LEARNING INVARIANT REPRESENTATIONS ON MULTILINGUAL LANGUAGE MODELS FOR UNSUPERVISED CROSS-LINGUAL TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in neural modeling have produced deep multilingual language models capable of extracting cross-lingual knowledge from unparallel texts, evidenced by their decent zero-shot transfer performance. While studies have attributed this success to cross-lingually shared representations, quantitative analyses are sparse. Towards a better understanding of the role of multilingual representations, in this work, we first make the following observations through empirical analysis: (1) invariance of the feature representations strongly correlates with transfer performance, and (2) distributional shift in class priors between data in the source and target languages negatively affects performance—an issue that is largely overlooked in prior work. Based on our findings, we propose an unsupervised cross-lingual learning method, called importance-weighted domain alignment (IWDA), that performs representation alignment, prior shift estimation, and correction. Experiment results demonstrate its superiority under large prior shifts. In addition, our method delivers further performance gains when combined with existing semi-supervised learning techniques.

1 INTRODUCTION

Many recent state-of-the-art results in natural language processing (NLP) have been achieved on Transformer-based deep neural language models (LMs) using the "pre-train then fine-tune" paradigm (Devlin et al., 2019; Conneau & Lample, 2019). Multilingual versions of these LMs are pre-trained on unannotated and unparallel texts from more than one language, such as multilingual BERT (mBERT) and XLM-R (Conneau et al., 2020a), and it is found that the learned representations are shared across languages. In particular, the cross-lingual knowledge acquired by these models has enabled *zero-shot* cross-lingual transfer, where after fine-tuned on data in a source language, not only do they work well when evaluated on the source, but also quite decently for the same task on almost all languages they have seen during pre-training (Wu & Dredze, 2019).

Their success on zero-shot cross-lingual transfer has prompted numerous studies on their multilingual abilities. Among the many explanations, a recent hypothesis suggests that the deep representation combined with parameter sharing induced intermediate feature representations that are shared across languages (K et al., 2019; Conneau et al., 2020b; Muller et al., 2021). Following this, we begin by studying the following question for mutilingual neural LMs: *What is the role of shared representations in unsupervised cross-lingual learning?*

In an effort to answer this question in a quantitative manner, through empirical analysis on mBERT, we find that (1) its cross-lingual transfer performance strongly correlates with the alignment of feature representations for the NLP task at hand between the source and target languages (Section 2.1). In addition, (2) distributional shift in class priors between data in the source and target languages negatively affects transfer performance (sections 2.2 and 4.2). Despite being a common issue with real-world data under unsupervised settings, it has been largely overlooked in prior work.

Based on these findings on the cross-lingual transfer of mutilingual neural LMs, towards a principled approach for the unsupervised setting where access to *unlabeled* data in the target language for the downstream task at hand is available in addition to labeled ones in the source (hereafter referred

to as UCL[1]), we propose importance-weighted domain alignment (IWDA). This method performs representation alignment, class prior shift estimation, and correction (Section 3). Experiment results show that IWDA outperforms the more popular semi-supervised learning methods under large prior shifts, and can be additionally combined with them for further performance gains (Section 4).

1.1 Related Work

Our proposed method is heavily inspired by Tachet des Combes et al. (2020), which studied the issues with alignment-based domain adaptation (DA) under the presence of class prior shifts, and proposed the inclusion of prior shifts estimation and correction procedures in DA. Our method also shares similar components with the DA-based UCL method by Keung et al. (2019), while highlighting its limitations discovered in the present work (Section 2.2).

Analyzing Multilingual Neural LMs. Most studies analyze mBERT's cross-lingual ability via probing experiments, and identified influencing factors including but not limited to linguistic similarity, size of pretraining corpus, domain similarity, parameter sharing and model depth (Pires et al., 2019; K et al., 2019; Lauscher et al., 2020; Conneau et al., 2020b; Dufter & Schütze, 2020). Another line of research analyze intermediate representations and observed that mBERT learns cross-lingually shared representations (Wu & Dredze, 2019; Conneau et al., 2020b; Muller et al., 2021).

Cross-Lingual Representation Alignment. The idea of learning invariant representations has powered a long history of approaches for cross-lingual transfer on neural models. Under zero-shot settings, it is achieved by aligning word embeddings on bilingual dictionaries that could be induced via adversarial training (Mikolov et al., 2013; Smith et al., 2017; Zhang et al., 2017; Artetxe & Schwenk, 2019; Cao et al., 2019), a method that has powered many machine translation techniques (Artetxe et al., 2018; Lample et al., 2018; Arivazhagan et al., 2019; Pham et al., 2019; Hu et al., 2021) It is found on static embedding models that the goodness of alignment affects transfer performance (Patra et al., 2019; Dubossarsky et al., 2020). On deep models such as mBERT, it could also take the form of shared feature representations in the Transformer body (Artetxe et al., 2020b; Pfeiffer et al., 2020). Under UCL and supervised settings, techniques based on representation alignment were proposed for models ranging from distributed word embeddings (Joty et al., 2017; Chen et al., 2018), RNN (Kim et al., 2017; Huang et al., 2019), to mBERT (Keung et al., 2019; Zheng & Lapata, 2019; Xia et al., 2021). For a survey, see Ruder et al. (2019) and Ruder (2019).

Connection to Domain Adaptation. Domain adaptation is a popular framework for handling domain shifts between training and test data, e.g. learning a POS tagger on news articles but to be deployed on biomedical papers with a significantly different word distribution (Ben-David et al., 2007). The concept of domain is extended naturally to cross-lingual learning by treating the distributions of sentence and document pertaining to the same task in each language as separate domains[2] (Ruder, 2019). Thereby, adaptation techniques proposed for domain shift can be applied to UCL. Besides methods aiming at extracting shared representations (Li et al., 2020; Vernikos et al., 2020), another family of successful approaches is semi-supervised learning, which include self-training (Dong & de Melo, 2019), tri-training (Ruder & Plank, 2018), knowledge distillation (Wu et al., 2020), and data augmentation (Wang et al., 2018; Maharana & Bansal, 2020; Bari et al., 2021).

2 Factors Affecting Cross-Lingual Transfer

To develop and provide evidence to the hypothesis that attributes the cross-lingual transfer success of multilingual neural LMs to shared representations, as well as to understand its implications on the design of unsupervised transfer algorithms, we attempt to quantify the relation between representation invariance and transfer performance in this section.

We briefly summarize the "pre-train then fine-tune" training paradigm and introduce notation along the way. Let $g : \mathcal{X} \to \mathcal{Z}$ denote the feature mapping provided by the multilingual LM, and $z =$

[1]With an abuse of notation. In this work, the acronym UCL refers to the learning setting considered, and is meant to distinguish from the umbrella term "unsupervised cross-lingual learning", which also includes the study of the zero-shot setting where no target language task data is given.

[2]Hence the terms *domain* and *language* are used interchangeably in this work.

$g(x) \in \mathcal{Z} \subseteq \mathbb{R}^d$ the feature vector computed on an input sentence or token $x \in \mathcal{X}$. In practice as well as in our studies and experiments, the feature is the last-layer wordpiece embeddings for token classification tasks, and the embedding of a special start-of-sentence marker (symbolized as `[CLS]`) for sequence classification tasks. Hence, on BERT-Base models, $d = 768$ and $\mathcal{Z} \approx \mathbb{S}^{d-1}$ sphere due to LayerNorm. Let $h : \mathcal{Z} \to \mathcal{Y}$ denote the output layer(s) added on top of g with an appropriate activation, usually a single linear layer with Softmax for classification tasks in practice. Hence, an end-to-end model is obtained, which computes predictions via $\hat{y} = h(g(x))$. By fine-tuning both g, h on labeled task data (x, y) sampled from the source language domain p_S only, we get a zero-shot transfer model for the downstream task at hand, where the term zero-shot points to the fact that the model does not use any task data from the target language $(x', y') \sim p_T$ during training.

Based on the following decomposition of the joint feature-label distribution into the product of class-conditional feature distribution and marginal prior distribution,

$$p(z, y) = p(z|y)p(y) := p(\{x \in \mathcal{X} : g(x) = z\}|y)p(y),$$

we examine how cross-lingual transfer performance is influenced by (1) invariance of the feature representation of data between the source and target languages, with perfect invariance being $p_S(z|Y = j) = p_T(z|Y = j)$ for all $j \in \mathcal{Y}$, and (2) the distributional shift in the class priors, with no shift being $p_S(Y = j) = p_T(Y = j)$ for all $j \in \mathcal{Y}$. On classification tasks, $\mathcal{Y} = \{1, 2, \cdots, k\}$ is discrete. For legibility in later sections, we use the shorthands to denote the class-conditional feature distribution and the marginal prior distribution: $p^{Z|Y}$ and p^Y.

