Revisiting Self-Training for Few-Shot Learning of Language Model

Yiming Chen[†,‡] **Yan Zhang**[†] **Chen Zhang**[†] **Grandee Lee**[†]
Ran Cheng[‡,*] **Haizhou Li**[†,⋆,⋆⋆]
[†]National University of Singapore [‡]Southern University of Science and Technology
[⋆]The Chinese University of Hong Kong (Shenzhen) [⋆⋆]Kriston AI Lab, China
{yiming.chen,chen_zhang,grandee.lee}@u.nus.edu,
{haizhou.li,eleyanz}@nus.edu.sg,
ranchengcn@gmail.com

Abstract

As unlabeled data carry rich task-relevant information, they are proven useful for few-shot learning of language model. The question is how to effectively make use of such data. In this work, we revisit the self-training technique for language model fine-tuning and present a state-of-the-art prompt-based few-shot learner, SFLM. Given two views of a text sample via weak and strong augmentation techniques, SFLM generates a pseudo label on the weakly augmented version. Then, the model predicts the same pseudo label when fine-tuned with the strongly augmented version. This simple approach is shown to outperform other state-of-the-art supervised and semi-supervised counterparts on six sentence classification and six sentence-pair classification benchmarking tasks. In addition, SFLM only relies on a few in-domain unlabeled data. We conduct a comprehensive analysis to demonstrate the robustness of our proposed approach under various settings, including augmentation techniques, model scale, and few-shot knowledge transfer across tasks. [1]

1 Introduction

Pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Yang et al., 2019; Lan et al., 2020; Raffel et al., 2020; Clark et al., 2020) have set new state-of-the-art performance in many downstream NLP tasks. However, such performance often relies on large-scale high-quality supervision. Unfortunately, labeled data are not always available in practice.

Recently, Brown et al. (2020) study how to facilitate the few-shot learning of language models via the GPT-3 model. It achieves remarkable performance on many NLP datasets without any gradient updates, by incorporating task-specific prompts into the text and reformulating the task as language modeling problems. However, GPT-3 has 175B parameters, that has a footprint too large for many real-world applications. Gao et al. (2020) applies the concept of prompt strategies in GPT-3 to small-footprint language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). After fine-tuned on a few annotated samples, the small-footprint models exhibit comparable performance to that of the large-footprint GPT-3 model. However, the performance of these models is still lagging behind those under supervised learning, which have a much smaller footprint. Intuitively, unlabeled data also carry rich information of downstream tasks and are more available than labelled data. In this paper, we focus on the few-shot learning of language model with a small amount of labeled and unlabeled data.

Semi-supervised learning benefits from partially labeled datasets. A common implementation of semi-supervised learning is self-training, which leverages supervision signals offered by labeled data to create pseudo-labels for unlabeled data. These pseudo labels serve as additional supervision to refine the models (Yarowsky, 1995; Blum and Mitchell, 1998; Zhu, 2005; Qiu et al., 2019; Zoph et al., 2020). Recent works (Schick and Schütze, 2020, 2021) apply self-training to language model few-shot learning in an iterative manner whereby multiple generations of models are trained on data pseudo-labeled by the ensemble of previous generations. However, the amount of in-domain unlabeled data required by these methods is quite large, that limits the scope of the applications, especially for low-resource downstream tasks. Du et al. (2020) try to retrieve more task-relevant unlabeled data from open-domain corpus, but the method depends on a quality sentence encoder.

To better address the above issue, we revisit the **S**elf-training techniques and introduce a data-efficient **F**ew-shot learner of **L**anguage

* Corresponding author.

[1]Our code is publicly available at https://github.com/MatthewCYM/SFLM

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9125–9135
November 7–11, 2021.

Model (SFLM). Inspired by recent advances in semi-supervised representation learning for images (Sohn et al., 2020), SFLM combines pseudo-labeling with consistency regularization.

Next, we briefly describe the workflow. Given each unlabeled sentence, we construct two views through weak augmentation (random dropout) and strong augmentation (token masking), respectively. The weakly-augmented view is first passed to a prompt-based language model (Gao et al., 2020) to derive the pseudo-label, while the strongly-augmented view is passed through the model to predict the probability distribution over classes, which is compared to the pseudo-label to derive a cross-entropy loss. This learning procedure encourages the model to capture the information that is almost outside of the data distribution, leading to effective utilization of data.

We evaluate SFLM on two groups of tasks – sentence classification and sentence-pair classification. Experiments show that our model outperforms other supervised and semi-supervised baselines. We also conduct a detailed analysis of the data efficiency of our model by examining its performance w.r.t various ratios of the amount of the unlabeled data to that of the labelled data. We find out that the performance gain diminishes as more unlabeled data are used. We further extend our method for a more challenging scenario: few-shot transfer across tasks, where the model is first trained on the labeled data of a source task and the unlabeled data of a target task, then evaluated on the target task. We provide the analysis of the factors that affect the model performance to motivate future research.

