# Towards Unifying the Label Space for Aspect- and Sentence-based Sentiment Analysis

**Yiming Zhang**
Zhejiang University
`yimingz@zju.edu.cn`
**Sai Wu**
Zhejiang University
`wusai@zju.edu.cn`

**Min Zhang**
Zhejiang University
`min_zhang@zju.edu.cn`
**Junbo Zhao (Jake)**
Zhejiang University
`j.zhao@zju.edu.cn`

## Abstract

The aspect-based sentiment analysis (ABSA) is a fine-grained task that aims to determine the sentiment polarity towards targeted aspect terms occurring in the sentence. The development of the ABSA task is very much hindered by the lack of annotated data. To tackle this, the prior works have studied the possibility of utilizing the sentiment analysis (SA) datasets to assist in training the ABSA model, primarily via pretraining or multi-task learning. In this article, we follow this line, and for the first time, we manage to apply the Pseudo-Label (PL) method to merge the two homogeneous tasks. While it seems straightforward to use generated pseudo labels to handle this case of label granularity unification for two highly related tasks, we identify its major challenge in this paper and propose a novel framework, dubbed as Dual-granularity Pseudo Labeling (DPL). Further, similar to PL, we regard the DPL as a general framework capable of combining other prior methods in the literature (Rietzler et al., 2019; Bai et al., 2020). Through extensive experiments, DPL has achieved state-of-the-art performance on standard benchmarks surpassing the prior work significantly (Liu et al., 2021).

## 1 Introduction

### 1.1 Aspect-based Sentiment Analysis

The aspect-based sentiment analysis (ABSA) task aims to recognize the sentiment polarities centered on the considered aspect terms occurring in the sentence. The establishment of the ABSA task echoes the long-standing literature of conventional sentence-level sentiment analysis (SA). For instance, as shown in Figure 1, a normal ABSA data annotation tags sentiment score on specific aspect terms in the sentence, like "surroundings" as positive and "food" as negative. Meanwhile, in the conventional sentence-based sentiment analysis, the whole sentence is labeled as negative at a coarser granularity.



Figure 1: Sentiment Analysis (SA) and Aspect-based Sentiment Analysis (ABSA). The sample on the above is the ABSA task, while the sample on the bottom is the SA task. Both tasks aim at analyzing the sentiments carried by the objects in the box.

Due to its much finer granularity, the annotation cost is significantly higher than its conventional counterpart. Essentially, many of the existing SA datasets (He et al., 2018) can be crawled and curated straightforwardly from the review websites such as Amazon[1] or Yelp[2]. The five-star rating system comes in handy to accomplish the annotation. Thus, the SA datasets are often presented at a large scale. By contrast, the ABSA annotation has no such "free lunch". It has to require human annotators to participate. Coupling with its higher complexity on labeling, the ABSA datasets are ubiquitously at considerably smaller scales (Pontiki et al.; He et al., 2018; Yu et al., 2021b). To this date, the available datasets for conventional sentiment analysis are generally larger to several orders of magnitude than the ABSA.

For instance, the commonly used ABSA benchmark SemEval 2014 task 4 has less than 5000 samples (Pontiki et al.), while there are 4,000,000 sentences in the Amazon Review dataset[3] for SA. Due to the similarity between the SA task and the ABSA task, it is natural to use SA datasets as auxiliary datasets for the ABSA task (He et al., 2018). Most, if not all, previous work has focused on pretraining and multi-task learning methods (He et al., 2018,
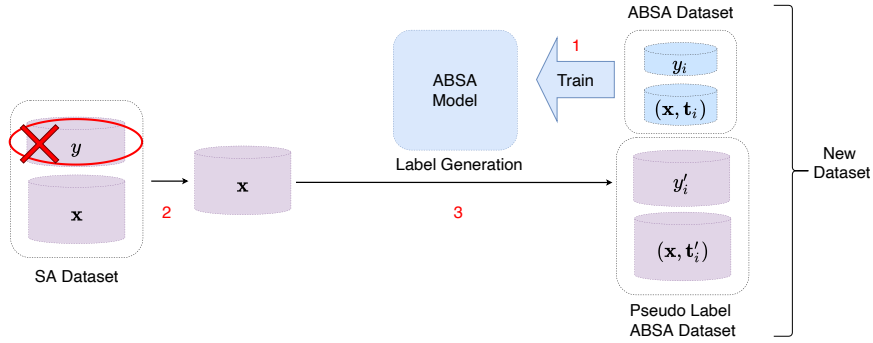
---

[1] `https://www.amazon.com/`
[2] `https://www.yelp.com/`
[3] `https://www.kaggle.com/bittlingmayer/amazonreviews`

Figure 2: Dataset Generation in the Pseudo-Label (PL) Method. This figure shows a pipeline of the traditional Pseudo-Label method. $\mathbf{x}$ is the input data, a sentence in the SA dataset, while $y$ is the sentiment carried by a sentence. $\mathbf{t}_i$ indicates the position of an aspect term in a sentence, and $y_i$ is the label for that aspect term. $\mathbf{t}'_i$ and $y'_i$ are pseudo labels generated by the ABSA model. As we can see, in the PL method, the sentence sentiment labels are dropped, and the SA dataset is regarded as an unlabeled dataset.