Setup. Our empirical analysis is performed on the representations learned by mBERT zero-shot fine-tuned on English for two multilingual tasks: sentiment analysis on the Multilingual Amazon Reviews Corpus (MARC), and named-entity recognition (NER) on the WikiANN dataset (see descriptions in Section 4). To simulate the effects of varying amount of representation misalignment and class prior difference, evaluation is performed on 500 smaller datasets randomly subsampled from the original MARC according to different class priors, and roughly 700 datasets from WikiANN. MARC covers five high-resource languages (excluding English), and (our selected subset of) WikiANN covers 39 languages of distinct linguistic properties and varying resourcefulness.

2.1 Invariance of Feature Representations

For studying the effects of representation invariance, we compare model performance to the discrepancy between the class-conditional feature distributions on the source and target domains, $p_S^{Z|Y}$ and $p_T^{Z|Y}$. We take the average to aggregate the discrepancies computed when conditioned on each class, $\frac{1}{k}\sum_{j=1}^{k} D(p_S^{Z|Y=j}, p_T^{Z|Y=j})$, and refer to this quantity by *conditional feature shift*.

The discrepancy measure we use computes the ℓ_1-distance between the mean feature vectors[3],

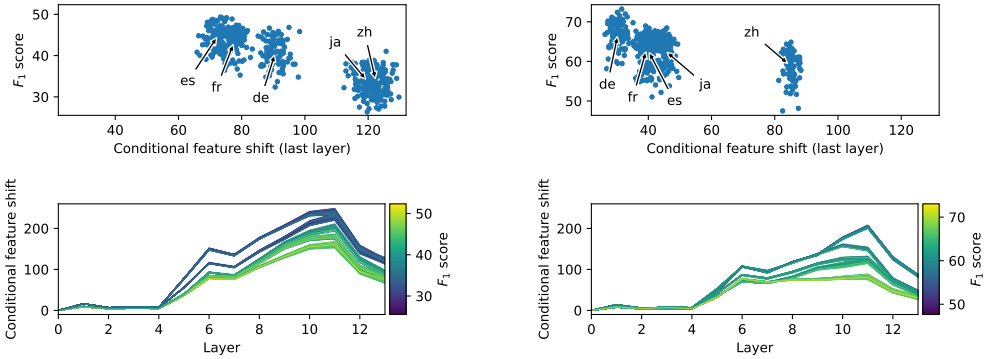
$$D(p, q) := \left\| \mathbb{E}_{x\sim p}[x] - \mathbb{E}_{x'\sim q}[x'] \right\|_1 = \sum_{i=1}^{d} \left| \mathbb{E}_{x\sim p}[x_i] - \mathbb{E}_{x'\sim q}[x'_i] \right|.$$

Measuring the distance between the means of representations for the purpose of analyzing mBERT has appeared in prior work. A similar quantity is referred to as *language centroids* by Libovický et al. (2020). The important distinction here is that we explicitly remove the effects of class prior shift by conditioning on the class labels, which we argue is an artifact of the dataset since arbitrary amounts of *unconditioned* feature shift can be generated by altering the class priors.

First, we consider zero-shot transfer. In Fig. 1a, we plot the F_1 score[4] attained by the zero-shot model against conditional feature shift between English features and those of the target language, and observe that better representation alignment is indicative of better transfer performance. Findings in prior work on the cross-lingual ability of mBERT are reflected. For instance, linguistic similarity, a factor commonly argued to be important for cross-lingual transfer, could be related to the alignment

[3]Comparisons to more sophisticated metrics such as RBF-kernel MMD and linear CKA were made in our preliminary studies (Gretton et al., 2012; Kornblith et al., 2019), and we found that they correlate tightly with our first-moment discrepancy measure for analyzing conditional feature shift.

[4]To study the effects of class prior shift, we use macro-averaged F_1 as the performance metric in this section, because classification accuracy could be inflated on datasets with skewed priors (Azizzadenesheli et al., 2018).



(a) Zero-shot (trained on English only)

(b) Bilingually supervised (English and target)

Figure 1: Cross-lingual transfer performance of mBERT on MARC sentiment analysis under two learning settings, compared to the (class-conditional) (mis)alignment of feature representations between English and target languages (lower shift means more aligned). **Top panels:** Performance (y-axis; higher the better) against alignment of last-layer features (x-axis). Average F_1 scores and feature shifts are marked by language. **Bottom panels:** Performance (now indicated by color; lighter means better) against alignment of features (y-axis) on all Transformer layers (x-axis). It is observed that performance correlates with representation alignment on all later layers. Note that the near-perfect alignment in lower layers is an artifact of using `[CLS]` embeddings as features, which are not informative until contextualized in later layers; see Fig. 6a for results on word embeddings.

of features: Spanish, French and German are more linguistically similar to English than Japanese and Chinese, whose features are less aligned with those of English. These results also reflects the successes of zero-shot learning approaches on mBERT that aim to achieve representation invariance.

While the above finding shows that representation alignment is desirable for zero-shot transfer, is it also suitable for UCL settings where task data from the target language are available? We start by recognizing that the objective function of most unsupervised transfer algorithms consists of two parts: loss on labeled source data, and terms involving unlabeled data that in theory are linked to the loss on target domain. Based on this, we compared against conditional feature shift the performance of the model fine-tuned supervisedly on labeled data from source and target languages, as it would be the ideal objective for unsupervised algorithms. The results are plotted in Fig. 1b. Instead of observing increases in conditional feature shifts, which would have implied that language-specific features are preferred by over shared ones for good performance on each language, we observe much-improved representation alignments between English and each of the target languages. This means that achieving representation invariance is also appropriate under UCL settings.

To demonstrate that the correlation between last-layer class-conditional alignment and performance is not an artifact of the linear classification head, we show that model performance is also correlated with (1) conditional feature shift in intermediate layers, plotted on the bottom panels of Fig. 1, and (2) *class-balanced* feature shift, plotted in Fig. 5 in the appendix, where the removal of prior shift effects is achieved by reweighting the class priors to be uniform (eq. 1). Furthermore, we include results on the XNLI dataset, a higher-level semantic task (Fig. 7), and results on MARC with the more powerful XLM-R Large model (Fig. 8).

2.2 Distributional Shift in Class Priors

Most benchmarks for evaluating cross-lingual transfer are constructed so that the class prior distributions between data in the source and target languages are the same. This is desired for assessing zero-shot learning approaches, as the focus should not be on "'tricks' for how to best handle the class imbalance" (Schwenk & Li, 2018). However, for UCL settings with unlabeled target data, we argue that making the best use of the additional data involves the detection and correction of prior shift. Otherwise, as we show in this and later sections, negative transfer could occur.

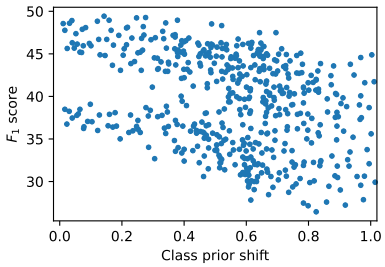


Figure 2: Distributional shifts in class priors between English and target datasets vs. zero-shot performance on MARC. Performance negatively correlates with prior shift.

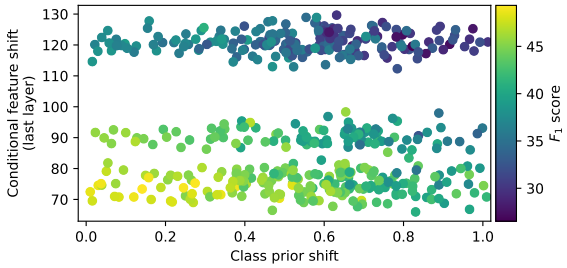


Figure 3: Conditional feature representation alignment and class prior shift vs. zero-shot performance (indicated by scatter color; lighter means better) on MARC. Stronger correlations of each factor with performance than those in figs. 1 and 2 are revealed.

