **C2F-GPT-LR** use GPT (Radford et al., 2018) as the generative language model and BERT, Logistic Regression as classifiers respectively.

For a fair comparison, we make coarse-grained annotated data available for all baselines and run them individually on each coarse-grained label c to assign a fine-grained label $\mathcal{F}_j \in f_\downarrow(c)$ and the predictions are accumulated at the end to compute aggregated results. We provide label surface names as seed words for seed-word-driven baselines like ConWea and WeSTClass.

We also present the performance of BERT in a supervised setting which is denoted as **BERT-Sup**. The results of BERT-Sup reported are on the test set which follows an 80-10-10 train-dev-test split.

4.3 Experimental Settings

While fine-tuning GPT-2, we experiment with learning rates $\alpha \in \{5e^{-5}, 5e^{-4}, 5e^{-6}\}$, with $\alpha = 5e^{-4}$ being found optimal, and continue the label-conditioned language model training for 5 epochs.

Generation from the model is done via nucleus sampling (Holtzman et al., 2020), with a budget of $p = 0.95$ and a length limit of 200 subwords. The prompt given for generation is simply the tag sequence corresponding to the intended fine-grained label of the sample to be generated. Since fine-grained class ratios are apriori unknown, an equal number of examples are sampled for each fine-grained class within the same coarse-grained class.

For the hierarchy-aware regularization, we set the hinge loss margin $\epsilon = \log 5$ and $\lambda = 0.01$. For hyperparameter selection of ϵ, we sweep over the sequence of values in $\{log n\}_{n=1}^{n=10}$. Further searching is done through two levels of binary search. The decision to initially sweep over values in logarithmic fashion is taken based on two intuitions: i) Larger jumps were found to skip over the domain of variation of epsilon too quickly ii) ϵ is essentially a margin on logarithmic probabilities.

Table 2: Micro and Macro f1 scores and their respective standard deviations on two datasets are presented. The statistical significance test results between C2F and all other baselines are showed in Appendix A. All the p-values are less than 10^{-15}, making the performance improvement over baselines significant.

	NYT		20 Newsgroup	
Methods	Mi-F_1(%)	Ma-F_1(%)	Mi-F_1(%)	Ma-F_1(%)
Word2Vec	32.50 (2.50)	17.50 (1.50)	11.03 (1.30)	11.03 (0.90)
ConWea	76.23 (0.97)	69.82 (0.54)	56.14 (0.76)	56.21 (0.32)
WeSTClass	73.96 (0.49)	65.03 (0.31)	55.46 (0.19)	55.53 (0.38)
LOTClass	15.00 (1.20)	20.21 (0.76)	34.18 (0.64)	33.63 (0.71)
X-Class	91.16 (0.56)	81.09 (0.39)	73.15 (0.23)	73.06 (0.12)
C2F	**92.62 (0.54)**	**87.01 (0.72)**	**77.50 (0.96)**	**77.57 (0.89)**
C2F-NoHier	90.44 (0.91)	85.50 (0.82)	76.27 (0.85)	76.13 (0.78)
C2F-Ind	91.60 (0.45)	86.82 (0.44)	74.62(0.96)	74.50 (0.97)
C2F-Ind-NoHier	90.95 (0.59)	85.75 (0.17)	74.59 (0.63)	74.48 (0.57)
C2F-1IT	82.10 (0.21)	58.12 (0.31)	71.06 (0.38)	70.94 (0.53)
C2F-GPT-BERT	89.09 (0.34)	85.07 (0.28)	75.10 (0.64)	75.02 (0.59)
C2F-GPT-LR	88.57 (0.48)	84.98 (0.53)	75.03 (0.63)	74.89 (0.69)
BERT-Sup	98.00 (0.27)	94.00 (0.57)	96.39 (0.43)	96.36 (0.72)

4.4 Quantitative Results

We evaluate our framework using Micro-f1(Mi-F_1) and Macro-f1(Ma-F_1) as performance metrics. The evaluation results of all methods run on three random seeds are summarized in Table 2 along with their respective standard deviations. We can observe that our proposed framework achieves superior performance compared to all other baselines. We discuss the effectiveness of C2F as follows:

- C2F demonstrates the best performance among all compared baselines. By utilizing the generative language model through label-conditioned fine-tuning and regularizing it with hierarchical hinge loss to leverage the hierarchy, it is able to generate good quality pseudo training data, which helped in achieving the best performance.
- C2F outperforms X-Class with a significant margin. X-Class doesn't take advantage of label hierarchy and requires class names to be one word whereas our framework has no such limitation and leverages rich language models to understand informative label surface names.
- We have to note the significantly low performance of LOTClass. LOTClass queries replacements of label surface names and consider those to be indicative of the label. This is a valid assumption for the coarse-grained classification but when the classes become fine-grained, the replacements may not be indicative of its respective class. For e.g., consider the sentence "I won a baseball game.". If "baseball" is replaced by "tennis", it is still a valid and meaningful statement but "tennis" is not indicative of "baseball". Therefore, LOTClass performs low in the fine-grained text classification task. Our framework

separates the weak supervision for each label initially and fine-tunes the language model in a label conditioned way. Therefore, it is able to distinguish between fine-grained labels as well.
- The comparison between C2F, C2F-Ind and C2F-NoHier, C2F-Ind-NoHier shows that the hinge loss helped in leveraging the constraints from hierarchy to improve the language model.
- The comparison between C2F and C2F-Ind shows that the fine-grained classification benefits from the hierarchical structure and joint training with other coarse-grained classes.
- We can observe that C2F perform significantly better than C2F-1IT. This shows that the fine-grained classification improves with bootstrapping, where the samples with high predicted probabilities are selected and used them as weak supervision for the next iteration.
- The comparison between C2F and C2F-GPT-BERT, C2F-GPT-LR shows that the performance improves with larger language models. This also demonstrates that C2F is compatible with different generative language models and classifiers.
- We observe that the performance of C2F is quite close to supervised method BERT-Sup, for e.g., on the NYT dataset. This demonstrates that C2F is quite effective in closing the performance gap between weakly supervised and supervised setting with just label surface names as supervision.

4.5 Performance increase with bootstrapping

The f1-scores of fine-grained labels in three coarse-grained labels "computer", "politics", "religion" across iteration-0 and iteration-1 are plotted in Fig 3. We see that performance increases significantly from iteration-0 (blue) to iteration-1 (red). We attribute this increase to our bootstrapping.

4.6 Sensitivity to ϵ

A potential concern with the experimental setup can be overtly high sensitivity of C2F to the hinge loss margin parameter, i.e ϵ. However, from the plot in Figure 4, we clearly see that F1 scores aren't drastically sensitive to epsilon - with standard deviations of 0.00515 and 0.00517 for Macro and Micro-F1 scores respectively.

4.7 Qualitative Analysis

Given a particular coarse label (say sports) and its data subsets $X = \{\mathcal{D}_i | c_i = \text{"sports"}\}$ and $X_f = \{\mathcal{D}_i | c_i = \text{"sports"}, f_i = f \in f_\downarrow(\text{"sports"})\}$, as a matter of post-hoc analysis,
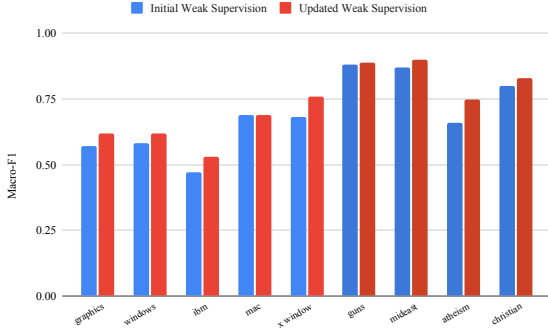
Figure 3: Performance increase in fine-grained text classifier from initial to updated weak supervision.
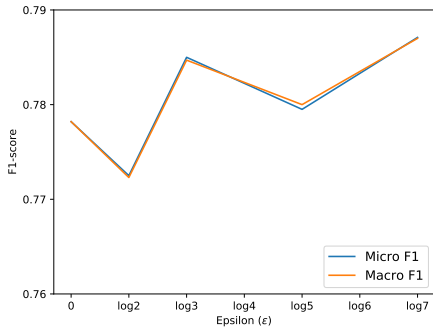

Figure 4: Micro and Macro-F1 scores vs ϵ on 20News.

we can compare three distinct "supervised" splits a classifier could've been trained on:

1. **Gold**: Data along with gold fine-grained labels, which is not actually available in our setting.
2. **C2F-Init**: This is the subset of X for which the initial weak supervision strategy assigns fine labels based on label surface names.
3. **C2F-Gen**: This is the data sampled from our trained language model as the generator for each of the respective fine-grained labels.