2019b). In this paper, we first take the Pseudo-Label method to utilize the SA datasets to solve the challenge faced by the ABSA task.

## 1.2 Pseudo-Label

The family of Pseudo-Label methods has had wide success in multiple fields (Pham et al., 2020; Ge et al., 2020; Mallis et al., 2020; Zoph et al., 2020; He et al., 2019a). The core of this family is to "trust" the generated fake labels by running the unlabeled samples through a teacher network that is trained by using the limited number of labeled samples. The generated labeled samples are then combined with the original set of supervised datasets and fed to the final model training.

In this article, our core mission is to incorporate the large-scale datasets into the sentiment analysis with the targeted ABSA task. While there have been works on this line, such as He et al. (2018) and He et al. (2019b), exploring the Pseudo-Label methods has been very much untapped. Indeed, a very straightforward technological solution is depicted in Figure 2. One can apply the traditional Pseudo-Label method to generate a bunch of pseudo-aspect-based sentiment labels from the SA or even the unlabeled datasets. However, a consequence of this is the total abandonment and waste of the provided coarse-grained labels. While seemingly acceptable, we argue that due to the homogeneous root for the ABSA and SA tasks, the under-exploiting of the sentence-level coarse-grained sentiment labels is sub-optimal. It will be unnecessary if the traditional framework throws away the coarser-grained labels containing finer-

grained task-relevant information. We argue that the Pseudo-Label family of approaches is limited to fit a *uniform granularity* situation. They ought to evolve and further adapt to the discrepancy of granularity in the label space.

## 1.3 Dual-granularity Pseudo Labels

To solve the aforementioned problem, we propose the Dual-granularity Pseudo Labeling framework (DPL). In essence, the DPL augments the original PL framework and is capable of leveraging the labels drawn from both granularities. Briefly, the DPL relies on two teacher models obtained from datasets from both granularities, respectively, then generates pseudo labels for both sides. As a result, datasets from both granularity levels can be merged into a whole, with every sentence sample being tagged by both finer- and coarser-set of labels. To facilitate the employment of both sets of labels, we set a few standard conditions as the design principle of DPL. Slightly more concretely, DPL establishes two separate pathways leading to prediction for both granularities. Together, the two pathways interact in the representation space and ideally may possess controlled information flow that respectively corresponds and only correspond to the considered granularity. We incorporate an adversarial module to accomplish this functionality.

On the widely used benchmarks, SemEval 2014 task 4 subtask 2 (Pontiki et al.), the DPL method significantly surpasses the current state-of-the-art method. We deem our simple but effective framework DPL pioneering a bi-granularity level dataset

merging. In what follows, we empirically validate that DPL is a framework that can be seamlessly combined with the previous pre-training or multi-task learning methods in terms of ABSA and SA dataset merging.

To sum up, we make the following contributions in this paper:

1. Among those works to solve the lack of labeled data in the ABSA task, we pioneer to adopt and enhance a pseudo-label framework to leverage the coarser-grained SA labels.

2. We propose a novel general framework called Dual-granularity Pseudo Labels (DPL). Just like the vanilla PL method, the DPL is established as a general framework. We validate that DPL is also compatible with previous work on this line, such as pre-training or multi-task learning (MTL). DPL has achieved excellent performances on the standardized ABSA benchmarks such as SemEval 2014, which significantly outperforms the prior works.

## 2 Related Works

### 2.1 Aspect-based Sentiment Analysis (ABSA)

ABSA is a finer-grained task of Sentiment Analysis (SA). It is a pipeline task, including aspect term extraction and aspect term sentiment classification. Aspect term sentiment classification is the true target task in this paper. For convenience, we use ABSA to refer to this task in the remaining parts.

Like other application tracks in NLP, the family of neural network models has wide successes in this task (Jiang et al., 2011; Vo and Zhang, 2015; Zhang et al., 2016; Ma et al., 2017; Li et al., 2018; Wang et al., 2018; Huang et al., 2018; Song et al., 2019). Wang et al. (2016) introduce attention mechanism into an LSTM to model the inter-dependence between sentence and aspect term. Tang et al. (2016) apply Memory Networks in this task.

Syntax-based models have also been explored widely in this domain (Dong et al., 2014; Tai et al., 2015; Nguyen and Shirai, 2015; Liu et al., 2020; Li et al., 2021; Pang et al., 2021). Sun et al. (2019) and Zhang et al. (2019) introduced graph convolution networks (GCN) to leverage the structured information from the dependency tree. Huang and Carley (2019) used graph attention networks (GAT) to improve the performance. Bai et al. (2020) and Wang et al. (2020) took the syntax relations as edge

features and introduced them into the Relational Graph Attention Network (RGAT).

In addition, pretrained language models like BERT (Devlin et al., 2018) have greatly promoted the development of ABSA (Li et al., 2018; Gao et al., 2019; Song et al., 2019; Rietzler et al., 2019; Yang et al., 2019).