2 Related Work

2.1 Few-Shot Learning of Language Model

It is desirable to reduce the amount of labeled data for language model fine-tuning, a.k.a., language model few-shot learning. The popular methods usually address this problem with meta-learning (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017), which first pre-trains a model on a set of auxiliary tasks, then fine-tunes on the task of interest (Yu et al., 2018; Han et al., 2018; Bao et al., 2020; Bansal et al., 2020).

Recently, Brown et al. (2020) proposes GPT-3 and demonstrates that the language model itself has a great potential for few-shot learning through task demonstrations and prompts. As GPT-3 (Brown et al., 2020) has an extremely large footprint, that limits its scope of applications.

More recent studies explore few-shot learning with pre-trained language models (Gunel et al., 2020; Schick and Schütze, 2020; Gao et al., 2020) of smaller size. A representative example is the LM-BFF by (Gao et al., 2020), which explores automatic prompt generation and prompt-based fine-tuning with a RoBERTa-large (Liu et al., 2019) language model in a few-shot setup. LM-BFF has achieved comparable results w.r.t methods fine-tuned with the full annotated dataset.

We are motivated to study few-shot learning of language model with prompt-based language model fine-tuning. We exploit the rich information in unlabeled data with semi-supervised learning. Furthermore, we adopt an even smaller RoBERTa-base model as the backbone of our framework.

2.2 Self-Training

Self-training refers to the process of creating pseudo-labels on unlabeled data with a pre-trained teacher model, then applying these labeled data to train a student model. It is a simple and effective semi-supervised approach, which has benefited a wide range of tasks, such as image classification (Xie et al., 2020b), neural sequence generation (He et al., 2019), and parsing (McClosky et al., 2006). Generally, sophisticated learning algorithms (Sohn et al., 2020), and a large corpus of task-relevant data (Xie et al., 2020a) are required for self-training to work well.

From the algorithm perspective, FixMatch (Sohn et al., 2020) is a simple and effective self-training framework for image classification, which unifies consistency regularization and pseudo-labeling. In our work, we transfer this useful framework to language model few-shot learning by exploring various text augmentation techniques for fine-tuning the pre-trained language model.

From the data perspective, several recent works have shown the effectiveness of self-training for language model fine-tuning (Du et al., 2020; Schick and Schütze, 2020, 2021) leveraging a large amount of unlabeled data. PET (Schick and Schütze, 2020) adopts prompt-based fine-tuning and self-training for language model few-shot learning. This approach assumes the presence of a large number of unlabeled in-domain data (roughly 10,000 examples per class). In addition, Du et al. (2020) propose to retrieve task-relevant unlabeled data from a large-

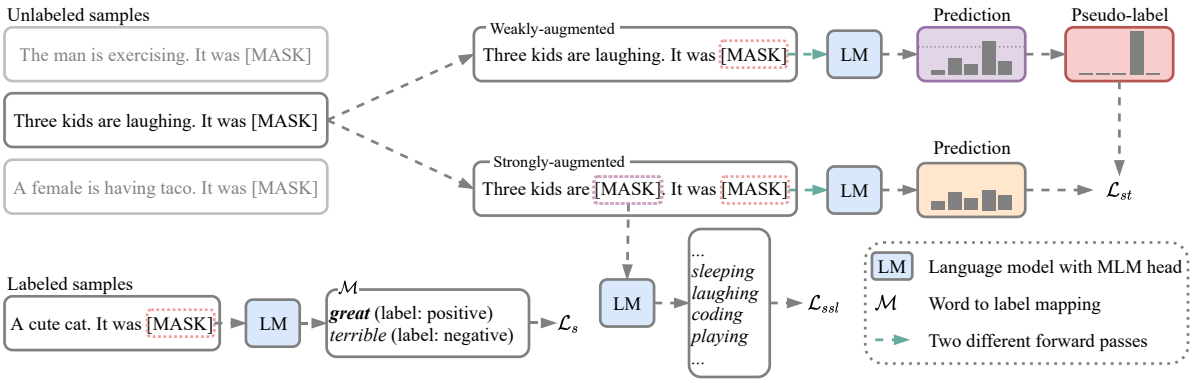

Figure 1: The learning process of SFLM on both labeled and unlabeled samples with three loss terms. For the supervised loss term $\mathcal{L}_s$, in SFLM, a pre-trained language model with a MLM head is used to get the predicted word from the template. Then the predicted word is mapped to the corresponding label with manually defined task-specific word to label mapping $\mathcal{M}$. Two loss terms are computed upon the unlabeled data: (1) We use masked language modeling to compute the self-supervised loss. (2) We use a weak augmented (dropout) sentence to get the pseudo-label, then force the prediction given by a strongly-augmented (random mask) view against the pseudo label via the self-training loss.

scale open-domain sentence bank. A paraphrase-based universal sentence encoder is designed to output sentence-level vectors for computing cosine similarity between labeled sentences and unlabeled ones in the sentence bank.