To examine the effects of class prior shift on cross-lingual transfer, we compare zero-shot transfer performance to the difference in prior distributions, $D(p_S^Y, p_T^Y)$, equivalent to total variation (TV). The results are presented in Fig. 2. We also plot the effects of representation alignment and prior shift jointly in Fig. 3, where stronger correlations are revealed: (1) transfer performance generally takes a hit when prior shifts are large, regardless of how well the features are aligned, and (2) on target domains of the same amount of prior shift, performance decreases when features are less aligned. The findings suggest that in addition to achieving representation invariance, good transfer performance also requires the detection and handling of potential prior shifts. While this is not possible under zero-shot settings without any data from the target domain, the unlabeled target task data provided under UCL settings could be leveraged toward this goal.

We examined existing approaches proposed for multilingual neural LMs for the UCL settings, but did not find procedures for prior shift correction, and moreover, there is a lack of attention to the potential issues and effects of class prior shift to the algorithms. Specifically, for alignment-based methods, Zhao et al. (2019) and Tachet des Combes et al. (2020) showed that adaptation could lead to negative transfer when prior shifts are not accounted for due to erroneous alignment. For methods based on semi-supervised learning, negative transfer is observed empirically under large shifts in our studies (Section 4.2). These cases of negative transfer call for more considerations to the effects of class prior shift in future work on cross-lingual transfer that leverage target data.

Interestingly, although the alignment-based method proposed by Keung et al. (2019) did not account for prior shift, it avoided negative transfer on the CoNLL NER task (which exhibits shifts) by performing feature alignment on the `[CLS]` sentence embeddings instead of the word embeddings (recall that predictions are made on word embeddings on token classification tasks). Since sentence embeddings on Transformer-based models could be loosely related to a weighted sum of the word embeddings, the procedure of Keung et al. (2019) is performing coarse-grained representation alignment, which avoids negative transfer while providing decent performance gain. However, as we compare in Table 1, this misses out opportunities for larger gains from word embedding alignment. If alignment were performed on word embeddings without prior shift correction, deteriorating performance would have been observed (Fig. 11).

3 Importance-Weighted Domain Alignment

For unsupervised cross-lingual learning on multilingual neural LMs with access to *unlabeled* target language data (which we refer to as the UCL setting), the findings in Section 2 suggest that good performance can be achieved via aligning representations and correcting class prior shifts. Based on these principles, we propose class-importance-weighted domain alignment (IWDA) for UCL, following closely the work by Tachet des Combes et al. (2020). In this section, we provide a high-level description of the algorithm and its connection to the findings above, and refer interested readers to the aforementioned work and Appendix C for further implementation details.

We have seen that class-conditional feature alignment is desired for cross-lingual transfer, however, this is not possible to achieve under the UCL setting since the target labels are unknown. Assume for now that the target class prior distribution is known, then we can compute the true ratio of the class priors, $w_j^* := p_T(Y=j)/p_S(Y=j)$, also known as the importance weights (IWs), and aim for a weaker *importance-weighted* feature alignment, $p_S^{w^*,Z} = p_T^Z$, where

$$p_S^{w^*}(x, Y=j) = p_S(x|Y=j)p_S(Y=j)w_j^* = p_S(x|Y=j)p_T(Y=j). \tag{1}$$

Besides as shown in Fig. 5 that importance-weighted feature alignment also correlates with transfer performance, it is also a necessary condition for the ideal goal of class-conditional alignment (Tachet des Combes et al., 2020).

Since the true w^* are unknown under UCL, they are estimated by matching first-moment statistics using the confusion matrix (Saerens et al., 2002; Lipton et al., 2018). Assume now that the features are conditionally aligned, $p_S^{Z|Y=j} = p_T^{Z|Y=j}$ for all $j \in \mathcal{Y}$ (known as the *label shift* assumption), and the label space is discrete, $\mathcal{Y} = \{1, \cdots, k\}$. Then for each class $i \in [k]$, we have the following relation between *predicted* target prior distribution, the source confusion matrix, and w^*:

$$p_T(\hat{Y}=i) = \sum_{j=1}^{k} p_T(\hat{Y}=i|Y=j)p_T(Y=j) = \sum_{j=1}^{k} p_S(\hat{Y}=i|Y=j)p_T(Y=j) = \sum_{j=1}^{k} p_S(\hat{Y}=i, Y=j)w_j^*, \tag{2}$$

where $\hat{Y} := h(g(X))$ are label predictions. Since we have the same number of equations as unknowns, w^* could be recovered under mild assumptions (Lipton et al., 2018).

Careful readers may have noticed that achieving either goal above requires having the other as an assumption. Indeed, attempting to estimate IW without conditional alignment could fail, as well as alignment without accurate IWs. The hope is that the optimization will be predominantly guided by the implicit bias of the pre-trained multilingual LM for selecting largely invariant features (hence conditionally aligned), so that small improvements to feature alignment and updates to prior shift estimates could bring incremental improvements to both components.

Optimization. The two goals above define an optimization problem over g, h and w,

$$\min_{g,h} \Big(\mathbb{E}_{(x,y)\sim p_S}[\ell(h(g(x)), y)] + \lambda D(p_S^{w,Z}, p_T^Z) \Big), \quad \min_{w} \sum_{\hat{y}\in[k]} \Big(p_T(\hat{y}) - \sum_{y} p_S(\hat{y}, y) w_y \Big)^2,$$

where alignment objective minimizes source task loss and the discrepancy of w-importance-weighted feature distributions (the metric should be more fine-grained than the first-moment metric used in our analysis), and the IW estimation objective solves eq. 2 using current statistics based on h, g with the constraints $w \geq 0$ and $w^\top p_S^Y = 1$.

In our implementation, we use the minimax formulation of the alignment objective optimized with adversarial training, where D is replaced by a shallow ReLU network Wasserstein-1 critic and zero-centered gradient penalty (Arjovsky et al., 2017; Thanh-Tung et al., 2019). We found the Wasserstein-1 adversarial loss to be more responsive to importance weighting than cross-entropy loss (Ganin et al., 2016).

Improving Stability of IW Estimates. In Tachet des Combes et al. (2020), the IWs estimation objective is solved by quadratic programming (QP), and updated once every epoch of training in their implementation. Since accurate IWs estimates are integral to successful adaptation, we (1) update w using projected gradient descent at every training step for maintaining their optimality. This, however, led to the discovery of previously unexplored practical issues with the IW estimator. In our preliminary experiments, we observed that w tends overestimate the majority class on the target domain, and diverges during longer training episodes.

These issues are unique to unsupervised adaptation settings, and are due to inefficiencies of the critic and noises introduced by the alignment objective. They could also be reproduced with the previous QP procedure. We address these issues by (2) ℓ_2-regularizing w (Azizzadenesheli et al., 2018), and (3) using a decaying learning rate schedule for w. These modifications are necessary for achieving best performance, as long training episodes are required for fine-tuning on many NLP tasks, and also needed to maintain the critic near optimality for guiding properly the alignment.

Table 1: CoNLL NER (macro-averaged F_1).

	en	de	es	nl
Supervised results (in-language)				
mBERT	91.97	82.82	87.38	90.94
XLM-R	92.92	85.81	89.72	92.53
Zero-shot baselines				
mBERT	90.57	69.77	74.14	78.28
XLM-R	92.49	72.37	77.06	82.03
Unsupervised learning				
Keung et al. (2019)	-	71.9	74.3	77.6
Wu et al. (2020)	-	73.22	76.94	**80.89**
mBERT IWDA	90.77	72.56	76.11	78.63
+ KD	90.89	**73.71**	**77.14**	79.72
mBERT IWDA (oracle)	90.75	72.58	76.48	79.17
Bari et al. (2021)	-	80.99	83.24	85.32
XLM-R IWDA	92.55	75.42	79.08	81.79
XLM-R IWDA (oracle)	92.42	75.32	78.94	81.82

Table 2: WikiANN NER (macro-averaged F_1).

	en	bn	my	ur
Supervised results (in-language)				
mBERT	83.73	92.69	69.08	93.29
XLM-R (Hu et al., 2020)	84.7	97.8	76.8	97.1
Zero-shot baselines				
mBERT	83.73	68.55	50.16	35.78
XLM-R	83.55	76.24	58.12	62.66
Unsupervised learning				
Wu et al. (2020)	-	71.78	46.32	44.52
mBERT IWDA	83.93	71.05	55.86	75.79
+ KD	83.07	**73.15**	56.05	**77.58**
mBERT IWDA (oracle)	83.93	72.19	**60.22**	74.98
Bari et al. (2021)	-	82.68	70.61	84.99
XLM-R IWDA	83.43	78.03	57.10	78.89
XLM-R IWDA (oracle)	83.35	78.81	57.37	79.21

4 EXPERIMENTS

In our experiments, we compare importance-weighted domain alignment (IWDA) to two baseline methods for cross-lingual transfer on three NLP tasks, under the unsupervised setting where access to *unlabeled* task data from the target language is provided (UCL setting). Besides supporting our findings and our motivated algorithm, these experiments are aimed at investigating the limitations of existing approaches under distributional shifts in prior distributions between source and target. We also illustrate the potential benefits of IWDA in light of stronger cross-lingual learning methods.

