From Mimicking to Integrating: Knowledge Integration for Pre-Trained Language Models

Lei Li[1], Yankai Lin[2,3], Xuancheng Ren[1], Guangxiang Zhao[1], Peng Li[4], Jie Zhou[5], Xu Sun[1]
[1]MOE Key Lab of Computational Linguistics, School of Computer Science, Peking University
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods , Beijing, China
[4]Institute for AI Industry Research (AIR), Tsinghua University, China
[5]Pattern Recognition Center, WeChat AI, Tencent Inc., China
lilei@stu.pku.edu.cn xusun@pku.edu.cn

Abstract

Investigating better ways to reuse the released pre-trained language models (PLMs) can significantly reduce the computational cost and the potential environmental side-effects. This paper explores a novel PLM reuse paradigm, Knowledge Integration (KI). Without human annotations available, KI aims to merge the knowledge from different teacher-PLMs, each of which specializes in a different classification problem, into a versatile student model. To achieve this, we first derive the correlation between virtual golden supervision and teacher predictions. We then design a Model Uncertainty–aware Knowledge Integration (MUKI) framework to recover the golden supervision for the student. Specifically, MUKI adopts Monte-Carlo Dropout to estimate model uncertainty for the supervision integration. An instance-wise re-weighting mechanism based on the margin of uncertainty scores is further incorporated, to deal with the potential conflicting supervision from teachers. Experimental results demonstrate that MUKI achieves substantial improvements over baselines on benchmark datasets. Further analysis shows that MUKI can generalize well for merging teacher models with heterogeneous architectures, and even teachers major in cross-lingual datasets.[1]

1 Introduction

Large-scale pre-trained language models (PLMs), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) have recently achieved promising results after fine-tuning on various natural language processing (NLP) tasks. Many fine-tuned PLMs are generously released for facilitating researches and deployments. Reusing these PLMs can greatly reduce the computational cost of retraining the PLM from scratch and alleviate the potential environmental side-effects like carbon footprints (Strubell et al., 2019), thus making NLP systems greener (Schwartz et al., 2020). A commonly adopted model reuse paradigm is knowledge distillation (Hinton et al., 2015; Romero et al., 2015), where a student model learns to mimic a teacher model by aligning its outputs to that of the teacher. In this way, though achieving promising results with PLMs (Sun et al., 2019; Jiao et al., 2020), the student is restricted to perform the same task as the teacher model, thus restricting re-utilization of abundant available PLMs fine-tuned on different tasks, e.g., models fine-tuned on various label sets or even different datasets.

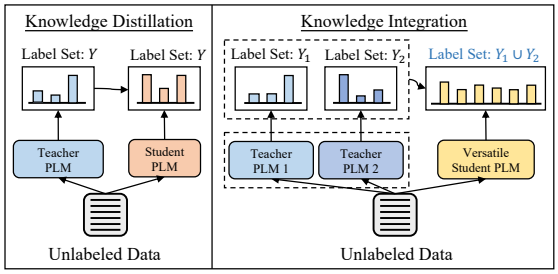

Figure 1: Comparison of knowledge distillation (KD) and knowledge integration (KI). KD assumes that the student performs predictions on the identical label set with the teacher, while KI trains a student model that is capable of performing classification over the union label set of teacher models.

In this paper, we generalize the idea of KD from mimicking teachers to integrating knowledge from teachers, and propose *Knowledge Integration* (KI) for PLMs. Given multiple fine-tuned teacher-PLMs, each of which is capable of performing classification over a unique label set, KI aims to train a versatile student that can make predictions over the union of teacher label sets. As the labeled data for training the teachers may not be publicly released due to data privacy issues, we assume no human annotations are available during KI. The benefits of KI are two-fold. First, compared to KD, KI can make full use of the released PLMs

[1]Our code is available at https://github.com/lancopku/MUKI. Part of the work was done while Yankai Lin and Peng Li were working at Tencent.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6420 - 6431
December 7-11, 2022

specializing different tasks. Besides, the ability of the versatile student, i.e., the label set coverage, can be improved over time by integrating newly released teacher models. Figure 1 illustrates the main difference between KD and KI.

As no annotations are available, the core challenge of KI lies in the integration of outputs from teachers to form golden supervision, i.e., the class probability distribution over the union label set, for guiding the student. Through theoretical derivation, we first build the bridge between the teacher predictions and the golden supervision, which indicates that the key to recovering such supervision is to identify the adequate teacher for each instance. However, due to the over-confident problem of PLMs (Desai and Durrett, 2020), selecting qualified teachers for unlabeled instances is non-trivial, and our exploration shows that prediction entropy is misleading. Inspired by Monte-Carlo Dropout (Gal and Ghahramani, 2016), we inject parameter perturbations to the teacher models during inference and then estimate the model uncertainties over averaged predictions for indicating the possible correct teacher model. Our Model Uncertainty–aware Knowledge Integration (MUKI) framework is then proposed based on the estimated model uncertainty. Specifically, the golden supervision is approximated by either taking the outputs of the most confident teacher, or softly integrating different teacher predictions according to the relative importance of each teacher. Furthermore, for instances on which teachers achieve close uncertainty scores, we introduce a re-weighting mechanism based on the margin of uncertainty scores, to down-weight the contribution of instances with potential conflicting supervision signals.

Experimental results show that MUKI can successfully achieve the goal of knowledge integration, significantly outperforming baseline methods, and even obtaining comparable results with models trained with labeled data. Further analysis shows that MUKI can produce supervision close to the golden one and generalize well for merging knowledge from heterogeneous teachers with different architectures, or even cross-lingual teacher models.