Unlike the prior studies, which rely on a large amount of unlabeled data and expensive computation resources, we do not assume the availability of abundant in-domain unlabeled data. Instead, we tackle the in-domain data constraint via improving data efficiency, i.e., proposing a scalable and effective self-training framework leveraging only a few unlabeled data.

3 Methodology

Problem setup: Our goal is to adapt pre-trained language models to downstream tasks in a few-shot setting. The model, m should correctly classify unseen examples leveraging very few labeled data points from each class. Let $\mathcal{X}$ denote a small set of labeled training data with N samples per class and an unlabeled dataset, $\mathcal{U}$ from the same task domain as $\mathcal{X}$. Assume that this unlabeled dataset has very limited size μN per class, where μ is the ratio between the size of $\mathcal{X}$ and that of $\mathcal{U}$.

During training, let each batch consist of B labeled data points, $\mathcal{X}_\mathcal{B}$, and μB unlabeled data points, $\mathcal{U}_\mathcal{B}$:

$$\mathcal{X}_\mathcal{B} = \{(x_i, y_i) : i \in (1, ..., B)\} \quad (1)$$

$$\mathcal{U}_\mathcal{B} = \{u_i : i \in (1, ..., \mu B)\} \quad (2)$$

Figure 1 illustrates the learning process with an example, containing one labeled and three unlabeled data samples. SFLM is optimized with following loss function:

$$\mathcal{L} = \mathcal{L}_s + \lambda_1 \mathcal{L}_{st} + \lambda_2 \mathcal{L}_{ssl} \quad (3)$$

where $\mathcal{L}_s$ is the prompt-based supervised loss applied to the labeled data (Gao et al., 2020), $\mathcal{L}_{st}$ and $\mathcal{L}_{ssl}$ refer to self-training loss and self-supervised loss applied to the unlabeled data accordingly, while λ_1 and λ_2 are fixed scalar hyper-parameters controlling the relative weight of the unlabeled loss terms.

Prompt-based supervised loss: The prompt-based supervised loss is motivated by LM-BFF (Gao et al., 2020). The classification is reformulated as a language modeling task, in which the probability of class prediction $y_i \in \mathcal{Y}$ is,

$$p_m(y_i|x_i) = p_m([\text{MASK}] = \mathcal{M}'(y_i|x_i^{prompt}) \quad (4)$$

where $\mathcal{M}'$ refers a mapping from task labels to the corresponding words[2], and x_i^{prompt} is the reconstructed input sentence with task-specific template. For instance, in a sentence-level binary classification task, the input sentence x_i is reconstructed as:

$$x_i^{prompt} = x_i \circ \text{It was [MASK]}. \quad (5)$$

where $\circ$ denotes the string concatenation operation.

Instead of using an additional classifier, the pre-trained masked language modeling head decides which word to be filled in the masked position.

[2] $\mathcal{M}'$ is the inverse operation of $\mathcal{M}$ as shown in figure 1

Then we could fine-tune the model with the standard cross-entropy loss:

$$\mathcal{L}_s = \frac{1}{B}\sum_{i=1}^{B} H(y_i, p_m(y_i|x_i)) \tag{6}$$

Self-training loss: For each unlabeled sentence u_i, we obtain the weakly-augmented version $\alpha(u_i)$ and the strongly-augmented version $\mathcal{A}(u_i)$, where α and $\mathcal{A}$ refers to different augmentation strategies. The self-training process consists of two stages. Firstly, we assign a pseudo label to each unlabeled sentence in the batch by computing the output probability distribution corresponding to the weakly-augmented input sentence $\alpha(u_i)$, defined as $q_i = p_m(y_i|\alpha(u_i))$. The pseudo label, $\hat{q}_i$, is obtained by $\hat{q}_i = \arg\max(q_i)$. Secondly, we compute the prompt-based cross-entropy loss between $\hat{q}_i$ and the prediction corresponding to the strongly-augmented input sentence $\mathcal{A}(u_i)$. The self-training loss is defined as,

$$\mathcal{L}_{st} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbb{1}(max(q_i) \geq \tau) \cdot \mathrm{H}(\hat{q}_i, p_m(y_i|\mathcal{A}(u_i))) \tag{7}$$

where τ defines the threshold above which we retain a pseudo-label.