Datasets. The tasks considered are named-entity recognition (NER), sentiment analysis, and textual entailment. **NER:** Two benchmark multilingual datasets are used, CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003) and WikiANN (Pan et al., 2017; Rahimi et al., 2019). The former includes English, German, Spanish and Dutch, and the latter covers over 200 languages, from which we select a subset of 40 according to the XTREME benchmark (Hu et al., 2020). **Sentiment analysis:** Multilingual Amazon Reviews Corpus (MARC) is used (Keung et al., 2020), a *balanced* dataset with product reviews in English, German, Spanish, French, Japanese and Chinese. **Textual entailment:** The XNLI dataset is used (Conneau et al., 2018), but due to space constraints, its setup and results are relegated to Appendix B. As in most work on cross-lingual transfer, English is used as the source language for transfer in our evaluations.

Baselines. We perform extensive comparative experiments with two semi-supervised learning (SSL) techniques. Both require having a model fine-tuned on source data only. **Knowledge distillation (KD):** It distills and transfers knowledge of the source teacher model to the target domain by training a newly initialized student network to mimic teacher outputs on target data (Hinton et al., 2015). A variant of KD was proposed for NER by Wu et al. (2020), which also freezes the bottom three layers of mBERT during training (Wu & Dredze, 2019). **Self-training (ST):** It iteratively assigns pseudo-labels to unlabeled target data and train on the most confident ones. Our implementation follows closely the work by Dong & de Melo (2019), which considered sequence classification.[5]

For IWDA experiments, the model is first fine-tuned on source only, then continue fine-tuned with the IWDA objective on both source and unlabeled target data. We also include results for when the true IWs are known and hard-coded as an upper bound on IWDA performance (**IWDA (oracle)**).

Experiments are performed on BERT-Base Multilingual Cased and XLM-R Large models. More details on IWDA and our baselines are included in Appendices C and D along with hyperparameter settings. Unless otherwise noted, results are from our implementation.

[5]These baselines are chosen because they require no more resources than the unlabeled target data provided, are not tailored to any particular model, and can be easily implemented on any existing models, which are qualities shared by IWDA. As many elaborate methods are also based on the same fundamental ideas as these baselines, we hope that they are representative of approaches based on SSL, if not exhaustive.

Table 3: MARC sentiment analysis (accuracy).

	en	de	es	fr	ja	zh
Supervised results (in-language)						
mBERT	58.50	61.19	57.76	57.05	57.97	53.74
XLM-R	62.05	64.66	60.70	59.40	60.40	56.24
Zero-shot baselines						
mBERT	58.50	44.80	46.49	46.02	37.37	38.48
XLM-R	62.05	62.52	59.04	56.43	54.32	51.42
Unsupervised learning						
Wu et al. (2020)	-	47.43	49.56	49.06	35.77	38.54
mBERT ST	58.25	50.74	48.80	47.96	42.10	41.40
mBERT IWDA	56.87	51.94	49.77	49.78	42.62	44.04
+ ST	56.41	**53.11**	**51.00**	49.91	43.59	45.27
mBERT IWDA (oracle)	55.96	51.95	50.83	**50.01**	**44.91**	**45.96**
XLM-R IWDA	60.20	61.34	56.64	55.58	55.00	50.36
XLM-R IWDA (oracle)	60.10	61.50	57.16	56.10	56.36	52.52

4.1 Results on Standard Benchmarks

To begin, through evaluations on standard benchmarks, we demonstrate that IWDA brings performance gains via improving representations to be better-aligned across languages. Results on CoNLL and low-resource WikiANN NER are presented in Tables 1 and 2, and those on MARC in Table 3. Remaining results on WikiANN with mBERT are included in Table 4 in the appendix.

In all mBERT IWDA experiments, improvements upon zero-shot baseline are observed, and are accompanied by improved representation alignment. The average decrease in conditional feature shift is 21.60% and 22.33% on CoNLL with IWDA and IWDA (oracle) respectively (with 100% indicating perfect alignment), 51.53% and 64.77% on MARC, and 32.91% and 36.34% on WikiANN.

However, improvements sometimes fall short of SSL baselines, which we suspect is due to the noise introduced during the alignment process. Potential sources include fluctuations in IW estimates and insufficiently trained critics. By training IWDA with more passes on the data and improving the optimality of the critic, further improvements on CoNLL NER are observed (Table 6).

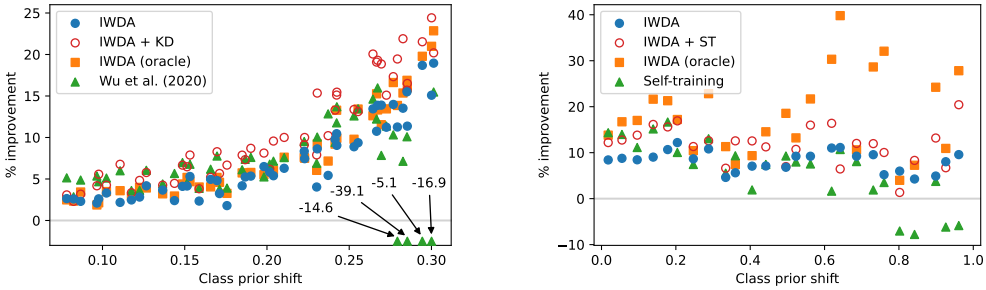
Another way to reduce noise is to perform a final-stage SSL training on the IWDA-aligned model. These experiments are labeled **IWDA + KD** and **IWDA + ST**, on which we observe some best results with mBERT on both tasks. This is an appealing strategy, because SSL algorithms theoretically depend on the assumption that the source and target domains are the same, IWDA training can be viewed as a preprocessing step on the feature representations to satisfy this assumption. We will revisit this point in the next section where we consider cross-lingual transfer under prior shifts.

On XLM-R models, IWDA also brings improvements on NER. However on MARC, gains are only observed for languages distant from English. Here, we remark that the performance gain from IWDA (if any) will come from enhancements to the alignment of last-layer features in their current configurations, but not from learning entirely different representations with fundamentally better class-conditional alignment. This cannot be achieved absent target labels, additional supervision or resources. If the representations learned by the powerful XLM-R model during zero-shot training are already efficient, then the potentials of IWDA will be limited. Opportunities for further improvement in this case could come from additional resources, e.g. Bari et al. (2021) perform unsupervised data augmentation by leveraging the multilingual mask-filling ability of XLM-R.

4.2 Results Under Class Prior Shifts

In this section, we evaluate on datasets with large prior shifts, a likely property of real-world data that is insufficiently investigated for UCL, and contrast the brittleness of SSL baselines with the robustness of IWDA. Because of the lack of multilingual benchmarks exhibiting prior shifts, similar to Section 2.2 , we evaluate on smaller datasets subsampled from the German portion of CoNLL and the Japanese portion of MARC according to varying amounts of shifts.

Results plotting percent improvements over zero-shot baselines against the amount of prior shift are presented in Fig. 4. Although SSL baselines are generally better than IWDA under mild to no shifts,



(a) CoNLL English to German

(b) MARC English to Japanese

Figure 4: Percent improvement of IWDA and SSL over zero-shot baselines (y-axis; in macro-averaged F_1) under varying class prior shifts (x-axis). Performance of SSL deteriorates as prior shift increases, because it does not account for prior shift.

their performances quickly deteriorate as prior shift increases, until eventually resulting in negative transfer. IWDA on the other hand does not cause negative transfer from prior shifts owing to its estimation and correction capabilities. The relative success of IWDA demonstrates that it is possible to estimate prior shift from unlabeled target data. We include XLM-R MARC results in Fig. 10.