The main contributions of this work can be summarized as follows: (1) We explore knowledge integration for PLMs, which is capable of making full use of released PLMs with different label sets and has great extendability. (2) We present MUKI, a generalizable KI framework, which integrates the knowledge from teachers according to model uncertainty estimated via Monte-Carlo Dropout and re-weights the instance contribution based on the uncertainty margin. (3) Experimental results demonstrate that MUKI is effective and generalizable, significantly outperforming baselines.

2 Knowledge Integration for PLMs

In this section, we first give the task formulation for knowledge integration, followed by the elaboration on the proposed MUKI framework.

2.1 Problem Formulation

Given N teacher PLM models $TS = \{T_1, \ldots, T_N\}$, where each teacher T_i specializes in a specific classification problem, i.e., a set of classes Y_i, knowledge integration aims to train student model S for performing predictions over the comprehensive class set $Y = \bigcup_{i=1}^{N} Y_i$, with an unlabeled dataset $\mathcal{D}$. We assume that for each instance in $\mathcal{D}$ there is at least one teacher capable of handling it and we focus on a practical setting where the teacher specialties are totally disjoint, i.e., $Y_i \cap Y_j = \emptyset, \forall i \neq j$, as merging teachers with overlapping classes can be easily converted in to the disjoint situation.

2.2 Model Uncertainty–Aware Knowledge Integration

As there are no annotated data available due to the data privacy issue, we need to construct supervision for guiding the student. Given a golden label distribution $\mathcal{T}(x)$ for each instance x over Y, we can train the student by minimizing the KL-divergence:

$$\mathcal{L} = \sum_{x \in \mathcal{D}} \mathrm{KL}\left(S\left(x\right) || \mathcal{T}\left(x\right)\right), \qquad (1)$$

where $S(x)$ denotes the output distribution of the student for input x. As we only operate on the output distribution level, thus this framework is generalizable for PLMs that potentially differ in the model architectures and training data distribution. To estimate the golden supervision $\mathcal{T}(x)$, we first derive the correlation between $\mathcal{T}(x)$ and the prediction $T_i(x)$ of teacher model T_i. Specifically, as teacher T_i specializes in label set Y_i, it can only predict $T_i(y \mid x)$ for instance x when $y \in Y_i$. Therefore, the correlation between $T_i(y \mid x)$ and global probability $\mathcal{T}(y \mid x)$ over the full class set

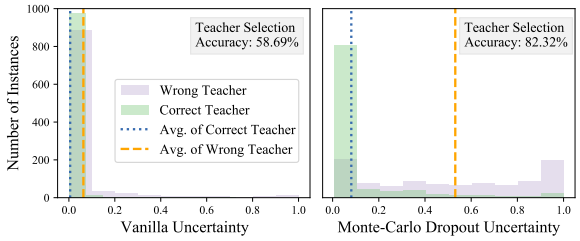

Figure 2: Model uncertainty (normalized) distributions evaluated with 1000 instances randomly sampled from the AG News dataset. The vanilla prediction entropy distributions of two teacher models overlaps greatly (left), while Monte-Carlo Dropout produces a more accurate uncertainty approximation for distinguishing the correct teacher model (right). Best viewed in color.

can be derived as:

$$T_i(y \mid x) = \mathcal{T}\left(y \mid x, y \in Y_i\right) \tag{2}$$

$$= \frac{\mathcal{T}\left(y, y \in Y_i \mid x\right)}{\mathcal{T}\left(y \in Y_i \mid x\right)}. \tag{3}$$

The above derivation indicates that we can recover the golden probability distribution by (1) getting the teacher predictions, and (2) estimating the denominator, which means how likely the instance x lies in the teacher T_i specialty Y_i. As instances associated with classes not in Y_i can be treated as the out-of-distribution data for the teacher T_i, the teacher predictions would be more uncertain about these instances than that of in-distribution instances (Hendrycks and Gimpel, 2017). We thus propose to approximate the denominator in an opposite direction, i.e., estimating how likely the instance is not belong to teacher T_i via model uncertainty. Followingly, we first explore different uncertainty estimations for recovering the golden supervision, and then introduce how we incorporates teacher predictions according to the estimated uncertainty scores.

2.2.1 Uncertainty Estimation

A naïve estimation is directly taking the statics like prediction entropy of predicted class distribution. However, due to the over-confident issues of over-parameterized models like PLMs (Guo et al., 2017; Desai and Durrett, 2020), this simple estimation can be unreliable. We investigate this by first splitting the instances of the AG News dataset (Zhang et al., 2015) into two sets with disjoint labels, and then fine-tuning teacher models on each set separately. For each instance, there is a correct teacher that is capable of handling it and a wrong teacher that is not qualified for processing it. We plot the prediction entropy distributions of the correct teacher and the wrong teacher in the left part of Figure 2. It can be found that the wrong teacher also produces confident predictions even for instances that are not in its speciality with nearly zero uncertainty scores, exhibiting a great overlap with the correct teacher model. This indicates that utilizing the simple metric will mislead the identification of the adequate teacher. To remedy this, inspired by recent progress in Bayesian neural networks (Blundell et al., 2015; Gal and Ghahramani, 2016), we propose to add small perturbations to the model weights during inference to find out the correct teacher model. The intuition behind is that, as the instance is well fitted by the parameter of the qualified teacher model, the teacher can produce confident results consistently in the multiple predictions even with small perturbed parameters. On the contrary, small perturbations on the model weights of the wrong teacher will lead to a drastic change in the output probabilities, resulting in more uncertain predictions on average. Therefore, we can estimate the model uncertainty more accurately according to the average predictions under parameter perturbations. Specifically, we adopt Monte-Carlo Dropout (Gal and Ghahramani, 2016), where the output distribution of an instance x with T_i is calculated as:

$$p_i\left(y \mid x, \mathcal{D}\right) \approx \frac{1}{K}\sum_{k=1}^{K} p_i\left(y \mid \mathbf{W}_k^i, x\right) \tag{4}$$

$$= \frac{1}{K}\sum_{k=1}^{K} T_i\left(x, \mathbf{W}_k^i\right), \tag{5}$$

where $\mathbf{W}_k^i$ is the k-th masked weights of T_i sampled from the Dropout distribution (Srivastava et al., 2014), and K is the sampling number. The model uncertainty of teacher model T_i thus can be summarized as the entropy of the averaged probability distribution p_i:

$$u_i = H\left(p_i\right) = -\sum_{y=1}^{|Y_i|} p_i^y \log p_i^y. \tag{6}$$

As shown in the right part of Figure 2, the uncertainty distributions of the correct teacher and the wrong teacher model estimated via Monte-Carlo Dropout exhibit a clearer difference than vanilla prediction entropy, indicating its great potential for guiding the probability combination.

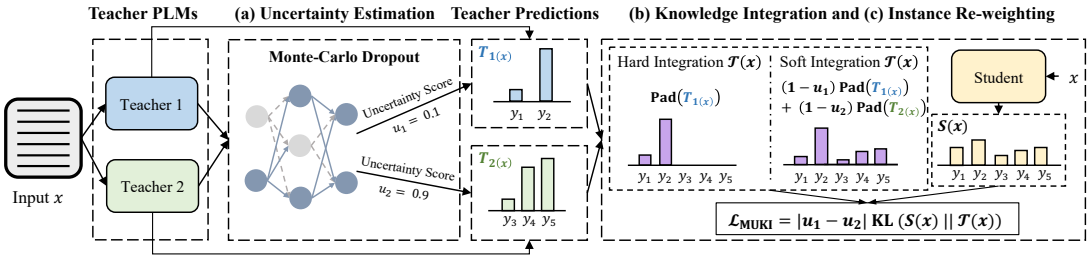

Figure 3: Overview of the proposed MUKI framework, which consists of (a) Model uncertainty scores estimation with Monte-Carlo Dropout. (b) Knowledge integration for estimating the golden supervision according to uncertainty scores with (c) instance-wise re-weighting mechanism. Best viewed in color.

2.2.2 Knowledge Integration

With the accurately estimated teacher uncertainties $U = \{u_1, \ldots, u_N\}$ for each instance at hand, we design two methods for approximating the golden supervision to guide the student model:

MUKI-Hard which directly takes the supervision provided by the teacher with the lowest uncertainty as the golden distribution:

$$i^* = \arg\min_i u_i / \log |Y_i| \tag{7}$$

$$\mathcal{T}(x) \approx \text{Pad}\left(T_{i^*}(x)\right) \tag{8}$$

$$\text{Pad}\left(T_{i^*}(y \mid x)\right) = \begin{cases} T_{i^*}(y \mid x) & y \in Y_{i^*} \\ 0 & y \in Y - Y_{i^*} \end{cases} \tag{9}$$

where $\log |Y_i|$ is a normalizing factor. As T_{i^*} only provides the label relation over the class set Y_{i^*}, the probabilities of classes not in Y_{i^*} are set to zeros, denoted by the Pad operation. In this way, we are actually set $\mathcal{T}(y \in Y_{i^*} \mid x) = 1$, and thus the student can learn from the teacher model that is most confident about x.

MUKI-Soft which estimates the golden supervision as a weighted sum of teacher model predictions by taking the relative uncertain level into consideration:

$$c_i = 1 - u_i / \log |Y_i| \tag{10}$$

$$w_i = \frac{\exp\left(c_i/\tau\right)}{\sum_{j=1}^{N} \exp(c_j/\tau)} \tag{11}$$

$$\mathcal{T}(x) \approx \sum_{i=1}^{N} w_i \text{Pad}\left(T_i(x)\right) \tag{12}$$

where c_i denotes the confidence score which indicates how likely x belongs to C_i, and τ is a hyper-parameter for controlling the smoothness of the weights. In this way, the teacher with a higher confidence score contributes more to the estimated golden supervision signal. Besides, the difference between the confidence scores reflects the inner correlation between the classes in different label groups, thus providing extra information for the classes in disjoint label sets.

2.2.3 Instance Re-weighting

Furthermore, the uncertainty distribution overlapping in the right part of Figure 2 indicates that there is still a small portion of instances on which teacher models achieve similar confidence levels with Monte-Carlo Dropout. For these instances, MUKI-Hard may wrongly select the supervision source, and MUKI-Soft would assign close weights to all teacher predictions, thus providing a vague even conflicting supervision signal. To remedy this, we devise an instance re-weighting mechanism by modifying the objective in Eq. (1):

$$\mathcal{L}_{\text{MUKI}} = \sum_{x \in \mathcal{D}} v(x) \,\text{KL}\left(S(x) \,||\, \mathcal{T}(x)\right) \tag{13}$$

$$v(x) = c_{\text{max}} - c_{\text{sec}}, \tag{14}$$

where c_{max} and c_{sec} denotes the largest and the second large teacher confidence score for instance x, respectively. By minimizing the instance-level weighted objective, the student is encouraged to focus more on the pivotal instances with clearer supervision signals, thus reducing the effect of potential confusing instances.