Sohn et al. (2020) adopt AutoAugment (Cubuk et al., 2018) for image augmentation, and highlight the importance of applying proper augmentation techniques in self-training. Text augmentation techniques can be tricky due to the discrete nature of text data. The recent successes in representation learning (Devlin et al., 2019; Gao et al., 2021) motivate us to purely rely on dropout for our weak augmentation, and random token masking for our strong augmentation.

Specifically, the surface forms of weakly-augmented sentences remain unchanged: $\alpha(u_i) = u_i$. For strong augmentation, we randomly replace 15% of the tokens in $\mathcal{A}(u_i)$ with the special mask token, [MASK]. Then, we input $\alpha(u_i)$ and $\mathcal{A}(u_i)$ to the language model separately. Therefore, the two input sentences will undergo independent dropout operations (0.1 dropout rate by default), which can be considered as part of the data augmentation process (The green arrow in Figure 1). We empirically show that the performance of our proposed augmentation techniques is superior against other common text augmentation techniques in Section 4.

Self-supervised loss: We also include an auxiliary self-supervised loss term, $\mathcal{L}_{ssl}$, for regularization purpose. The masked language model loss is used for its simplicity and efficiency.

4 Experiment

4.1 Setup

We evaluate our model on two groups of tasks: (1) 6 standard single sentence classification tasks (SST-2 (Socher et al., 2013), SST-5 (Socher et al., 2013), MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), MPQA (Wiebe et al., 2005), Subj (Pang and Lee, 2004)) and (2) 6 sentence pair classification tasks (MNLI (Williams et al., 2018), MNLI-mm (Williams et al., 2018), SNLI (Bowman et al., 2015), QNLI (Rajpurkar et al., 2016), RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), MRPC (Dolan and Brockett, 2005)). These tasks are adapted from the benchmarks in (Conneau and Kiela, 2018; Wang et al., 2018).

We set N to 16 and μ to 4. Following (Gao et al., 2020). We randomly sample five different splits of ($\mathcal{X}_{train}$, $\mathcal{X}_{dev}$, $\mathcal{U}$) from the original training set. Five different models are trained with these splits. Then, we report the average performance of these five models on the original development set. As in the previous work (Schick and Schütze, 2020), the sampled unlabeled splits are carefully constructed to account for class balance. As few-shot learning can be unstable, and extremely sensitive to hyper-parameter selection, we also perform a grid search over several hyper-parameters (learning rate, batch size B, controlling weight of loss λ and confidence threshold τ) across different tasks. Finally, Adam (Kingma and Ba, 2015) is used as the optimizer.

4.2 Baselines

We consider three baselines, namely standard fine-tuning (**FT**), supervised learning (Gao et al., 2020) (**LM-BFF**), semi-supervised learning (Schick and Schütze, 2021) (**PET**). We use RoBERTa-base (Liu et al., 2019), which has 125M parameters, and the same task-specific manual prompt from (Gao et al., 2020), including template and word-to-label mapping, for prompt-based fine-tuning.

FT: We directly fine-tune the language model with the sequence classification head on the few-shot labeled dataset.

LM-BFF: We choose current state-of-the-art

	SST-2 (acc)	SST-5 (acc)	MR (acc)	CR (acc)	MPQA (acc)	Subj (acc)
FT	78.3 (3.7)	36.4 (2.1)	69.5 (4.3)	78.6 (4.3)	69.2 (7.2)	89.6 (0.8)
LM-BFF	89.9 (0.5)	45.8 (3.1)	84.1 (1.7)	89.5 (0.6)	84.3 (1.1)	88.3 (3.5)
PET-few♠	89.8 (0.9)	46.7 (0.8)	84.2 (1.2)	89.4 (0.7)	84.9 (0.9)	90.0 (1.7)
PET-full♡	90.2 (0.8)	46.0 (1.2)	85.0 (1.3)	88.9 (1.0)	84.1 (1.0)	91.4 (0.7)
SFLM	**91.0** (0.7)	**47.7** (1.1)	**86.6** (0.4)	**90.8** (0.6)	**86.5** (0.3)	**92.4** (0.7)

	MNLI (acc)	MNLI-mm (acc)	SNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (F1)
FT	41.2 (2.2)	42.9 (2.6)	43.9 (3.1)	60.3 (3.7)	50.3 (1.9)	70.8 (19.9)
LM-BFF	60.2 (1.7)	62.3 (1.4)	65.8 (2.6)	60.6 (2.1)	66.2 (3.4)	77.7 (0.8)
PET-few♠	60.0 (1.8)	61.6 (1.4)	66.8 (2.8)	60.8 (2.5)	62.5 (1.7)	77.4 (5.0)
PET-full♡	**62.6** (2.9)	**64.8** (2.2)	**67.8** (3.5)	**61.3** (4.0)	65.5 (2.3)	77.5 (4.5)
SFLM	**62.6** (1.5)	64.7 (1.3)	67.4 (2.7)	61.0 (4.6)	**67.3** (2.7)	**81.8** (1.2)