### 2.2 Using Extra Dataset for ABSA

Due to the dataset scale challenge of the ABSA task, there have been some methods exploring how to utilize the auxiliary dataset.

Some of them (Xu et al., 2019; Rietzler et al., 2019; Yu et al., 2021b) achieve decent ABSA performance by post-processing or fine-tuning BERT (Devlin et al., 2018) with an additional unlabeled dataset. For these methods, we argue that the cost of computation is too high. Moreover, DPL does not conflict with it and can accommodate the results of these works. We take Rietzler et al. (2019)'s work as an example for comparison in experiments.

The others (He et al., 2018, 2019b; Chen and Qian, 2019; Liang et al., 2020; Yang et al., 2019; Oh et al., 2021; Yu et al., 2021a; Yan et al., 2021) utilize some labeled datasets and propose (later extend) a framework applying multitask learning methods. These auxiliary labeled datasets mainly include the sentiment analysis (SA) task and other subtasks of ABSA, such as Aspect Term Extraction, Opinion Term Extraction, and so on (Yan et al., 2021). DPL is more similar to these methods, using labeled datasets. However, we argue that the datasets of other subtasks can't solve the problem of the high annotation cost. Thus, DPL utilizes the SA task as auxiliary datasets and is the first to apply the PL method to this problem.

### 2.3 Pseudo-Label

Pseudo-label (PL), often associated with self-training, is a semi-supervised learning method. PL has been utilized and further developed by many studies (Ge et al., 2020; Mallis et al., 2020; Zoph et al., 2020; He et al., 2019a). It has been successfully applied in many tasks, such as image classification (Pham et al., 2020; Xie et al., 2020), object detection (Ge et al., 2020), text classification (Mukherjee and Awadallah, 2020), Etc.

However, these PL methods are inapplicable under a non-uniform granularity situation; that is, there are massive available coarse-grained datasets for fine-grained tasks. These existing methods can only discard the coarse-grained labels and treat

them as unlabeled datasets. Thus, we argue that these PL methods cause loss of information and are definitely unreasonable.

## 3 Preliminary

### 3.1 Pseudo-Labels

The traditional PL method generally involves a labeled set denoted by $D$ and an unlabeled set $D_u$. A teacher model is trained on $D$ by cross-entropy loss:

$$\mathcal{L}(\Theta_T) = \sum_{(x,y) \in D} [-\log P_{\Theta_T}(y|x)] \qquad (1)$$

where $\Theta_T$ denotes the parameters of the teacher model. The cross-entropy loss function is adopted for general classification problems, including image classification, detection, and semantic segmentation (Ge et al., 2020; Pham et al., 2020; Xie et al., 2020; Zoph et al., 2020).

In what follows, on the unlabeled dataset $D_u$, one can obtain the corresponding labels via running the unlabeled input through an inference procedure of the teacher model. The yielded label set for $D_u$ forms a pseudo-labeled dataset that can later be combined with the original dataset with gold annotations. A student model $M_S$ is trained by the newly merged dataset:

$$\mathcal{L}(\Theta_S) = \sum_{(x,y) \in D} [-\log P_{\Theta_S}(y|x)] + \\ \lambda \sum_{(x_u,y') \in D'_u} [-\log P_{\Theta_S}(y'|x_u)] \qquad (2)$$

where $y'$ indicates the pseudo label corresponding to the sample $x_u$ generated by the teacher model. $D'_u$ are the pseudo-label augmented version of $D_u$. $\lambda$ is a weighing term.

## 4 Dual-granularity Pseudo Labeling

In short, our work focuses on expanding the traditional PL method to utilize coarse-grained datasets. To achieve this goal, we draw inspiration from the multi-task learning community and augment the PL method with a different modeling pathway. Consequently, we obtain a framework where two separate pathways are trained synergistically targeted at labels of both granularities.

### 4.1 Setup

Our work is based on two datasets, the fine-grained and the coarse-grained datasets in the same domain.

Let us use $D_{\text{fine}}$ and $D_{\text{coarse}}$ to denote two datasets respectively. For the coarse-grained dataset $D_{\text{coarse}}$, the task is to learn a mapping:

$$f_{\text{coarse}}(\mathbf{x}) \to y, \qquad (3)$$

For the fine-grained dataset $D_{\text{finer}}$, the target mapping is:

$$f_{\text{fine}}(\mathbf{x}, \mathbf{t}_i) \to y_i, i \in \{1, ..., m\} \qquad (4)$$

where $(\mathbf{x}, y) \in D_{\text{coarse}}$ and $(\mathbf{x}, \mathbf{t}_i, y_i) \in D_{\text{fine}}$. $\mathbf{x}$ is the input data, and $y$ is the corresponding label for $\mathbf{x}$. $\mathbf{t}_i \subseteq \mathbf{x}$. $m$ means that $\mathbf{x}$ has m sub-parts totally, and $y_i$ is the corresponding label for $\mathbf{t}_i$.