In mBERT experiments, the average percent correction of prior shift, computed from the decrease in TV between true target domain priors and end-of-training estimates, normalized by the amounts of shift, is 87.33% on CoNLL (with 100% indicating perfect IW estimation), and 28.08% on MARC under shifts larger than 0.8 in TV. On MARC with smaller shifts, we observe estimates that are worse than assuming a uniform prior distribution (-335.80% in correction). This is because these are harder problems with noisy labels (as it is naturally difficult to distinguish a 2-star review from a 3-star one), the suboptimality of the source model from which IWDA is trained means that the initial representations may be conditionally misaligned, thereby affecting the accuracy of the initial as well as subsequent IW estimates.

Since IWDA brings improvements to class-conditional alignment, as already pointed out in the previous section, it is appealing to complement SSL and potentially more elaborated methods with IWDA to improve robustness against prior shift. As demonstrated in our IWDA + KD and ST experiments, SSL no longer suffer from negative transfer when preceded by IWDA, and altogether provide further performance gains.

5 Conclusion and Future Work

From the observations on multilingual neural LMs that cross-lingual transfer performance is strongly correlated with representation invariance and negatively affected by class prior shift, we proposed importance-weighted domain alignment (IWDA) for unsupervised cross-lingual learning. Experiment results support our findings, and show that IWDA is effective at the estimation and correction of class prior shift from unlabeled data, and robust to large shifts, under which in contrast existing UCL approaches based on SSL could result in negative transfer. As prior shift is a prevalent issue under unsupervised settings, more care to it is necessary in the design of future UCL algorithms for them to achieve good performance not only on curated benchmarks, but also on real-world data.

While our present implementation of IW estimation with the moment-matching approach is largely effective, future work on the search and study of potentially more stable estimation methods would enable IWDA to achieve better and more consistent transfer performance. Since the strength of IWDA lies in its ability to detect prior shifts and improve on class-conditional representation alignment, with our experiments on combining IWDA and SSL baselines as a preliminary demonstration of its potential usefulness as a feature preprocessing step and in improving robustness to prior shifts, it would be of interest to investigate its benefits on stronger models and methods towards better cross-lingual transfer.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. In *International Conference on Machine Learning*, pp. 222–232. PMLR, November 2020.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. The Missing Ingredient in Zero-Shot Neural Machine Translation. *arXiv:1903.07091 [cs]*, March 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, July 2017.

Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, September 2019. doi: 10.1162/tacl_a_00288.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised Neural Machine Translation. In *International Conference on Learning Representations*, February 2018.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation Artifacts in Cross-lingual Transfer Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7674–7684, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A Call for More Rigor in Unsupervised Cross-lingual Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7375–7388, Online, July 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.658.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized Learning for Domain Adaptation under Label Shifts. In *International Conference on Learning Representations*, September 2018.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. UXLA: A Robust Unsupervised Data Augmentation Framework for Zero-Resource Cross-Lingual NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1978–1992, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.154.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*, September 2019.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018. doi: 10.1162/tacl_a_00039.

Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv:1809.05053 [cs]*, September 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Xin Dong and Gerard de Melo. A Robust Self-Learning Framework for Cross-Lingual Text Classification. In *Q*, pp. 6306–6310, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1658.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The Secret is in the Spectra: Predicting Cross-lingual Task Performance with Spectral Similarity Measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2377–2390, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.186.

Philipp Dufter and Hinrich Schütze. Identifying Elements Essential for BERT's Multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.358.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. ISSN 1533-7928.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A Unified View of Label Shift Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3290–3300, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, July 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, March 2015.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *International Conference on Machine Learning*, pp. 4411–4421. PMLR, November 2020.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit Alignment Objectives for Multilingual Bidirectional Encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3633–3643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.284.

Lifu Huang, Heng Ji, and Jonathan May. Cross-lingual Multi-Level Adversarial Transfer to Enhance Low-Resource Name Tagging. 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)*, pp. 3823–3833, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1383.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. Cross-language Learning with Adversarial Neural Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 226–237, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1024.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*, September 2019.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER. 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)*, pp. 1355–1360, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1138.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4563–4568, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.369.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2832–2838, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1302.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, May 2019.

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

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363.

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. Unsupervised Domain Adaptation of a Pretrained Cross-Lingual Language Model. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 4, pp. 3672–3678, July 2020. doi: 10.24963/ijcai.2020/508.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1663–1674, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.150.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and Correcting for Label Shift with Black Box Predictors. In *International Conference on Machine Learning*, pp. 3122–3130. PMLR, July 2018.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Adyasha Maharana and Mohit Bansal. Adversarial Augmentation Policy Search for Domain and Cross-Lingual Generalization in Reading Comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3723–3738, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.333.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*, September 2013.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2214–2231, Online, April 2021. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 184–193, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. Improving Zero-shot Translation with Language-Independent Constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 13–23, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5202.

Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, NUI Galway, 2019.

Sebastian Ruder and Barbara Plank. Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1044–1054, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1096.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, January 2002. ISSN 0899-7667. doi: 10.1162/089976602753284446.

Holger Schwenk and Xian Li. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, April 2017.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J. Gordon. Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19276–19289, 2020.

Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, and Ion Androutsopoulos. Domain Adversarial Fine-Tuning as an Effective Regularizer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3103–3112, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.278.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. SwitchOut: An Efficient Data Augmentation Algorithm for Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 856–861, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1100.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6505–6514, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.581.

Shijie Wu and Mark Dredze. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. 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)*, pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077.

Shijie Wu and Mark Dredze. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. MetaXL: Meta Representation Transformation for Low-resource Cross-lingual Learning. In *Proceedings of the 2021 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pp. 499–511, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.42.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On Learning Invariant Representations for Domain Adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, May 2019.

Hao Zheng and Mirella Lapata. Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6236–6247, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1628.

A Additional Analyses and Findings from Section 2

In this section, we provide additional results to support our findings by considering more languages (WikiANN NER), a higher-level task (XNLI), and a more powerful model (XLM-R Large).

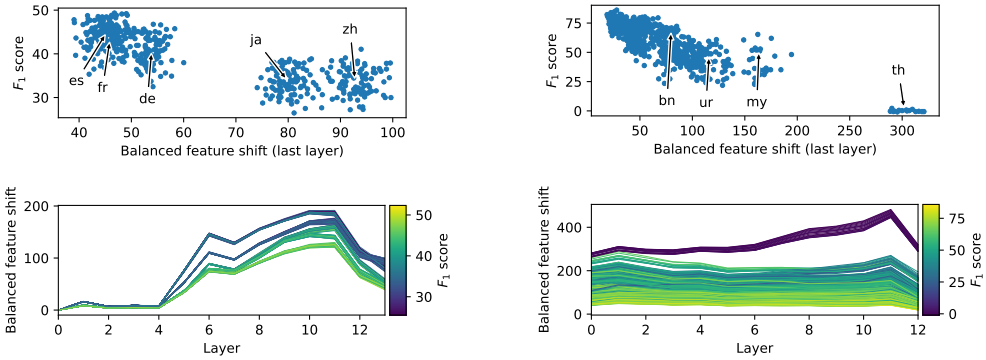
First and foremost, to demonstrate that the correlation between class-conditional representation alignment and performance is not confounded by the clustering structure induced by the linear classifier trained on last-layer features, we plot and observe similar correlation between transfer performance and *class-balanced* representation alignment in Fig. 5.

WikiANN. As an attempt to minimize the effects from potential confounders on the empirical findings in the main sections, we perform the same experiments for studying the relation between transfer performance, representation alignment and prior shift by zero-shot transferring mBERT from English to 39 target languages included in the WikiANN NER dataset. The results are provided in Fig. 6. These languages have distinct linguistic properties and varying degrees of resourcefulness. A summary of them can be found in Table 5 of Hu et al. (2020), and sizes of their pretraining corpora in Table 1 of Wu & Dredze (2020).

It is observed that the same correlation between representation alignment and performance still holds, but the effects of prior shift are mostly mild. We suspect that this is because the task we evaluated on, named-entity recognition, is a structured prediction problem, hence mBERT's ability to leverage dependencies between words made it robust to prior shifts on these tasks. We later observed that if the source dataset on which the model is trained on has a skewed prior distribution, then the effects of prior shift will be pronounced. The mBERT WikiANN results under this setting are shown in Fig. 9b, and more discussions are coming up in Section A.1.