Table 1: We use RoBERTa-base and report the average scores in all experiments, where *acc* denotes the accuracy (%), and *F1* denotes the F1 score. The standard deviation is included in the bracket. We use $N = 16$ (# labeled examples per class), and $\mu = 4$ (ratio of unlabeled data to labeled data) for few-shot experiments. Upper block shows the results on single sentence tasks, while lower block shows the results on sentence pair tasks. ♠: we re-implement the PET (Schick and Schütze, 2021) based on LM-BFF (Gao et al., 2020). ♡: We treat the full traning set as the unlabeled dataset and fine-tune PET with it. The size of the full training set is kept to 10,000 samples for all tasks.

LM-BFF (Gao et al., 2020) as our supervised baseline. We use the prompt with demonstration (Gao et al., 2020) implementation across all tasks for fair comparison. We retrain the model with the official code [3].

PET: For fair comparison, We re-implement our own version of PET based on LM-BFF (Gao et al., 2020), since LM-BFF largely benefits from prompt-based fine-tuning. Specifically, we remove the knowledge distillation on the standard sequence classifier in the original implementation. Instead, we fine-tune the prompt-based language model (Gao et al., 2020) with a mixed training set of labeled and pseudo-labeled data. We iteratively increase the amount of pseudo-labeled data in the training set for model fine-tuning. Through our extensive experiments, we find that our implementation outperforms the official implementation[4] across various tasks. In addition, we evaluate PET under two different settings: (1) using reduced unlabeled dataset, which is the same as our SFLM; (2) using the full training set for self-training, while we limit the number of unlabeled samples to 10,000 per class such that the amount of unlabeled samples across different tasks are kept in the same range.

4.3 Main Results

Table 1 presents the performance of SFLM against baselines across various benchmarking tasks. Overall, our proposed SFLM consistently outperforms the supervised and semi-supervised methods by 2% on average with the same amount of data, and by 1.2% with 45 times less unlabeled data. Next, we summarize our observations over the experiment results.

First, we find that self-training can greatly improve the performance of vanilla prompt-based fine-tuning under few-shot setting, either using PET or SFLM. With a large amount of in-domain unlabeled data, even a simple iterative self-training approach can boost the performance by 3.13% on Subj, and 0.88% on average (PET-full vs. LM-BFF). This demonstrates the effectiveness of exploiting the rich information carried in the unlabeled data.

Second, in-domain unlabeled data is crucial to the success of semi-supervised methods. As we down-sample the unlabeled dataset size to 64 samples per class, The performance of PET barely has

[3] https://github.com/princeton-nlp/LM-BFF

[4] https://github.com/timoschick/pet

any improvement w.r.t LM-BFF. The performance even degrades in some tasks. For example, in tasks[5] such as RTE, CR, and MRPC, PET doesn't perform as well as LM-BFF even using the full unlabeled dataset.

Third, unlike PET-few, SFLM can still bring a significant improvement w.r.t LM-BFF even when the size of unlabeled data is limited. This observation implies that our approach utilizes the unlabeled data in a much more efficient way. SFLM also outperforms PET on 8 tasks out of 12. The exceptions are the four natural language inference tasks, of which the unlabeled datasets contain 10,000 unlabeled data samples. However, the performance difference between PET-full, which uses all the 10,000 unlabeled data samples, and SFLM in these four tasks is insignificant (less than 0.4%). In addition, SFLM generally has lower variance compared to the baselines.

The major difference between SFLM and LM-BFF is that we utilize the rich information carried in the unlabeled data. The experiment results confirm our hypothesis about the usefulness of semi-supervised learning in language model few-shot learning. Furthermore, the major difference between SFLM and PET is the self-training algorithm. We include strong data augmentation technique for consistency regularization. This confirms that our proposed text augmentation techniques are crucial to the success of SFLM.

4.4 Analysis of Data Efficiency

One of the key questions in this study is how many labeled and unlabeled data are required. We provide an answer in Figure 2, which illustrates the performance of SFLM with different combinations of N and μ. It can be observed that the error rate reduces generally as μ increases, that suggests SFLM benefits from more unlabeled data, with an exception in SST-5, where the best performance is achieved at $\mu = 2$. a similar trend is also spotted for PET (see Table 1).