The traditional PL method is limited to fit a uniform granularity situation. The first step to resolve this limitation is to merge the coarse-grained dataset with the fine-grained dataset. Like the traditional PL method, we train a teacher model on one dataset and generate pseudo labels for the other dataset. We repeat this process at two granularities. Here, we suppose that $\mathbf{x}_i$ for each $\mathbf{x}$ in the $D_{\text{coarse}}$ have been extracted. After pseudo labels generation, two new datasets are generated, donates as $D'_{\text{fine}}$ and $D'_{\text{coarse}}$, and a new dataset $D'$ are merged by these two datasets. Specifically,

$$D' = D'_{\text{fine}} \cup D'_{\text{coarse}}, \qquad (5)$$

where $(\mathbf{x}, \mathbf{t}_i, y, y'_i) \in D'_{\text{coarse}}$ and $(\mathbf{x}, \mathbf{t}_i, y', y_i) \in D'_{\text{fine}}$. $y'$ and $y'_i$ are the generated pseudo labels.

Up to now, we get a new dataset with a much larger scale. Our goal translates into obtaining a model trained by the new dataset $D'$ with high performance on the fine-grained task. In other words, compared with the traditional PL method, the key problem is: how to utilize the coarse-grained labels to improve the model's performance on the fine-grained task.

### 4.2 DPL Skeleton

As we mentioned, the core challenge for adapting the vanilla PL method is to utilize coarse-grained labels. As displayed in Figure 3, we set dual pathways corresponding to each granularity. Both pathways are finished by setting a proper softmax-based classifier. Using $\mathbf{z}$ and $\mathbf{h}$ to denote the internal representation vectors for both pathways, we decompose the design philosophy of DPL by the following three conditions:
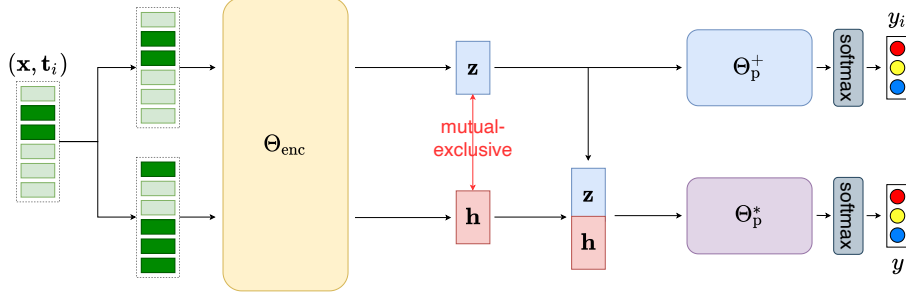
Figure 3: The Model for the DPL Framework. $(\mathbf{x}, \mathbf{t}_i)$ is the input data. $\mathbf{t}_i$ indicates the aspect terms, which are painted by the dark green. We first generate $(\mathbf{x}, \mathbf{t}_i)$ and $(\mathbf{x}, \mathbf{1} - \mathbf{t}_i)$ as the input of the upper and lower pathways, respectively. In this case, $\mathbf{t}_i = (0, 1, 1, 0, 0, 0)$ and $\mathbf{1} - \mathbf{t}_i = (1, 0, 0, 1, 1, 1)$. "$\Theta_{\text{enc}}$" is an encoder that outputs $\mathbf{z}$ and $\mathbf{h}$. "$\Theta_{\text{p}}^+$" is a predictor for the fine-grained task, and "$\Theta_{\text{p}}^*$" is a predictor for the coarse-grained task. Correspondingly, $y_i$ is the prediction for the fine-grained task, and $y$ is the prediction for the coarse-grained task. "mutual-exclusive" means the information carried by $\mathbf{z}$ and $\mathbf{h}$ has little overlap.

- $\mathbf{z}$ carries adequate information to determine the label at the fine-grained level. More formally, there exists a function $f_{\Theta_{\text{p}}^+}$ in the overall functional space that is able to map the $\mathbf{z}$ to $y_i$.

- The union set of $\mathbf{h}$ and $\mathbf{z}$ is capable of determining the label at the coarse level. There exists another function $f_{\Theta_{\text{p}}^*}$ in the overall functional space that is sufficient to map the $[\mathbf{z} \circ \mathbf{h}]$ to $y$.

- $\mathbf{h}$ and $\mathbf{z}$ are mutually exclusive in terms of the carried information. That means we cannot train a function $f_{\Theta_{\text{p}}^*}$ to map $\mathbf{h}$ to $y_i$, due to the lack of information contributed from $\mathbf{z}$.

The main rationale behind these three conditions may include but is not limited to: (i)-the information passing through the pathway with $\mathbf{z}$ is only required in the fine-grained task; (ii)-the other information needed by the coarse-grained task passes through the pathway with $\mathbf{h}$; (iii)-the prediction at coarse-grained level is based on the concatenation of $\mathbf{h}$ and $\mathbf{z}$, while either of them is insufficient to accomplish the prediction of coarse-grained labels.

In order to satisfy the model to these three conditions, our loss function consists of three terms. Among them, the two terms are the classification loss terms for the fine- and coarse-grained tasks, respectively, fulfilling conditions 1&2. For condition 3, we draw inspiration from adversarial training (Lample et al., 2017) to reduce the fine-grained task-relevant information carried by $\mathbf{h}$.