In summary, MUKI consists of supervision estimation and instance re-weighting mechanism based on model uncertainty for integrating the knowledge from different teachers. Figure 3 gives an overview of MUKI framework.

3 Experiments

3.1 Experimental Settings

Datasets We conduct evaluations on four classic text classification benchmarks, including three

Method	Model Size	AG News	THUCNews	Google Snippets	5Abstracts Group	Average
Supervised	110M	94.6 ± 0.00	97.8 ± 0.00	89.3 ± 0.00	90.7 ± 0.00	93.10
Teacher 1	110M	49.9 ± 0.00	48.8 ± 0.00	50.2 ± 0.00	42.0 ± 0.00	47.73
Teacher 2	110M	47.5 ± 0.00	49.8 ± 0.00	43.5 ± 0.00	51.5 ± 0.00	48.08
Ensemble	220M	59.8 ± 0.00	93.1 ± 0.00	80.4 ± 0.00	62.3 ± 0.00	73.90
Vanilla KD	110M	63.1 ± 0.81	94.9 ± 0.18	83.7 ± 1.30	67.0 ± 1.14	77.18
DFA	110M	66.4 ± 2.33	94.4 ± 0.22	82.6 ± 0.18	57.7 ± 4.41	72.78
CFL	110M	61.4 ± 1.18	95.1 ± 0.21	84.5 ± 0.45	61.6 ± 0.12	75.65
UHC	110M	78.8 ± 1.42	92.1 ± 0.63	86.3 ± 0.39	71.4 ± 0.67	82.15
MUKI-Hard (Ours)	110M	87.0 ± 0.40	**97.2** ± 0.12	**88.4** ± 0.32	79.0 ± 0.82	**87.90**
MUKI-Soft (Ours)	110M	**87.1** ± 0.19	**97.2** ± 0.08	87.9 ± 0.32	**79.3** ± 0.85	87.88

Table 1: Comparisons on the benchmark datasets. The results are classification accuracy averaged by three seeds, and standard deviations are reported. Both MUKI variants achieve statistically significant improvements over the best-performing baselines ($p < 0.01$). Best results are shown in bold.

English datasets: AG News (Zhang et al., 2015), Google Snippets (Phan et al., 2008), 5Abstracts Group (Liu et al., 2018), and a Chinese dataset THUCNews (Sun et al., 2016). We randomly split 5% of data from the training set for datasets without a validation set to form a validation set for model selection. The statistics of datasets can be found in Appendix A.

Compared Methods We implement various baselines to evaluate our proposal, as follows:

Simple Baselines, which require no additional training, including: (1) Original Teacher: The teacher models are used independently for prediction. We set the probabilities of classes out of the teacher speciality to zeros. (2) Ensemble: The output logits of teachers are directly concatenated for predictions over the union label set.

Distillation Methods, which assume internal states of the teacher model are available and the student is trained via aligning the states of teacher models on $\mathcal{D}$, including: (1) Vanilla KD (Hinton et al., 2015): The student is trained to mimic the soft targets produced by logits combination of all teacher models, via minimizing the vanilla KL-divergence objective. (2) DFA (Shen et al., 2019): DFA designs a layer-wise feature adaptation mechanism for providing extra guidance based on Vanilla KD. The student aligns its features to the merged features of multiple teachers layer by layer. (3) CFL (Luo et al., 2019): CFL first maps the hidden representations of the student and the teachers into a common space. The student is trained by aligning the mapped features to that of the teachers, with supplemental supervision from the logits combination. (4) UHC (Vongkulbhisal et al., 2019): UHC splits the student logits into subsets corresponding to the class sets of teacher models. Each subset is trained to mimic the corresponding output of the teacher model.

A supervised learning method with labeled data is also included, to serve as a performance upper-bound for better understanding of the results.

Implementation Details We implement our framework using the HuggingFace transformers library (Wolf et al., 2020). In our main setting, we set the teacher number N to 2 and we explore integrating multiple teachers in Section 3.4. For each dataset, the classes are randomly split into two non-overlapping parts, and two teachers are fine-tuned on each set separately to imitate the actual applications. Detailed class split can be found in Appendix A. The teacher and student models for English datasets and THUCNews are BERT-base-uncased (Devlin et al., 2019) and BERT-wwm-ext (Cui et al., 2020), respectively. We first fine-tune the teacher models with the split labeled data for 3 epochs with a learning rate 2×10^{-5}. The trained teacher model weights are frozen during the student training process. We set the forward number K of Monte-Carlo Dropout uncertainty estimation to 16 and the dropout rate is set to 0.1. Temperature τ in Eq. (11) is set to 0.2 according to our hyper-parameter analysis results in Appendix B. The student model then is learned by optimizing the KL-divergence objective for 3 epochs, with a 2×10^{-5} learning rate and 32 batch size. The student is evaluated on the validation set every 100 step. We select the best performing checkpoints for final evaluation. The results are replicated with 3 random seeds and we report the averaged accuracy.

Method	AG News	THUCNews
MUKI-Hard	87.0 ± 0.40	97.2 ± 0.12
w/o Monte-Carlo Dropout	65.1 ± 1.67	97.0 ± 0.13
w/o Instance Re-weighting	85.5 ± 0.51	96.9 ± 0.06
MUKI-Soft	87.1 ± 0.19	97.2 ± 0.08
w/o Monte-Carlo Dropout	74.2 ± 0.28	95.4 ± 0.20
w/o Instance Re-weighting	86.7 ± 0.30	96.8 ± 0.03

Table 2: Ablation analysis of MUKI. The removed modules both lead to deteriorated performance.