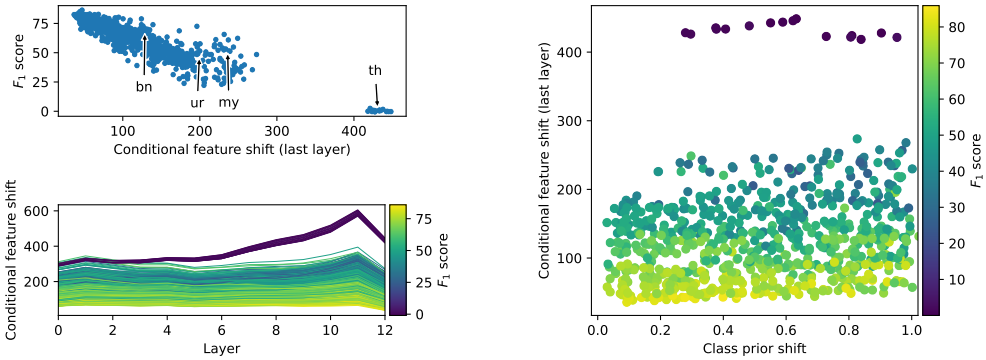
XNLI. So far, we have seen empirical evidence on sentiment analysis and NER tasks, and one may wonder if the same conclusions still hold when considering cross-lingual transfer to high-level semantic tasks. For thoroughness, we include a plot in Fig. 7 comparing zero-shot transfer performance to representation alignment when mBERT is trained on a textual entailment task, from English to 14 languages included in the XNLI dataset. Here, the same observations are made, demonstrating that the findings hold across NLP tasks.

XLM-R. Another variable in our empirical analyses is the model used for evaluation. While the lightness of mBERT afforded us to perform extensive experiments, it would be of practical interest to see whether our findings also hold on larger and more powerful models. In Fig. 8, we compared performance to alignment and prior shift on XLM-R Large model on MARC. Again, the observations support our findings, and show that they also hold across models.



(a) MARC (b) WikiANN

Figure 5: mBERT zero-shot transfer performance on two datasets vs. class-balanced alignment.



(a) Last and intermediate feature alignment (b) Feature alignment and class prior shift

Figure 6: mBERT zero-shot transfer performance on WikiANN NER dataset vs. class-conditional alignment and prior shift.

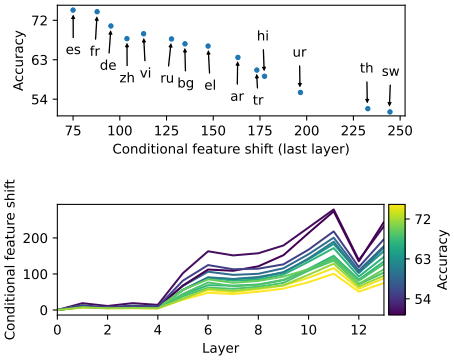
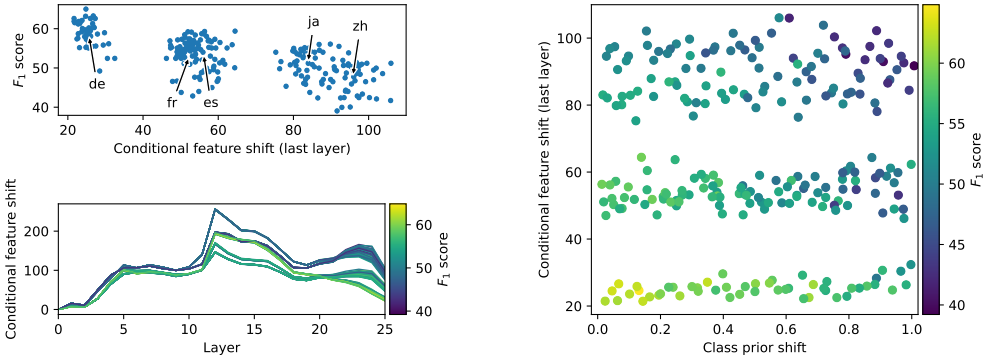


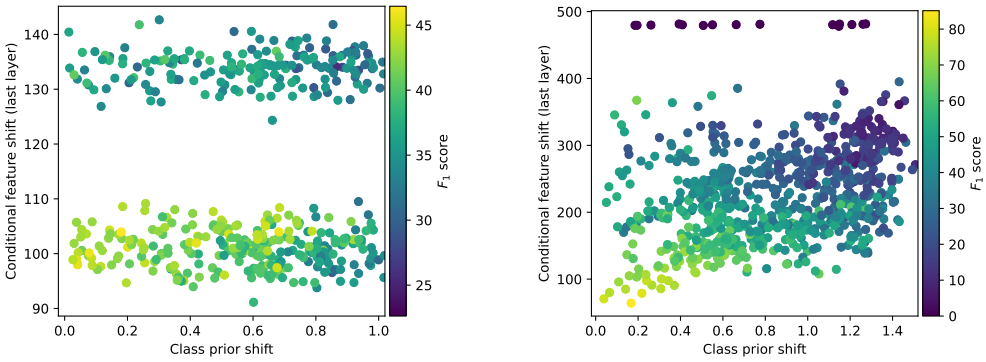
Figure 7: mBERT zero-shot transfer performance on XNLI dataset vs. class-conditional alignment and prior shift.



(a) Last and intermediate feature alignment

(b) Feature alignment and class prior shift

Figure 8: XLM-R zero-shot transfer performance on MARC vs. class-conditional alignment and prior shift.



(a) MARC

(b) WikiANN

Figure 9: mBERT zero-shot transfer performance on two datasets vs. class-conditional alignment. **Model is trained on skewed source training data.** Effects of class prior shifts are pronounced.

A.1 Effects from Skewed Source Prior Distributions

As briefly mentioned above, we have only considered the cross-lingual transfer from relatively balanced source datasets in previous sections. Here, we show that when the model is fine-tuned on source datasets with skewed prior distributions, the effects of class prior shift are more pronounced.

We first experiment on MARC. We generate an skewed English source dataset via subsampling according to the prior distribution of 14.86%, 6.43%, 8.03%, 19.28% and 51.41%[6], corresponding to the proportion of 1- through 5-star reviews. Then, we repeat the previous empirical analysis by evaluating an mBERT fine-tuned on this skewed English dataset, and the results are shown in Fig. 9a. Compared to when the source is class-balanced (Fig. 3), here the correlation between performance and class prior shift is much stronger.

On WikiANN, although the original English portion of the dataset is already considerably skewed (with class O makes up 50.75% of the labels), the correlation between prior shift and model performance is weak (Fig. 6b), likely due to NER being a structured prediction problem where the dependencies between words in the same sentence helped the model to be more robust. Repeating the same experiment above with the skewed English source dataset generated by keeping only sentences containing tokens with an O label (so that now 80.92% of the tokens belong to class O), we again observe pronounced effects of prior shift on performance (results in Fig. 9b).

[6]This is the actual distribution of product ratings on Amazon.com.

Table 4: WikiANN NER full mBERT results (macro-averaged F_1).

	en	af	ar	bg	bn	de	el	es	et	eu
Supervised	83.73	91.36	87.62	92.64	92.69	89.47	90.55	91.42	91.58	91.72
Zero-shot	83.73	77.63	40.68	77.89	68.55	79.01	71.19	76.73	76.74	59.76
IWDA	83.93	77.50	37.59	78.57	71.05	78.86	72.19	83.37	79.16	67.11
IWDA (oracle)	83.93	77.18	37.29	79.65	72.19	79.07	72.13	83.72	80.23	72.69
	fa	fi	fr	he	hi	hu	id	it	ja	jv
Supervised	92.11	91.20	90.15	85.19	86.10	92.67	91.88	91.44	73.18	73.71
Zero-shot	40.27	77.87	80.47	56.55	66.08	76.55	60.67	80.67	27.49	61.51
IWDA	28.32	78.43	79.99	56.57	64.37	76.00	51.35	80.46	38.05	61.48
IWDA (oracle)	46.33	79.56	83.78	57.62	67.01	76.69	72.00	80.99	34.98	60.57
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru
Supervised	86.55	85.06	87.71	82.68	85.42	93.27	69.08	91.37	91.45	88.28
Zero-shot	66.41	46.87	60.52	53.33	57.97	67.63	50.16	82.16	80.07	65.61
IWDA	67.14	48.15	61.18	53.75	61.80	69.78	55.86	81.03	82.40	66.00
IWDA (oracle)	67.39	49.11	63.29	59.47	61.55	69.01	60.22	81.05	83.66	72.00
	sw	ta	te	th	tl	tr	ur	vi	yo	zh
Supervised	88.98	83.53	78.53	76.84	93.64	92.35	93.29	90.83	82.93	81.14
Zero-shot	69.35	50.22	50.63	1.31	69.57	73.78	35.78	71.19	49.29	43.29
IWDA	67.56	53.93	52.09	1.60	75.16	74.36	75.79	80.20	52.07	49.90
IWDA (oracle)	67.40	55.46	53.84	8.39	76.50	78.64	74.98	78.85	54.60	56.26