### 4.2.1 Fine- and Coarse-grained Tasks

As shown in Figure 3, the model consists of an encoder, $\Theta_{\text{enc}}$, together with two predictors, $\Theta_{\text{p}}^*$ and $\Theta_{\text{p}}^+$. In particular, $\Theta_{\text{enc}}$ encodes each input data $(\mathbf{x}, \mathbf{t}_i)$ into two intermediate results, $\mathbf{z}$ and $\mathbf{h}$. In the figure, the top line with $\mathbf{z}$ is the pathway for the fine-grained task-relevant information flow, while the bottom line with $\mathbf{h}$ is the pathway for the fine-grained task-irrelevant information flow.

The fine-grained predictor $\Theta_{\text{p}}^+$ spits out prediction based on $\mathbf{z}$, with a cross-entropy loss:

$$
\begin{aligned}
&\mathcal{L}_{\text{fine}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^+) \\
&= \sum_{(\mathbf{x}, \mathbf{t}_i, y, y_i) \in D'} [-\log P_{\Theta_{\text{p}}^+}(y_i|\mathbf{z})],
\end{aligned}
\tag{6}
$$

Another crucial design in the DPL is that the concatenation of $\mathbf{h}$ and $\mathbf{z}$, $[\mathbf{h} \circ \mathbf{z}]$, is fed to decide the prediction of the sequence-level prediction:

$$
\begin{aligned}
&\mathcal{L}_{\text{coarse}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^*) \\
&= \sum_{(\mathbf{x}, \mathbf{t}_i, y, y_i) \in D'} [-\log P_{\Theta_{\text{p}}^*}(y|\mathbf{h} \circ \mathbf{z})].
\end{aligned}
\tag{7}
$$

The gradient of this loss will update the model parameters on both pathways. To prevent the degenerated case where the two pathways act completely separately, we introduce another crucial part to DPL in the next subsection.

### 4.2.2 Adversarial Training

The current version of DPL could still work as two separate systems, which is deemed a degenerated case. Therefore, to guarantee the mutual exclusiveness between the $\mathbf{h}$ and the $\mathbf{z}$, we introduce an

24

adversarial training loss term to maximally reduce the fine-grained task-relevant information carried by $\mathbf{h}$:

$$\mathcal{L}_{\text{enc}}(\Theta_{\text{enc}}) = \sum_{(\mathbf{x},\mathbf{t}_i,y,y_i)\in D'} [-\log P_{\Theta_{\text{p}}^+}(1-y_i|\mathbf{h})], \tag{8}$$

$$\mathcal{L}_{\text{dis}}(\Theta_{\text{p}}^+) = \sum_{(\mathbf{x},\mathbf{t}_i,y,y_i)\in D'} [-\log P_{\Theta_{\text{p}}^+}(y_i|\mathbf{h})], \tag{9}$$

$$\mathcal{L}_{\text{adv}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^+) = \mathcal{L}_{\text{dis}}(\Theta_{\text{p}}^+) + \lambda\mathcal{L}_{\text{enc}}(\Theta_{\text{enc}}), \tag{10}$$

where $\lambda$ weighs the trade-off between $\Theta_{\text{enc}}$ and $\Theta_{\text{p}}^+$. The adversarial training was first introduced in Lample et al. (2017) and has been widely used (Zhao et al., 2018; Fu et al., 2018; Shen et al., 2017; Melnyk et al., 2017). The loss term trains $\Theta_{\text{enc}}$ to fool $\Theta_{\text{p}}^+$ by removing fine-grained task relevant information from $\mathbf{h}$. Considering that $\mathbf{z}$ is only required by the fine-grained task, the less fine-grained task-relevant information the $\mathbf{h}$ has, the less overlap there is between the $\mathbf{h}$ and $\mathbf{z}$. As a result, the adversarial training makes h and z more mutually exclusive in terms of the carried information.

### 4.2.3 Loss Function

The overall loss function to optimzie DPL combines as below:

$$\begin{aligned}\mathcal{L}(\Theta_{\text{enc}}, \Theta_{\text{p}}^*, \Theta_{\text{p}}^+) =&\mathcal{L}_{\text{fine}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^+)\\ &+\alpha\mathcal{L}_{\text{coarse}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^*)\\ &+\beta\mathcal{L}_{\text{adv}}(\Theta_{\text{enc}}, \Theta_{\text{p}}^+)\end{aligned} \tag{11}$$

where $\alpha$ and $\beta$ are weighing terms. With this design of the loss functions, we posit that all three philosophies should be satisfied. The ideal result for it is that (i)-$\mathbf{z}$ only carries information dedicated at the fine-level; (ii)-$\mathbf{h}$ carries the information of the entire coarse level (i.e., the whole sequence) excluding the information of $\mathbf{z}$; (iii)-neither $\mathbf{h}$ nor $\mathbf{z}$ is sufficient on deciding the whole-sequence coarse-level prediction, but with the concatenation of them, $\mathbf{h} \circ \mathbf{z}$, the information is just adequate.