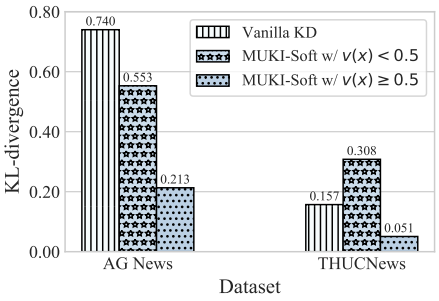

Figure 4: Supervision quality measured by the KL-divergence to the golden supervision of different methods. For MUKI-Soft, we divide instances into two groups according to the uncertainty margin $v(x)$.

3.2 Main Results

The model performance comparison on the four datasets and the corresponding model size are listed in Table 1. Our findings are: (1) Simple baselines fall far behind, showing that it is necessary to design a proper integration strategy for amalgamating the knowledge from different teachers. (2) While extra feature alignment objectives are adopted, DFA and CFL cannot achieve consistent improvements over Vanilla KD. We speculate the reason is that the supervision based on feature alignments is unstable, as teacher features are fine-tuned for specializing in different semantic classes. (3) UHC achieves better average results than Vanilla KD, while performing relatively worse on the THUCNews dataset. It indicates there exists potential supervision conflict as UHC matches the student output independently to that of teachers, thus limiting its generalizability on different datasets. (4) Two variants of MUKI both significantly outperform the previous baseline models on all the datasets, and the average accuracy of MUKI-Hard is achieves a 5.75 points gain over the best performing baseline model. On the THUCNews dataset, while no label information is included during the knowledge amalgamation, MUKI can obtain a 97.2 accuracy, which is very close to 97.8 of the supervised learning method. We attribute the success to that MUKI provides the student with the accurately estimated golden probability distribution over the union label set according to model uncertainty, which can effectively transfer the knowledge and alleviate potential supervision conflicting. These promising results indicate that our MUKI framework can produce better supervision for training the student model, thus has great potentials for reusing PLMs with different label sets.

3.3 Ablation Studies

We conduct ablation experiments on two large datasets, i.e., AG News and THUCNews for stable results, to explore the following two questions.

How Monte-Carlo Dropout benefits knowledge source identification? We replace the Monte-Carlo Dropout estimation of model estimation with a single forward estimation, i.e., setting $K = 1$ in Eq. (5). As shown in Table 2, we find that the performance is degraded on both datasets. Interestingly, we find that the accuracy drop is much clearer on the AG News than that on the THUCNews. To explore this, we compute the average ECE score (Guo et al., 2017) of two teacher models on the out-of-distribution samples, where higher ECE scores indicate more severe over-confident predictions. The teacher models of AG News achieve an average 45.42 ECE score, while that of THUCNews is 19.52. Therefore, the teacher models of AG News exhibit a much more serious over-confident issue than that of THUCNews. This result verifies that our adoption of the Monte-Carlo Dropout technique is effective for accurately identifying the adequate teacher model, especially when teachers tend to make over-confident predictions.

How instance-wise re-weighting benefits supervision integration? In Table 2, we find that removing the instance re-weighting mechanism leads to deteriorated results for both MUKI variants. We further probe whether the re-weighting mechanism is capable of resolving the vague supervision issue when teachers achieve similar uncertainty scores. Specifically, we train a PLM with all labeled data as the proxy of an oracle model, which thus can provide golden supervision over the union label set. We then calculate the KL-divergence between the golden probability distribution and the approximated one of different combination methods. Lower KL-divergence indicates the combined predictions are more correct. We discard results of MUKI-Hard as KL-divergence is not defined

Method	3 Teachers {3,3,4}	4 Teachers {2,2,2,4}	5 Teachers {2,2,2,2,2}
Teacher 1	29.9 ± 0.00	20.0 ± 0.00	20.0 ± 0.00
Teacher 2	29.8 ± 0.00	19.1 ± 0.00	19.1 ± 0.00
Teacher 3	39.8 ± 0.00	19.9 ± 0.00	19.9 ± 0.00
Teacher 4	N / A	39.8 ± 0.00	20.0 ± 0.00
Teacher 5	N / A	N / A	20.0 ± 0.00
Ensemble	91.0 ± 0.00	74.0 ± 0.00	80.6 ± 0.00
Vanilla KD	92.5 ± 1.13	76.5 ± 0.57	82.6 ± 1.64
DFA	91.1 ± 1.25	79.6 ± 1.34	82.0 ± 2.04
CFL	93.9 ± 1.07	77.4 ± 1.34	84.0 ± 0.24
UHC	84.4 ± 0.91	71.6 ± 2.86	69.4 ± 2.73
MUKI-Hard	**94.7** ± 0.20	93.4* ± 0.58	**90.5*** ± 0.32
MUKI-Soft	**94.7** ± 0.14	**93.6*** ± 0.30	**90.5*** ± 1.13

Table 3: Results of merging multiple teacher models on the THUCNews dataset. * denotes the improvement over the best performing baseline is significant ($p < 0.05$). N / A means that the teacher model does not exist in the corresponding setting.

Method	T_1: *BERT-base*	T_2: *BERT-large*
	AG News	THUCNews
Teacher 1	49.7 ± 0.00	48.8 ± 0.00
Teacher 2	47.2 ± 0.00	49.8 ± 0.00
Ensemble	76.6 ± 0.00	79.3 ± 0.00
Vanilla KD	79.6 ± 0.22	82.9 ± 1.36
DFA	78.2 ± 0.30	84.7 ± 1.74
CFL	75.9 ± 0.63	81.9 ± 1.37
UHC	78.3 ± 2.65	92.3 ± 1.08
MUKI-Hard	78.3 ± 1.58	**95.4*** ± 0.45
MUKI-Soft	**80.6** ± 0.17	**95.4*** ± 0.29

Table 4: Results of merging BERT-base and BERT-large. * denotes results are statistically significant ($p < 0.05$).

for distributions with zeros, and divide MUKI-Soft into two groups according to the uncertainty score margin $v(x)$, i.e. instances with $v(x) \geq 0.5$ and that with $v(x) < 0.5$.