We note that SST-5 is the most difficult one among the six tasks. We observe that, with a relatively weak teacher model, self-training doesn't benefit from more unlabeled data. We speculate that by increasing unlabeled data, we introduce more noise into the training process when we have a weak starting point.

[5]The full unlabeled datasets for these tasks contain less than 5,000 sentences.

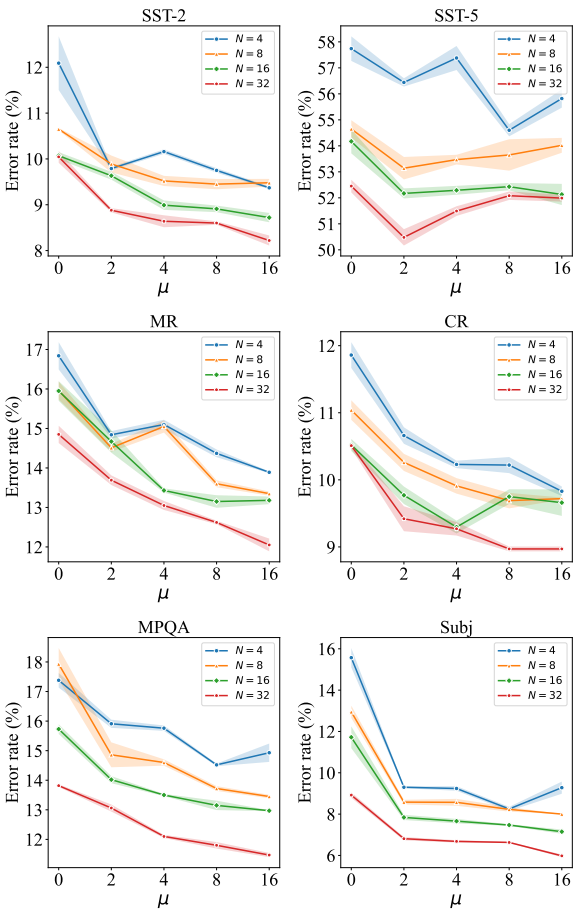

Figure 2: Error rates of our SFLM under different N (# instances per class), and μ (ratio between unlabeled and labeled data)[6]. $\mu = 0$ refers to vanilla LM-BFF.

In Figure 2, we also observe that, for the simpler tasks, e.g., SST-2 and Subj, the gain obtained by more unlabeled data rapidly saturates when four times of unlabeled data are given, which is consistent with the finding in (Sohn et al., 2020). We are encouraged to see that SFLM continues to improve as μ increases for the other three tasks and saturates later. In general, the trend of performance improvement by varying μ is consistent across different value of N. For instance, the performance gain is $\sim 0.65\%$ for increasing μ from 2 to 4 across different N.

To better understand what SLFM actually improves over the baselines, we further analyze the distribution of incorrectly-labelled examples corrected by SFLM, motivated by (Wei et al., 2020). First, we partition incorrectly-labelled examples into five bins based on the cosine similarity of their sentence embeddings given by SimCSE (Gao et al.,

[6]The performance in CR for ($N = 32$, $\mu = 16$) is put as the same as that of ($N = 32$, $\mu = 8$) due to limited amount of unlabeled data

2021) w.r.t. the average embedding of the training set. Next, we show the percentage of examples in each bin whose labels are corrected by SFLM. As shown in Figure 3, SFLM is more likely to correct examples within the gold-labelled data's neighborhood (the region surrounding a data point in the embedding space) than those away from the neighborhood. In other words, the performance gain of SFLM is not only related to the unlabelled data size but also to the data distribution.

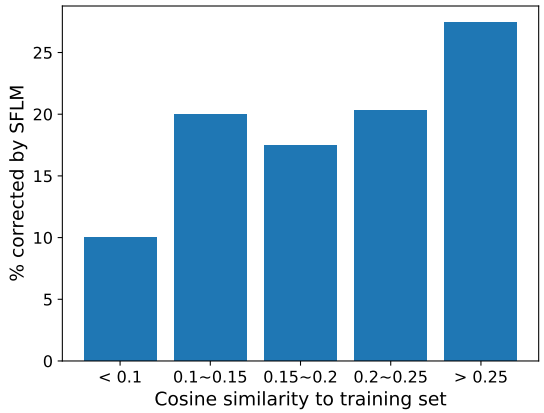

Figure 3: Percentage of examples with varying cosine similarities to training set corrected by SFLM.

SFLM is also scalable with different amounts of labeled data, and consistently outperforms LM-BFF by 3.2% for $N = 4$ and by 2% for $N = 32$ on average. Especially, SFLM exhibits significant improvement over LM-BFF under the extremely few-shot scenario. When $N = 32$, the performance of LM-BFF saturates in simple tasks, e.g., SST-2, CR. However, SFLM continues to narrow the gap between few-shot learning and fine-tuning with the entire labeled dataset, which further validates the benefits of self-training to supervised learning.