### 4.3 Grounding DPL to ABSA

#### 4.3.1 Document-level Sentiment Analysis.

The task aims to analyze the sentiments reflected by sentences. Given an ordinary labeled document-level dataset

| Dataset | positives | | neutral | | negative | |
|---------|-------|------|-------|------|-------|------|
| | Train | Test | Train | Test | Train | Test |
| Rest | 2164 | 727 | 637 | 196 | 807 | 196 |
| Laptop | 976 | 337 | 455 | 128 | 851 | 167 |

Table 1: Statistics of SemEval 2014 task 4 subtask 2.

$D = \{(\mathbf{x}^0, y^0), (\mathbf{x}^1, y^1)\ldots(\mathbf{x}^N, y^N)\}$, where $\mathbf{x}^i$ donates a sentence and $y^i$ donates the sentiment polarity of the sentence. The goal of the task is to learn a mapping function: $f_{\text{sent}}(\mathbf{x}^i) \to y^i$.

#### 4.3.2 Aspect-based Sentiment Analysis.

The ABSA task is to derive the sentiment polarity attached to specific aspect terms in the given sentence. Formally, one can draw a data point $(\mathbf{x}^i, \mathbf{y}^i)$ from the dataset $D$. We assign a separate variable indicating the aspect terms annotation, $\{\mathbf{t}^{i,1}, \ldots, \mathbf{t}^{i,N_i}\}$, where $N_i$ denotes the number of total aspect terms in $\boldsymbol{\tau}^i$. In addition, the label $\mathbf{y}$ is a combination of polarities corresponding to aspect terms, $\mathbf{y}^i = \{y^{i,1}, \ldots, y^{i,N_i}\}$. The goal for the ABSA is to learn the mapping $f_{\text{aspect}}(\mathbf{x}^i, \mathbf{t}^{i,k}) \to y^{i,k}$, where $k \in \{1, \ldots, N_i\}$.

#### 4.3.3 Implementation

Before implementing a specific DPL model, we first map the task objectives of the SA and ABSA tasks to the coarse- and fine-grained tasks in the DPL framework. The coarse-grained task is the SA task, while the fine-grained task is the ABSA task. In another word, the mapping $f_{\text{sent}}(\mathbf{x}^i) \to y^i$, is considered as the coarse-grained mapping $f_{\text{coarse}}(\mathbf{x}) \to y$, and the mapping $f_{\text{aspect}}(\mathbf{x}^i, \mathbf{t}^{i,k}) \to y^{i,k}$ is considered as $f_{\text{fine}}(\mathbf{x}, \mathbf{t}_i) \to y_i$.

Then we choose the model for $\Theta_{\text{enc}}$, $\Theta_{\text{p}}^+$ and $\Theta_{\text{p}}^*$. $\Theta_{\text{p}}^+$ and $\Theta_{\text{p}}^*$ are simple multilayer perceptron (MLP). It is worth noting that $\Theta_{\text{enc}}$ can be a prior ABSA model. Thus, we argue that the DPL framework can be applied to most ABSA methods. Typically, we choose Bai et al. (2020)'s and Rietzler et al. (2019)'s works and a multi-task learning baseline as examples to verify. The results are shown in Table 3.

## 5 Experiments

### 5.1 Experimental Setup

#### 5.1.1 Dataset

The experiments of the DPL framework require at least two datasets at different granularities. For the

ABSA task, we select the SemEval dataset (Pontiki et al.) as the fine-grained sentiment task dataset and the Amazon reviews dataset from Kaggle[4] as the coarse-grained sentiment task dataset. The SemEval datasets are used as our core task dataset, and the Amazon reviews dataset is used as an auxiliary dataset.

**Dataset SemEval.** This dataset is SemEval 2014 task 4 subtask2 (Pontiki et al.). It has two sub-datasets, the reviews in the restaurant and laptop domains. We show more details in Table 1.

**Dataset Amazon Reviews.** The dataset contains 3.6 million sentences in the training set and 0.4 million sentences in the test set. Considering the huge data volume gap, we only chose the test set as the auxiliary dataset for this experiment.

### 5.1.2 Generation of Pseudo Labels

Here we provide some details of the pseudo labels generation process.

As a result of the PL generation, the ABSA dataset has true aspect-level sentimental labels and pseudo-sentence-level sentimental labels, while the SA dataset has true sentence-level sentimental labels and pseudo-aspect-level sentimental labels.

To get aspect terms from the sentence in the SA dataset, we first performed aspect extraction using the model proposed by Li et al. (2019) and discarded sentences without aspect terms.

We train the model proposed by (Bai et al., 2020) as the teacher models on the aspect-level dataset with the accuracy scores of 86.05% and 79.53% respectively on the domain of Restaurant and Laptop.

We train a BERT+Linear as the teacher model on the document-level dataset, with a 94.45% accuracy score in the restaurant Domain and a 93.35% accuracy score in the laptop domain.

### 5.1.3 Implementation Details

In addition to the above introduction, some more important details of our experiments need to be clarified for ease of understanding.