As shown in Figure 4, we find that the predictions of instances with $v(x) \geq 0.5$, are much closer to the golden distributions than the Vanilla KD. This indicates that the estimated supervision of instances with a clearer confidence margin is of higher quality, thus paying more attention to these instances is effective for knowledge integration.

3.4 Results in Challenging Settings

KI with Multiple Teachers As MUKI is agnostic to the number of teacher models, we explore its adaptability with more teacher models. We conduct experiments on the THUCNews dataset as it has 10 classes, allowing us to train up to 5 teacher models specialized in different class sets. The 10 classes are split into $\{3, 3, 4\}$, $\{2, 2, 2, 4\}$ and $\{2, 2, 2, 2, 2\}$ for 3, 4 and 5 teachers, respectively. As shown in Table 3, the proposed MUKI framework generalizes well to this setting, outperforming previous baselines with a clear margin. Besides, we find that all baselines based on logits alignment perform poorly under the 4-teacher scenario. We attribute it to that when some teacher models have more classes than others, they usually produce a larger range of logits to make the prediction more distinguishable. Directly combing the teacher logits thus leads to a biased probability distribution. Our MUKI instead estimates the golden supervision according to model uncertainty scores, thus producing better supervision even teacher models exhibit different logit scales.

KI with Heterogeneous Teachers As MUKI only operates on the output distribution level, it is generalizable for heterogeneous teachers. We verify this by merging teachers with different model structures. Specifically, we adopt BERT-base (12 layers and 768 hidden units) and BERT-large (24 layers and 1024 hidden units) as the teachers, respectively. As shown in Table 4, we find that while a larger teacher tends to perform better, the student model performs worse on the THUCNews dataset than learning from two BERT-base teachers, indicating it is challenging to integrate knowledge in this setting. Our MUKI achieves the best results on these two datasets, showing its effectiveness for heterogeneous teachers.

KI with Cross-Dataset Teachers Specifically, we fine-tune teacher models on different datasets separately and then train a student to perform classification over the union label set of both datasets. The multilingual BERT-base is adopted for the teachers and the student in the cross-lingual setting. The results of merging knowledge from two English datasets, AG News and Google-Snippets and even cross-lingual datasets, AG News (in English) and THUCNews (in Chinese) are listed in Table 5. We find that MUKI still outperforms previous baseline models in both settings. Interestingly, we find that the MUKI-Hard is consistently better than MUKI-Soft in this setting. We speculate the reason is that the correlations between classes of different datasets are weak, thus modeling the label relation in these disjoint groups is unnecessary.

Method	T_1: *AG News (en)* T_2: *GS (en)*	T_1: *AG News (en)* T_2: *THUCNews (zh)*
Teacher 1	26.8 ± 0.00	0.50 ± 0.00
Teacher 2	63.0 ± 0.00	54.6 ± 0.00
Ensemble	74.3 ± 0.00	54.5 ± 0.00
Vanilla KD	75.1 ± 0.58	54.6 ± 0.05
DFA	74.7 ± 0.30	54.1 ± 0.20
CFL	74.0 ± 0.50	54.3 ± 0.65
UHC	72.0 ± 0.73	53.5 ± 0.17
MUKI-Hard	**76.3** ± 0.54	**68.1***± 0.20
MUKI-Soft	75.9 ± 0.47	65.6*± 0.40

Table 5: Cross-dataset results. GS is short for Google Snippets. (Left Column) Integrating teachers major in AG News and GS, respectively. (Right Column) Merging teachers major in AG News (in English) and THUCNews (in Chinese), respectively. * denotes results are statistically significant with $p < 0.05$.

Method	CoNLL 2003	OntoNotes 5.0
Teacher 1	55.4 ± 0.00	46.20 ± 0.00
Teacher 2	38.6 ± 0.00	25.34 ± 0.00
Ensemble	79.5 ± 0.00	49.85 ± 0.00
Vanilla KD	83.6 ± 0.66	53.33 ± 0.00
DFA	83.9 ± 0.45	52.66 ± 0.13
CFL	83.1 ± 0.79	53.39 ± 0.63
UHC	81.4 ± 0.24	55.17 ± 1.20
MUKI-Hard	**85.3***± 0.26	**59.83****± 0.31
MUKI-Soft	84.5 ± 0.30	59.33 ± 0.80

Table 6: F1-score results on two NER datasets. The results are statistically significant (* for $p < 0.05$ and ** for $p < 0.01$). MUKI outperforms all the baselines significantly, verifying its generalizability for complicated tasks like structured prediction.

3.5 Results for Structured Prediction

We extend the KI framework into a classic structured prediction task, i.e., named entity recognition (NER). The problem is modeled as a tagging problem following Devlin et al. (2019), and we conduct evaluations on CoNLL 2003 (Sang and De Meulder, 2003) and OntoNotes 5.0 (Pradhan et al., 2013). Specifically, we split the entity types of the dataset into two groups and train two teachers responsible for identifying the entities in each group, respectively. We refer readers to Appendix A for the detailed dataset statistics and the division of entity types. We adapt MUKI to the NER task by estimating the model uncertainty and integrating the predictions at the token level. Besides, we notice that the teacher will predict the non-entity tag with high confidence for tokens of entities it cannot handle. Therefore, we adjust the uncertainty estimation procedure by calculating the entropy of probability distribution over the tags of entity types. The knowledge integration remains the same with the classification problem. As shown in the in Table 6, our MUKI still performs the best among all methods, validating that the proposed framework is generalizable for structured predictions tasks like NER.