Table 5: XNLI textual entailment (accuracy; non-benchmark).

	en	ar	bg	de	el	es	fr	hi
Supervised results (English + machine translation)								
mBERT	73.60	65.32	68.58	68.94	66.32	71.18	69.28	63.02
XLM-R	84.36	77.48	79.08	80.42	80.26	81.18	80.16	70.82
Zero-shot baselines								
mBERT	73.60	60.64	60.60	67.12	63.14	67.58	67.06	57.30
XLM-R	84.36	74.72	77.42	78.58	77.68	80.26	79.14	68.12
Unsupervised learning								
mBERT ST	73.04	60.20	63.40	66.30	62.00	67.92	67.16	56.62
mBERT IWDA	73.03	**61.16**	64.16	**67.30**	62.68	68.28	**67.34**	57.92
+ ST	73.10	60.28	64.18	66.66	62.44	68.02	66.96	57.70
mBERT IWDA (oracle)	72.94	60.74	**64.62**	67.02	**63.22**	**68.88**	67.06	**58.88**
XLM-R ST	84.35	74.78	**78.14**	78.10	77.60	79.48	79.26	67.20
XLM-R IWDA	84.61	74.31	77.32	78.36	77.22	80.14	79.54	67.92
+ ST	84.51	**75.62**	77.36	**78.82**	77.46	79.94	79.20	67.94
XLM-R IWDA (oracle)	84.69	74.68	77.98	78.58	**78.22**	79.78	**79.74**	**68.24**
	ru	sw	th	tr	ur	vi	zh	
Supervised results (English + machine translation)								
mBERT	67.66	57.32	63.36	63.02	48.60	68.24	69.62	
XLM-R	78.24	68.14	77.24	74.02	49.68	77.20	77.08	
Zero-shot baselines								
mBERT	64.30	47.28	51.40	57.36	43.16	62.98	63.70	
XLM-R	76.84	64.30	73.82	72.30	45.82	74.92	75.20	
Unsupervised learning								
mBERT ST	64.32	47.76	53.40	57.80	46.04	62.88	63.90	
mBERT IWDA	**65.06**	47.78	**54.80**	58.00	44.24	63.60	65.08	
+ ST	64.56	**48.26**	53.78	57.92	**46.94**	63.40	65.12	
mBERT IWDA (oracle)	64.84	45.96	54.50	**59.36**	46.76	**64.28**	**65.82**	
XLM-R ST	76.90	64.60	73.94	72.00	47.26	73.82	74.54	
XLM-R IWDA	**76.94**	64.20	74.12	71.64	45.70	74.44	74.94	
+ ST	76.84	**65.70**	**74.14**	72.30	**47.12**	74.14	75.86	
XLM-R IWDA (oracle)	76.90	64.56	**74.14**	**72.84**	46.98	**75.20**	**76.10**	

Figure 10: Percent improvement of IWDA and SSL over zero-shot baselines on XLM-R and MARC Japanese under varying class prior shifts.

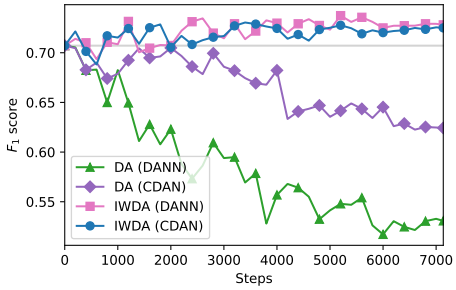


Figure 11: mBERT transfer performance on CoNLL NER from English to German vs. number of optimization steps. We compare domain alignment techniques applied on last-layer word embeddings with importance-weight estimation (IWDA) and without (DA).

B Additional Experiment Results from Section 4

By comparing the transfer performance of IWDA to SSL on standard benchmarks as well as datasets exhibiting prior shifts in the main sections, we demonstrated that better performance can be achieved with better representation alignment under the UCL setting, as well as the limitations with existing approaches under prior shifts. While the prior shifts experiments in Section 4.2 are performed on mBERT, in Fig. 10, we provide XLM-R Large results on MARC Japanese prior shift experiments.

Furthermore, we provide full mBERT WikiANN results of applying IWDA (Table 4), and results on the XNLI dataset, a high-level semantic task of textual entailment.

XNLI. Because XNLI does not provide human-generated training data for languages other than English, to obtain unlabeled target language task data for our studies on the UCL setting, we use the machine-translation of the English training set provided by Conneau et al. (2020a). To remove potential effects due to parallelism, similar to the setup considered in Bari et al. (2021), we partition the dataset into equal portions of size 20,000 (5% of the original size) and randomly select disjoint partitions to use as source and target data. In addition, since the original XNLI test set is human-translated but our target data is machine-translated, to avoid the effects from this discrepancy (or domain shift) between train and test, evaluations are performed on the machine-translated portion of the dataset in our experiments (we refer interested readers to Artetxe et al. (2020a) for a discussion on XNLI translation artifacts).

Due to our setup, our results are non-benchmark and cannot be compared to models evaluated on the original XNLI test set. However, the goal of this experiment is to study whether our principle of achieving domain alignment is beneficial on high-level tasks.

The results are included in Table 5. It is observed that IWDA generally improves transfer performance upon the zero-shot baseline, although the gains are not significant. As discussed in Section 4.1, the potentials of IWDA will be limited if the zero-shot representations are already largely aligned, which is reflected by the small performance gap between zero-shot and supervised results in our setup.

IWDA vs. DA. Besides demonstrating that SSL suffers from negative transfer under large prior shifts in our experiments in Section 4.2, we also discussed in Section 2.2 that domain alignment without prior shift correction leads to negative transfer.

In Fig. 11, we compare the performance on benchmark CoNLL German, which exhibits a prior shift from CoNLL English, from applying IWDA and domain alignment without prior shift correction (DA). CDAN refers to the conditioning technique by Long et al. (2018), which we use in our IWDA implementation by default. It is observed that without prior shift correction, as training progresses,

Table 6: CoNLL NER. Model is fine-tuned for 40 epochs with alternating updates.

	en	de	es	nl
IWDA	91.19	72.91	76.93	78.62
+ KD	90.68	73.83	77.87	79.96
IWDA (oracle)	91.32	72.47	77.24	79.31

transfer performance of vanilla DA deteriorates as the model makes more and more erroneous feature alignments caused by the shift.

Better Training of Adversary. As briefly mentioned in Section 4, while the IWDA models for the experiments in the main sections are trained for four epochs with simultaneous adversary and model updates for fair comparisons with SSL techniques, better performance could be obtained from IWDA if the optimality of the critic (adversary) can be better maintained throughout training with more passes over the data and alternating updates.

In Table 6, we observe that IWDA delivers much-improved performance on CoNLL when we perform four adversary updates per model update and allow 40 passes over the dataset (we also set `lambda_iw` to 1). Although this setting would lengthen the training time, this could be of practical interest for getting the best IWDA performance.

C Implementation Details of IWDA

A high-level description of IWDA was given in Section 3 with discussions that connect it to the motivating findings in Section 2. This section provides more details on the implementation of the algorithm along with the hyperparameter settings[7] used in our experiments (Table 7).

First, recall the joint objective, reproduced below, which consists of an alignment objective and an importance weight (IW) estimation objective,

$$\min_{g,h} \Big(\mathbb{E}_{(x,y)\sim p_S}[\ell(h(g(x)), y)] + \lambda D(p_S^{w,Z}, p_T^Z) \Big), \quad \min_{w} \sum_{\hat{y}\in[k]} \Big(p_T(\hat{y}) - \sum_{y} p_S(\hat{y}, y) w_y \Big)^2. \quad (3)$$

Note that the alignment objective contains the discrepancy between (important-weighted) source and target feature distributions and source task loss, added to prevent g from degenerating to a trivial representation with perfect source and target alignment, e.g. constant maps $x \mapsto c$.