**Evaluate Matrix**

The model for **ABSA** is tested on SemEval's test set. Like those who have performed this work before, we use the model classification accuracy (ACC) and macro-F1 (F1) scores as the evaluation criterion.

**Batch Loader**

| Model | | Restaurant | | Laptop | |
|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 |
| Auxiliary | He et al. (2018) | 78.73 | 68.63 | 71.91 | 68.79 |
| | Chen and Qian (2019) | 79.55 | 71.41 | 73.87 | 70.10 |
| | He et al. (2019b) | 83.89 | 75.66 | 75.36 | 72.02 |
| | Liang et al. (2020) | 84.93 | 76.66 | 77.51 | 73.42 |
| BERT | Bai et al. (2020)* | 86.04 | 80.27 | 79.53 | 74.54 |
| | Pang et al. (2021) | 87.66 | 82.97 | 80.22 | 77.28 |
| | Li et al. (2021) | 87.13 | 81.16 | 81.80 | 78.10 |
| | Rietzler et al. (2019) | 87.89 | 81.05 | 80.23 | 75.77 |
| Ours | DPL | 89.54 | 84.86 | 81.96 | 78.58 |

Table 2: Results of different methods. "BERT" represents the works that are also based on the BERT (Devlin et al., 2018), "Auxiliary" represents the methods that also utilize auxiliary datasets to help the ABSA task. "*" means our replication results. The results show that our method achieves state-of-the-art in this benchmark.

Since the size of the current auxiliary dataset is much larger than the existing dataset. To avoid the large auxiliary dataset changing the original dataset distribution, we adopt two asynchronous loaders and define the step ratio $k$, i.e., whenever the model is trained on the original dataset by 1 step, it is trained on the auxiliary dataset by $k$ steps. In general, we set $k = 1$.

**Model Implementation**

The encoder has three main structures for the **ABSA** task: BERT (Devlin et al., 2018), Relational Graph Attention Networks (RGAT) (Wang et al., 2020), and masking embedding module. The BERT and RGAT have been proved to have a good effect on this task. The mask embedding module is used to generate **z** and **h**. It is similar to the implementation of "segment_id" in the code of BERT.

### 5.2 Main Results

Table 2 shows that the DPL has achieved a state-of-the-art (SOTA) performance in terms of the average accuracy and F1-scores on the SemEval 2014 task 4 subtask 2 dataset. The group denoted as "Auxiliary Dataset is multi-task learning methods based on labeled datasets. Compared with them, our work shows the advantage of the PL method. "BERT-based" are some recently published works with good results. Obviously, our method achieves significant improvements over them.

It should be noted that our design is based on the BERT. Thus the comparison is not made with the methods based on a more powerful pre-trained model, such as Roberta (Liu et al., 2019), DeBERTa (Silva and Marcacini), and GPT-3 (Floridi and Chiriatti, 2020).

| Model | Restaurant | | Laptop | |
|---|---|---|---|---|
| | Acc | F1 | Acc | F1 |
| **RGAT (Bai et al., 2020)** | 86.04 | 80.27 | 79.53 | 74.54 |
| **RGAT+DPL** | 87.22 | 81.47 | 81.01 | 77.52 |
| **Improvement** | **+1.18** | **+1.20** | **+1.48** | **+2.98** |
| **Adapter(Rietzler et al., 2019)** | 87.89 | 81.05 | 80.23 | 75.77 |
| **Adapter+DPL** | 89.54 | 84.86 | 81.96 | 78.58 |
| **Improvement** | **+1.65** | **+3.71** | **+1.73** | **+2.81** |
| **MultiBERT** | 84.54 | 78.52 | 78.32 | 73.87 |
| **MultiBERT+DPL** | 85.52 | 79.61 | 79.75 | 75.80 |
| **Improvement** | **+0.98** | **+1.09** | **+1.43** | **+1.93** |

Table 3: Results of Combining DPL with Other Methods. Restaurant and Laptop are two benchmarks same as those in Table 2. RGAT (Bai et al., 2020), Adapter (Rietzler et al., 2019) are typical ABSA methods. Multi-BERT is a multi-task baseline implemented by us. It predicts the SA label based on the "[cls]" and predicts the ABSA task based on the specific word vector. We add the DPL framework to them, denoted as "+DPL", and achieve significant improvements.

## 5.3 DPL as a General Framework

As we mentioned, we promote DPL as a general framework capable of combining other methods on the ABSA task. Table 3 shows the performances of some typical methods before and after they combine the DPL framework. On the one hand, RGAT (Bai et al., 2020) is a model architecture based on GAT and BERT. Thus the improvement shows that the DPL framework fits other architectural designs, even without auxiliary datasets. On the other hand, for those methods involving auxiliary datasets, we take Adapter (Rietzler et al., 2019) and MultiB-ERT for demonstration. Previous works are mainly divided into two categories, pretraining and multi-task learning. Adapter (Rietzler et al., 2019) can be categorized into the pretraining class while Multi-BERT is a multi-task learning baseline inspired by He et al. (2018). Since the previous works using the multi-task method to combine the SA and the ABSA datasets were LSTM based, we implemented a better model based on the BERT. All the improvements verify that the DPL framework does not conflict with these methods and exhibits full compatibility for further performance gains.