4 Related Work

Our work is mainly related to knowledge distillation (KD), which aims to transfer the knowledge from a teacher model to a student model. Hinton et al. (2015) utilize the soft labels of the teacher model for the student to learn, and Romero et al. (2015) align the internal representations between the student and the teacher. Recent studies apply KD for PLMs successfully by matching the intermediate states (Sun et al., 2019; Wang et al., 2020) and enriching the training data with data augmentation (Jiao et al., 2020; Liang et al., 2021), learning from multiple teachers (Wu et al., 2021), and dynamically adjusting the learning objectives (Li et al., 2021). Nevertheless, all these KD studies assume that the student has an identical label set with the target teacher model(s). Instead, knowledge integration removes this restriction by merging knowledge from multiple teacher with various label sets to train a versatile student model.

Recently, the idea of integrating knowledge from models with different skills has been explored in computer vision (Shen et al., 2019; Ye et al., 2019; Luo et al., 2019; Vongkulbhisal et al., 2019) and graph neural networks (Jing et al., 2021), or extended to a semi-supervised setting (Thadajarassiri et al., 2021). To the best of our knowledge, we are the first to explore knowledge integration for PLMs, which is of great practical value as there are abundant released PLMs. Besides, different from previous methods relying on supervision from feature alignments (Shen et al., 2019; Luo et al., 2019) or independent logits matching (Vongkulbhisal et al., 2019), our MUKI framework operates on the distribution level by utilizing model uncertainty to approximate the golden supervision. Therefore, MUKI is more effective and generalizable for integrating knowledge from heterogeneous PLMs.

5 Error Analysis

The MUKI is built on the assumption that the model uncertainty estimation can faithfully reflect the ability of teacher model. We perform an error analysis

to investigate when this assumption will fail, i.e, the estimated model uncertainty misguides the teacher selection. Specifically, we probe the label distribution of instances on which MUKI assigns a higher uncertainty score to the correct teacher on THUCNews. As shown in the left part of Figure 5, the uncertainty-based teacher selection only fails on a small portion of training examples, i.e., 2.8% of total training instances. Interestingly, the labels of these instances are not uniformly distributed, e.g., *Estate* and *Tech.* have higher error rates than other classes. We further plot the label confusion matrix of an oracle model that is fine-tuned with labeled data with all categories, in the right part of Figure 5. We find that there are classes that can even confuse the oracle model, e.g., instances of *Estate* are tending to be classified into *Politics* and *Finance*, indicating that the mis-identification is partially due to the inherent class similarity. These findings suggest that the performance MUKI can be limited when the integrating teacher models whose classes are highly correlated. As a remedy, the proposed instance re-weighting mechanism is an effective ad-hoc strategy, as the average teacher uncertainty margin $v(x)$ is 0.04, which can greatly reduce the negative impact of these instances. Developing better model uncertainty estimation techniques like incorporating more class information into the estimation process is also promising. Besides, the uncertainty estimation can also be influenced by the model capacity. As deeper models tend to produce more confident predictions, the model selection based on uncertainty scores will favor stronger teacher models when there exist particularly weak teacher models. In such a case, applying model calibration techniques like temperature scaling according to the teacher model size (Guo et al., 2017; Desai and Durrett, 2020) before estimating the uncertainty can be beneficial.

6 Conclusion

In this paper, we explore knowledge integration for PLMs to promote better model reuse. We present MUKI, which integrates teacher predictions according to the model uncertainty estimated via Monte-Carlo Dropout, and dynamically adjusts the instance contribution according to the uncertainty margin. Extensive results on benchmark datasets demonstrate that MUKI can substantially outperform strong baselines, and perform well in challenging settings such as merging heterogeneous

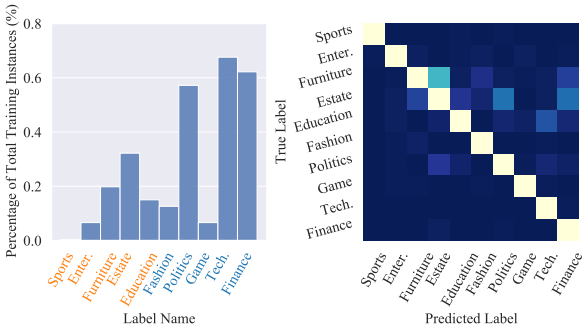

Figure 5: (Left) Label distribution of instances with wrongly selected teachers. Labels in the same color indicate they are in the same teacher specialty. (Right) The confusion matrix of an oracle model.

teachers. Further investigation shows that MUKI can be extended to sequence labeling. In the future, we are interested in developing better integration frameworks for more complex tasks.

7 Ethical Considerations

Our work faces several ethical challenges. As the released PLMs may exhibit potential biases against specific groups, e.g., gender or ethnic minorities (Kurita et al., 2019; Kennedy et al., 2020), these social biases can be propagated to the merged student model. Besides, users may collect unlabeled data from the web for conducting knowledge integration, which possibly contains offensive content and thus introduces new biases into the merged student model as well. We offer possible remedies to reduce the concerns. For the biases exhibited in the teacher PLMs, de-biasing techniques (Zmigrod et al., 2019; Liang et al., 2020; Schick et al., 2021) can be applied to eliminate the potential biases in the teachers before integration. For the offensive unlabeled data collected from the internet, simple template-based or human-in-the-loop data cleaning strategies can be adopted, to identify and filter potential biased data. Except for these techniques, developing a bias-aware knowledge transfer framework that can de-bias the supervision for the student model while maintaining task performance is also promising (Gupta et al., 2022).