4.5 Augmentation Techniques

It has been shown that strong data augmentation plays a crucial role in semi-supervised visual representation learning (Zoph et al., 2020; Xie et al., 2020b; Sohn et al., 2020). The images can be augmented easily by cutout, flipping, and cropping (Zoph et al., 2020). However, very few works have been done on augmentation techniques for text (Xie et al., 2020b)[7].

Here, we study how different augmentation techniques would affect the model performance. We fix *Dropout* as the weak augmentation, and present the results of another three strong augmentation approaches, including *Crop*, *Swap* and *Deletion* in Table 2. Specifically, *Dropout* is same as the weak augmentation, we directly forward the original sentence into the language model. *Crop* refers to randomly cropping the original sentence into a continuous span of 85% of the original length. In terms of *Swap*, we randomly swap two tokens in the sentence and repeat the same procedure 3 times. For *Deletion*, we randomly delete 15% of the tokens in a sentence. *Mask* refers to randomly replacing 15% of tokens in a sentence with the special [MASK] token. In this experiment, we keep $N = 16$ and $\mu = 4$.

We observe that the SFLM framework can also work with *Dropout* and *Swap*, as they still outperform PET on average by 0.4 and 0.2 respectively. However, they are less effective than *Mask*. Another interesting finding is that *Deletion*, which is similar to *Mask*, yields poor performance. We hypothesize that *Deletion* and *Crop* may adversely affect the original semantics, for example, deleting the word, *not*, may reverse the meaning of the original sentence. In contrast, *Mask* keeps the structure of sentences, and hence, it is easier to maintain the semantics by adding consistency regularization and MLM objectives.

Furthermore, we empirically study the effect of different masking ratio on SST-2. 90.62% of accuracy is obtained for 10% masking, 90.14% accuracy for 20% masking, and the best performance of 91.0% for 15% masking.

4.6 Model Scale

To put the SFLM framework under a stress test, we further apply SFLM on a smaller language model, distilled RoBERTa with 84M parameters, as reported in Table 3. While DistilRoBERTa shows a similar performance as RoBERTa, it performs much worse under few-shot learning scenarios, e.g., LM-BFF has a significant performance drop of 5% in CR and MPQA. We hypothesize that a robust language model of reasonable footprint is required for effective self-training.

While SFLM consistently outperforms LM-BFF with a smaller language model, the overall performance of SFLM based on DistilRoBERTa-base also degrades sharply w.r.t that based on RoBERTa-base. This suggests the crucial role of the language model in our approach.

[7]They adopt a back-translation system, which requires external parallel corpus. In SFLM, we focus on a few easy data augmentation strategies (Wei and Zou, 2019)

Augmentation	SST-2	SST-5	MR	CR	MPQA	Subj	Avg
Dropout	90.0	46.0	84.7	89.7	85.7	90.9	81.2
Crop	89.6	45.7	83.6	88.4	85.2	88.8	80.2
Swap	**91.1**	41.8	86.1	89.2	85.4	92.3	81.0
Deletion	90.3	44.1	83.7	89.8	86.3	90.0	80.7
Mask	91.0	**47.7**	**86.6**	**90.8**	**86.5**	**92.4**	**82.5**

Table 2: Comparison of different data augmentation on six single sentence classification datasets (accuracy %).

	CR	MPQA	Subj
*RoBERTa-base (**12-layers**)*			
FT	78.6	69.2	89.6
LM-BFF	89.5	84.3	88.3
PET-few♠	89.4	84.9	90.0
PET-full♡	88.9	84.1	91.4
SFLM	90.8	86.5	92.4
FT (full)	89.6	87.4	96.9
*DistilRoBERTa-base (**6-layers**)*			
FT	71.0	71.7	86.6
LM-BFF	84.8	79.8	87.5
PET-few♠	86.5	80.6	88.3
PET-full♡	87.2	80.6	88.0
SFLM	87.9	82.0	89.5
FT (full)	85.5	86.9	96.6

Table 3: Accuracy (%) for systems with language models of different size. We use $N = 16$ (# labeled examples per class), and $\mu = 4$ (ratio of unlabeled data to labeled data) for few-shot experiments. ♠, ♡ follow the same definition in Table 1. DistilRoBERa-base: 6-layer RoBERTa distilled from RoBERTa-base with 2 times speedup.