## 5.4 Ablation Study

We set up several sets of ablation experiments and present the results in Table 4 to explore the role of adversarial training and pseudo labels in the DPL framework.

The above experiments contain two types of BERT on the SemEval Restaurant dataset. To ensure the fairness of the ablation experiments, we

| Model | Restaurant+Pre | | Restaurant | |
|---|---|---|---|---|
| | Acc | F1 | Acc | F1 |
| **DPL** | 89.54 | 84.86 | 86.68 | 80.44 |
| **Traditional Pseudo-Label** | -1.43 | -2.09 | -1.60 | -2.73 |
| **- adversarial training** | -1.96 | -3.31 | -1.96 | -3.60 |
| **- coarse-grained pseudo labels** | -1.60 | -2.74 | -1.34 | -1.35 |
| **- fine-grained pseudo labels** | -1.96 | -2.84 | -0.79 | -1.79 |

Table 4: Results of ablation study. "Restaurant" takes plain BERT as the initial model while "Restaurant+Pre" takes Rietzler et al. (2019)'s BERT as the initial model. "DPL" denotes our method. "Traditional Pseudo-Label" represents we take the PL method for fine-grained tasks dropped out the coarse-grained labels. The last three cases named in the form of "- X" means that we deleted the "X" from the original DPL to evaluate the effect of "X".

use the same parameters when training the same group, and the parameter configurations are shown in Appendix.

The comparison with "Traditional Pseudo-Label" shows the advantages of our method. From the item "- adversarial training", the significant decline on F1 reflects that adversarial training plays an important role in the DPL framework. The items, "- coarse-grained pseudo labels" and "- fine-grained pseudo labels", show that only adding adversarial training at one granularity has less effect than adding it both ways.

Furthermore, we also take Chamfer Distance (CD) between the set of $\mathbf{h}$ and the set of $\mathbf{z}$ to provide an insight into the effect of the mutual exclusiveness. And the CD of the model with the adversarial training process is 30% larger than that of the model without this process. That means the adversarial training process increases the distance between the variable $\mathbf{h}$ and $\mathbf{z}$.

## 6 Conclusion

In this paper, we propose Dual-granularity Pseudo Labeling (DPL). DPL extends from the vanilla Pseudo-Label method and augments it to a dual-pathway system. It additionally enforces strong control of information flow directing to the data at different granularities of annotation. The results demonstrate the state-of-the-art performance of DPL on the data-scarce ABSA task. As a pioneering framework design, we also show that the DPL is compatible with pre-training and multi-task learning methods as published before. In the future, we hope to explore the possibility of DPL in other domains, such as computer vision problems where

the discrepancy of granularities possesses.

## References

Xuefeng Bai, Pengbo Liu, and Yue Zhang. 2020. Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299.

Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019a. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. *arXiv preprint arXiv:1806.04346*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019b. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1906.06906*.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606*.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: Manipulating images by sliding attributes. *arXiv preprint arXiv:1706.00409*.

Lishuang Li, Yang Liu, and AnQiao Zhou. 2018. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd conference on computational natural language learning*, pages 181–189.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. 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 6319–6329.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. An iterative knowledge transfer network with routing for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.01935*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Shu Liu, Wei Li, Yunfang Wu, Qi Su, and Xu Sun. 2020. Jointly modeling aspect and sentiment with dynamic heterogeneous graph neural networks. *arXiv preprint arXiv:2004.06427*.

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*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. 2020. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33.

Igor Melnyk, Cicero Nogueira dos Santos, Kahini Wadhawan, Inkit Padhi, and Abhishek Kumar. 2017. Improved neural text attribute transfer with non-parallel data. *arXiv preprint arXiv:1711.09395*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

Shinhyeok Oh, Dongyub Lee, Taesun Whang, IlNam Park, Gaeun Seo, EungGyun Kim, and Harksoo Kim. 2021. Deep context-and relation-aware learning for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.03806*.

Shiguan Pang, Yun Xue, Zehao Yan, Weihao Huang, and Jinhui Feng. 2021. Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2627–2636.

Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, and Quoc V Le. 2020. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*.

Maria Pontiki, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *arXiv preprint arXiv:1705.09655*.

Emanuel H Silva and Ricardo M Marcacini. Aspect-based sentiment analysis using bert with disentangled attention.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. 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 5683–5692.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence*.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018. Target-sensitive memory networks for aspect sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 957–967.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. 2021. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*.

Heng Yang, Biqing Zeng, JianHao Yang, Youwei Song, and Ruyang Xu. 2019. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing*, 419:344–356.

Guoxin Yu, Xiang Ao, Ling Luo, Min Yang, Xiaofei Sun, Jiwei Li, and Qing He. 2021a. Making flexible use of subtasks: A multiplex interaction network for unified aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2695–2705.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021b. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *International conference on machine learning*, pages 5902–5911. PMLR.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*.