Acknowledgements

We thank all the anonymous reviewers for their constructive comments, and Shuhuai Ren for his valuable suggestions in preparing the manuscript. This work was supported by a Tencent Research Grant. Xu Sun is the corresponding author.

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1613–1622. JMLR.org.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

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.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Umang Gupta, J. Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and A. G. Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. *ArXiv*, abs/2203.12574.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. 2021. Amalgamating knowledge from heterogeneous graph neural networks. In *CVPR*, pages 15709–15718.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin J. Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2021. Mixkd: Towards efficient distillation of large-scale language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Qian Liu, Heyan Huang, Yang Gao, Xiaochi Wei, Yuxin Tian, and Luyang Liu. 2018. Task-oriented word embedding for text classification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Sihui Luo, Xinchao Wang, Gongfan Fang, Yao Hu, Dapeng Tao, and Mingli Song. 2019. Knowledge amalgamation from heterogeneous networks by common feature learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3087–3093. ijcai.org.

Xuan Hieu Phan, Minh Le Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 91–100. ACM.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *ArXiv preprint*, abs/2103.00453.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. 2019. Amalgamating knowledge towards comprehensive classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3068–3075. AAAI Press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Maosong Sun, Jingyang Li, Zhipeng Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. *GitHub Repository*.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. 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 4323–4332, Hong Kong, China. Association for Computational Linguistics.

Jidapa Thadajarassiri, Thomas Hartvigsen, Xiangnan Kong, and Elke A. Rundensteiner. 2021. Semi-supervised knowledge amalgamation for sequence classification. In *AAAI*, pages 9859–9867.

Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, and Marco Visentini Scarzanella. 2019. Unifying heterogeneous classifiers with distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3175–3184. Computer Vision Foundation / IEEE.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. One teacher is enough? pre-trained language model distillation from multiple teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4408–4413, Online. Association for Computational Linguistics.

Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. 2019. Student becoming the master: Knowledge amalgamation for joint scene

parsing, depth estimation, and more. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2829–2838. Computer Vision Foundation / IEEE.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Datasets Details

The label sets of datasets we used in the main paper are first sorted according to the name and will be evenly divided into subsets according to the number of teacher models. Table 7 gives the dataset statistics and the class number for two teacher models experiments. Table 8 gives the label (or entity types for NER datasets) list on each dataset. For teacher number that cannot evenly dividing the label sets, the final label set will include the left labels. For example, when there are 4 teacher models are needed on THUCNews dataset, the label set will be split into {*Sports*, *Enter.*}, {*Furniture*, *Estate*}, {*Education*, *Fashion*} and {*Politics*, *Game*,*Tech.*, *Finance*}.

Dataset	#Class (Ent.)	#Train	#Test	$\{\|Y\|\}$
AG News	4	120k	7.6k	{2, 2}
5Abstracts Group	5	53k	1k	{2, 3}
Google Snippets	8	10k	2.2k	{4, 4}
THUCNews	10	50k	10k	{5, 5}
CoNLL 2003	4	14k	3.5k	{2, 2}
Ontonotes v5	18	115.8k	12.2k	{9, 9}

Table 7: Statistics of datasets used in our paper. Ent. denotes the entity types and $\{|Y|\}$ is the number of classes each teacher model specializes.

B Hyper-parameter Search for τ

We perform a hyper-parameter search experiment for the optimal τ in MUKI-soft. We conduct experiments on AG News and THUCNews for stable results. The values of τ are picked from $\{0.01, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$, and the results are shown in Figure 6. We observe that the accuracy drops significantly when τ is set to high values, where the teacher weights distribution is becoming a uniform distribution, while it reaches a peak when τ is set to a small value between 0.2 and 0.5. It indicates that slightly sharpening the teacher weights distribution is helpful for KI. Therefore, we adopt $\tau = 0.2$ in all the experiments.

Dataset	Label Order
AG News	World, Sports, Business, Sci/Tech
THUCNews	Sports, Entertainment, Furniture, Estate, Education, Fashion, Politics, Game, Technology, Finance
Google Snippets	Business, Computers, Culture-Arts-Entertainment, Education-Science, Engineering, Health, Politics-Society, Sports
5Abstracts Group	Business, CSAI, Law, Sociology, Trans
CoNLL 2003	PER, MISC, ORG, LOC
OntoNotes 5.0	LAW, TIME, DATE, LOC, ORG, GPE, NORP, LANGUAGE, EVENT, CARDINAL, MONEY, PRODUCT, PERCENT, FAC, PERSON, QUANTITY, WORK OF ART, ORDINAL

Table 8: Sorted label names (entity types) of datasets. Label names of THUCNews are translated into English.

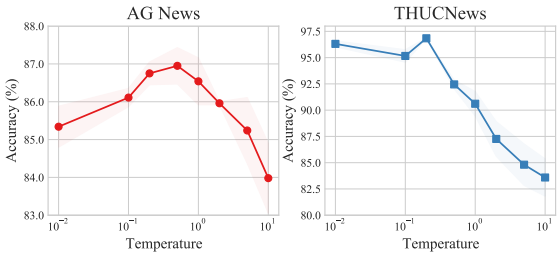

Figure 6: Varying temperature τ for MUKI-Soft. The average accuracy of three seeds are plotted with standard deviation in shade.