Meanwhile, we find that PET better suits the small language model setting. For instance, the performance gain of PET-full w.r.t. LM-BFF increasing from 0.8% to 1.1% under the RoBERTa-base and DistilRoBERTa-base settings respectively. We hypothesize that more unlabeled data benefit few-shot learning of smaller language model. However, the performance of PET is still worse than that of SFLM under the DistilRoBERTa-base setting.

Overall, the above observations confirm the effectiveness of our approach for language model few-shot learning regardless of the backbone language model.

4.7 Zero-shot Transfer Across Tasks

Lastly, we show that our proposed method can be easily extended to zero-shot transfer across tasks. Specifically, we assume that before few-shot learning on the target unlabeled dataset $\mathcal{U}$, the learner has access to an annotated dataset $\mathcal{D}$ known as the base dataset.[8] Accordingly, we modify the learning objective as:

$$\mathcal{L} = \mathcal{L}_s^{\mathcal{D}} + \lambda_1 \mathcal{L}_{st}^{\mathcal{U}} + \lambda_2 \mathcal{L}_{ssl}^{\mathcal{U}}, \tag{8}$$

where the last two terms are intended to encourage the model to capture knowledge specific to the target task.

We evaluate SFLM on three binary classification tasks, including sentiment analysis (MR and SST) and product reviews (CR). In the experiments, we use one of the three tasks as the source task and test on another two target tasks. The SFLM model is based on RoBERTa-base and trained on 64 samples per class for $\mathcal{D}$ and $\mathcal{U}$, respectively. We compare SFLM with a baseline, which is a language model fine-tuned directly on the source dataset with a single prompt-based loss $\mathcal{L}_s^{\mathcal{D}}$.

The results are reported in Table 4. SFLM, which adopts a small number of unlabeled target data, generally outperforms that with supervised fine-tuning on source datasets. In particular, with MR as the source and CR as the target, SFLM obtains 1.3 points higher than the baseline model, which demonstrates that our proposed method has the flexibility to be applied to the cross-task zero-shot learning scenario.

5 Conclusion and Future Work

In this paper, we present SFLM, a simple and effective self-training framework for few-shot learning of language model. Our approach addresses the

[8] The concept of task transfer has been successfully applied to natural language understanding. For example, pre-training language model on MNLI before fine-tuning on RTE (Liu et al., 2019) yields better performance.

	Transfer			SFLM		
	SST	MR	CR	SST	MR	CR
SST	-	87.3	89.7	-	87.3	90.8
MR	90.7	-	89.5	91.4	-	90.8
CR	89.7	84.5	-	90.3	85.5	-

Table 4: Accuracy (%) for zero-shot task transfer between three sentiment classification tasks with RoBERTa-base. Transfer: a language model fine-tuned on source dataset with a single prompt-based loss; We refer SST to SST-2.

few-shot language model fine-tuning problem with very limited labeled and unlabeled data with a self-training loss term unifying pseudo-labeling and consistency regularization. Through comprehensive experiments, we show that our approach outperforms previous state-of-the-art methods across tasks, data amount, and model scale. Despite its efficiency, SFLM also has several limitations. Compared to standard fine-tuning, SFLM requires more computational resources for unlabeled data. In addition, the performance gain by self-training is not proportional to the amount of unlabeled data. We leave it to future study. Namely, how to utilize large amount of unlabeled data efficiently through self-training.

Acknowledgements

We would like to thank all the anonymous reviewers for their constructive comments. This work is partly supported by Human-Robot Interaction Phase 1 (Grant No. 19225 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore; National Natural Science Foundation of China (Grant NO. 61903178, 61906081, and U20A20306); Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT03X386); Program for University Key Laboratory of Guangdong Province (Grant No. 2017KSYS008).

References

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *Proc. of ICLR*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. of COLT*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

T. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Proc. of NeurIPS*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *Proc. of ICLR*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

E. D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2018. Autoaugment: Learning augmentation policies from data. *ArXiv*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. page 177–190.

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

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Jingfei Du, E. Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. In *ArXiv*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of ICML*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *ArXiv*, abs/2011.01403.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proc. of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. In *Proc. of ICLR*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proc. of ICLR*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *ArXiv*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Zimeng Qiu, Eunah Cho, Xiaochun Ma, and William Campbell. 2019. Graph-based semi-supervised learning for natural language understanding. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 151–158, Hong Kong. Association for Computational Linguistics.

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

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. In *Proc. of NAACL*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proc. of NeurIPS*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. of NeurIPS*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proc. of NeurIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2020. Theoretical analysis of self-training with deep networks on unlabeled data. In *Proc. of ICLR*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. In *Proc. of NeurIPS*.

Qizhe Xie, E. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *Proc. of CVPR*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaojin Zhu. 2005. Semi-supervised learning literature survey. *world*.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. 2020. Rethinking pre-training and self-training. In *Proc. of NeurIPS*.