Which supervision is more apt from the purview of training? To answer this, we examine the entropy $H()$ of the word frequency distribution of the three datasets. Specifically, we examine reduction in value from entropy of the overall set $H(X)$ to the mean entropy on partitioning further by fine label, i.e $\bar{H}(X_f)$. The larger this drop, the more internally coherent are the label partitions.

As we can see from Figure 5, the drops $H(X) - \bar{H}(X_f)$ are greater for C2F-GEN compared to both C2F-INIT and GOLD, indicating that it produces more mutually discriminative examples than both of them. At the same time, we observe that overall entropy of C2F-GEN, i.e $H(Gen) = 6.631$ does not drastically differ in value from, though it is significantly lesser than, that of GOLD, $H(Gold) =$
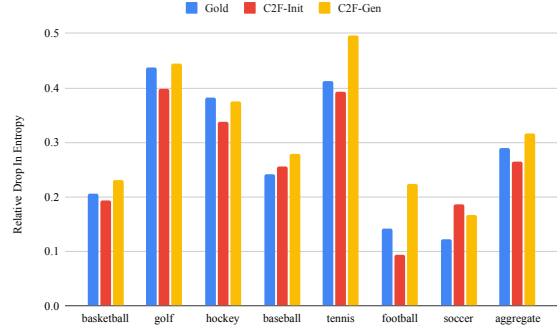

Figure 5: Relative drops in entropy $H(X) - H(X_{l_f})$ on splitting by fine label l_f, $\forall l_f \in f_{\downarrow}(\text{SPORTS})$, along with their aggregate. *C2F-Gen, C2F-Init, Gold* stands for the generator sampled, initial weakly supervised subset and the entire ground truth datasets respectively.

6.924, being 4.21% smaller. In summary, we see that C2F-GEN provides a more discriminative training signal without reducing example diversity.

A few samples of generated documents for fine-grained labels is shown in Table 3.

5 Related Work

We review the literature about different weakly supervised text classification methods.

There are three main sources of weak supervision: (1) a set of representative keywords for each class (Meng et al., 2018; Mekala and Shang, 2020; Mekala et al., 2020), (2) a few labeled documents (Tang et al., 2015; Miyato et al., 2017; Xu et al., 2017; Meng et al., 2018), (3) label surface names (Tao et al., 2018; Meng et al., 2020; Wang et al., 2021). Typically, weakly supervised text classification frameworks obtain pseudo-labeled data, train a classifier, and improve the classifier by bootstrapping over unlabeled data. Seed-driven frameworks obtain pseudo-labeled data from user-provided seed words. When a few labeled documents are provided as weak supervision, the above-mentioned pipeline similarly starts with these as pseudo-labeled data. In this paper, we focus on label surface names as the source of weak supervision. Along this line, Doc2Cube (Tao et al., 2018) expands label keywords from label surface names and performs multidimensional document classification by learning dimension-aware embedding; (Meng et al., 2020) identify keywords for classes by querying replacements for class names using BERT and pseudo-labels the documents by string matching with the selected keywords. (Wang et al., 2021) proposed an adaptive representation learning

Table 3: Example generated samples for fine-grained labels *hockey*, *basketball*, *cosmos*, and *economy* in NYT dataset and *autos*, *atheism*, *windows* in 20Newsgroup dataset.

Label	Sentence
hockey	the rangers' two injury-riddled offense and seven-game effort to defeat the florida panthers on saturday have prompted coach alain vigneault to call on the veteran right wing to start against the capitals. his message to the rookie goalie thomas vanek was that "he was fine today" and the rangers did not have to worry about his status in a game against another talented team.
basketball	the nets entered the n.b.a. season with such wide open space that even their most ardent supporters of the franchise might admit their disappointment. the season ended sunday with the los angeles lakers selecting the first-round draft pick, with the eventual no. 1 pick jason kidd, a dynamic penetrator who could challenge for a starting job, at least in a year.
cosmos	washington — the outermost reaches of space are barely perceptible from the surface, making it difficult to see the red planet, but a spacecraft based on this data could find something as yet elusive: an elusive galaxy that has just vanished.on thursday, kepler spacecraft, which is carrying back-to-back samples of solar starlight, will begin a journey that will lead the space agencyś curiosity.
economy	washington — the number of americans seeking unemployment benefits fell less than expected last week, the latest evidence that steady job growth is holding steady. jobless claims rose,000 to a seasonally adjusted,000, the labor department said on thursday. that is 146 more than previously reported.last week's drop in jobless claims came two months after steady increases in the previous week, when claims rose.
autos	re manual shift bigots in article (james bruder) writes another question that hasn't been addressed yet is how come the auto mated to the lever controlling selector is not mounted to the transmission? i would think that the mated to the shifter would be mounted in the passenger compartment. is this a problem with the manual transmission? and if so, is it a problem with the shifter's mounting point?
atheism	re why do people become atheists? in article, (kent sandvik) writes in article, (robert beauchaine) writes and i suppose i would have better evidence if i could. why would it be any different, for one thing? iṁ fairly new to this group, so perhaps this sort of question has already been asked, and answered before. but iv́e just started to think about it.
windows	re dos 6.0 in article 1qh61o, (russ sharp) wrote it's absolutely ludicrous for me to try and run dos 6.0 without the bloody help of at least 8 people. i've tried compiling it on several systems, and i've run it six times without a problem. dos 6.0 didn't mention a config.sys or anything else. there were a couple other windows' manuals which did mention about config.sys.

method for obtaining label and document embedding and these document embeddings are further clustered to pseudo-label the corpus. However, all these methods perform flat text classification. Although our method performs text classification using only fine-grained label surface names as supervision, we have coarse-grained annotated data and leverage it to improve fine-grained classification. There are a few methods that perform weakly supervised hierarchical classification (Meng et al., 2019; Zhang et al., 2021). However, our problem statement is different from hierarchical classification. We have coarse-grained annotated data and our framework utilizes it and label hierarchy to perform fine-grained text classification. Recently, (Hsieh et al., 2019) introduced coarse-to-fine weakly-supervised multi-label learning problem. However, they assume a few fine-grained labeled documents as supervision whereas we require only label surface names. Additionally, our framework is generative in nature i.e. instead of pseudo-labeling the corpus, we generate training data and train the classifier.

6 Conclusions and Future Work

Through this work, we introduced the task of coarse-to-fine grained classification and laid out its significance. Next, we showed the promise of incorporating pre-trained language models like GPT-2 into a weak supervision strategy which starts out with just label surface names. Finally, we showed a way to attune these models for our task even better, through explicit regularization based on coarse-fine label constraints which fall naturally out of our task definition. We outperform multiple SOTA zero-shot baselines on NYT and 20News, underscoring the utility both of incorporating pre-trained language models as well as task constraints.

We believe exploring newer ways of exploiting task agnostic knowledge sources and injecting task constraints into the weakly supervised learning process are promising avenues for future work.

7 Acknowledgements

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *EMNLP*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Cheng-Yu Hsieh, Miao Xu, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. 2019. A pseudo-label method for coarse-to-fine multi-label learning with limited supervision. *OpenReview*.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 1–10.

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.

Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. Meta: Metadata-empowered weak supervision for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8351–8361.

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, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. 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.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenReview*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861.

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.

Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance Kaplan, and Jiawei Han. 2018. Doc2cube: Allocating documents to text cube without labeled data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1260–1265. IEEE.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. 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.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical metadata-aware document categorization under weak supervision. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 770–778.

Appendix

A Stastical Significance Results

We perform a paired t-test between C2F and each of the other baselines on both datasets and the results are showed in Table 4. From these p-values, we can conclude that the performance improvement over baselines is significant.

Table 4: Statistical significance results.

Baseline	p-value NYT	p-value 20News
ConWea	8.34×10^{-131}	2.49×10^{-165}
WeSTClass	5.18×10^{-146}	1.97×10^{-166}
X-Class	6.45×10^{-71}	1.63×10^{-92}
C2F-NoHier	1.80×10^{-25}	2.33×10^{-55}
C2F-Ind	1.36×10^{-18}	7.92×10^{-110}
C2F-Ind-NoHier	3.46×10^{-24}	3.17×10^{-114}