Alignment Objective. We reformulate the alignment objective in eq. 3 as a minimax objective, and replace explicit evaluation of the discrepancy D with a critic function f (also known as discriminator or adversary; Ganin et al. 2016; Goodfellow et al. 2014). This turns the alignment objective into

$$\min_{g,h} \left(\mathbb{E}_{(x,y)\sim p_S}[\ell(h(g(x)), y)] + \lambda \max_{f\in\mathcal{F}} \left(\underbrace{\mathbb{E}_{x\sim p_S^{w,X}}[\ell_{\text{ad}}(f(g(x)), 0)] - \mathbb{E}_{x\sim p_T^X}[\ell_{\text{ad}}(f(g(x)), 1)]}_{(\triangle)} \right) \right),$$

where ℓ_{ad} is a specific choice of adversarial loss accompanied by an adversary function class $\mathcal{F}$, that takes as input the scores computed by the adversary and the domain labels (here, source domain is labeled 0). The role of the adversary is to distinguish features computed from inputs sampled on the source domain from those on the target domain, and output a score on each feature that reflects its domain identity. When the features are aligned, the adversary would not be able to distinguish source and target features.

In our implementation, we use Wasserstein-1 loss for ℓ_{ad}, whose corresponding adversary function class $\mathcal{F}$ is 1-Lipschitz functions (Arjovsky et al., 2017). It is shown that when f attains its maximum, the term $(\triangle)$ exactly computes the Wasserstein-1 distance between the importance-weighted feature

[7] We use the same settings on all IWDA experiments. For methodological discussions on hyperparameter tuning for unsupervised cross-lingual learning, see Artetxe et al. (2020c).

Table 7: IWDA hyperparameter settings.

Hyperparameter	lr	lambda_da	lr_ad	lambda_gp	lr_iw	lr_iw_stats	lambda_iw	lambda_iw_init
Value	1e-5	0.005	5e-4	10	5e-4	0.005	2	0.25

distributions, $p_S^{w,Z}$ and p_T^Z. We parameterize $\mathcal{F}$ as a shallow ReLU network of width 2048, and enforce the Lipschitz condition using soft constraints in the form of zero-centered gradient penalty, which penalizes the gradient of f between pairs of source and target features (Thanh-Tung et al., 2019).

Last but not least, for added stability, we condition the alignment of features on their (pseudo-)classes as proposed by Long et al. (2018) (CDAN). This is implemented as follows. Instead of feeding the adversary directly with features output by g, it receives the outer product between features and classifier output distributions, $g(x) \otimes h(g(x))$. This means that the input dimension of f is increased to the dimension of features multiplied by the number of output classes k. When k is large, Long et al. (2018) used random projections to reduce dimensionality, which we did not explore.

Putting everything together—minimax reformulation, Wasserstein-1 loss, zero-centered gradient penalty, and conditional alignment—upon receiving a mini-batch of source labeled examples, $(x_i, y_i)_{i=1}^n$, and target unlabeled examples, $(x'_j)_{j=1}^m$, the empirical alignment loss is

$$\frac{1}{n}\sum_{i=1}^{n} \ell(h(g(x_i)), y_i) + \lambda\left(\frac{1}{n}\sum_{i=1}^{n} f(g(x_i) \otimes h(g(x_i))) - \frac{1}{m}\sum_{j=1}^{m} f(g(x'_j) \otimes h(g(x'_i)))\right) - \frac{\mu}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \|(\nabla f)(\tilde{z}_i)\|^2,$$

where $\tilde{n} = \min(n, m)$ and $\tilde{z}_i$ is a random point on the line connecting $g(x_i)$ and $g(x'_i)$. Here, we have two hyperparameters: λ (lambda_da) controls the strength of alignment, and μ (lambda_gp) controls the strength on enforcing the Lipschitz condition. The learning rate of the adversary (lr_ad) is set to be larger than that of the model (lr; Heusel et al. 2017). AdamW optimizer is used throughout our experiments.

On this loss, we back-propagate gradients to f on maximizing the second and third terms, g on minimizing the first two terms, and h on minimizing the first term only. In our experiments in the main sections, all three models are optimized with simultaneous updates, but we note that alternating updates (standard in GAN literature) could allow IWDA to deliver further performance gains with longer training time. In Table 6, we observe improved mBERT CoNLL results when the training alternates between four updates to f and one update to g for 40 epochs.

IW Estimation Objective. We proposed several modifications to the IW estimation objective in Section 3 to improve its stability, and one of them is adding ℓ_2-regularization to w (Azizzadenesheli et al., 2018).

Given first moment statistics in the form of the source confusion matrix, $C_{ij} := p_S(\hat{Y} = i, Y = j)$, and target output distribution, $\nu_i := p_T(\hat{Y} = i)$, the regularized IWs objective minimizes

$$\sum_{i=1}^{k}\left(\nu_i - \sum_{j=1}^{k} C_{ij} w_j\right)^2 + \xi \sum_{i=1}^{k} (w_i - w^{(0)})^2, \tag{4}$$

where $w^{(0)}$ is an initial IW estimate (set to all-ones vector by default, i.e. no prior shift). Here, one hyperparameter that controls the strength of the regularization is introduced, ξ (lambda_iw). For simplicity, the ℓ_2-regularization is implemented with and documented as weight decay. Hence, ξ equals to lambda_iw multiplied by 2 times the learning rate of w.

Instead of initializing $w^{(0)} = \mathbf{1}$, because IWDA fine-tuning is preceded by zero-shot fine-tuning in our experiments and its decent performance reflects partial class-conditional feature alignment, we initialize it to a crude IW estimate w' obtained from the zero-shot model using the QP by Tachet des Combes et al. (2020). Since the initial alignment will not be perfect and w' not entirely reliable, we set $w^{(0)} = (1 - \rho)w' + \rho\mathbf{1}$, where ρ (lambda_iw_init) is a hyperparameter.

The statistics C, ν evolve during training as the feature representation computed by g changes, so it needs to be updated throughout the learning process. Re-evaluating them at each optimization step

on a held-out set would make training prohibitively slow, hence we update them with the statistics $\hat{C}, \hat{\nu}$ computed on training mini-batches: $C^{(t)} = (1-\eta)C^{(t-1)} + \eta\hat{C}$, and similarly for $\nu^{(t)}$. This update rate η (`lr_iw_stats`) is not the same as IW learning rate (`lr_iw`).

Lastly, to minimize potential effects from the miscalibration of neural models (Guo et al., 2017; Alexandari et al., 2020), hard confusion matrix and target output distribution are used in the estimation of IWs (Lipton et al., 2018; Garg et al., 2020).

D Description of SSL Baselines

Knowledge Distillation (KD). Our implementation follows the original paper by Hinton et al. (2015) closely (from which the Wu et al. (2020) variant is moderately different). Given a teacher LM fine-tuned either under zero-shot or using IWDA, we distill its knowledge on a newly initialized pre-trained student LM by minimizing the mean-squared-error difference of the teacher and student Softmax outputs on source and target data without labels. If labeled data from the source is available, we also train the student to minimize source task loss.

We set the temperature to 3, and the weight of the source task loss to 0.1. Learning rate is 1e-4, batch size is 32, and we train for 4 epochs.

In addition, Wu et al. (2020) observed that when distilling to pretrained LMs, better performance can be achieved if the bottom three layers of the student is frozen (motivated by findings from Wu & Dredze (2019)). We also apply this technique in our KD experiments.

Self-Training (ST). Our implementation follows Dong & de Melo (2019) closely. Given an LM fine-tuned on source domain only, ST continues fine-tuning the model by first assigning pseudo-labels to target data, collecting the ones with the highest confidence, and then train on them with the pseudo-labels. Source task loss is also minimized during self-training.

The hyperparameters of ST involve selecting how any pseudo-labeled examples to train on. On MARC with 200,000 training examples, at each epoch, we select top 4,000 most confident pseudo-labeled examples from the target domain conditioned on each (pseudo-)class, so that 20,000 examples are selected each round (MARC has five classes). These numbers are scales proportionally on datasets of other sizes. We perform self-training for 4 epochs, and set learning rate to 1e-5 and batch size to 128.

On IWDA + ST prior shift experiments, we select much less examples for self-training to avoid training on low-confidence examples from minority classes. This adjustment of hyperparameters relies on knowledge about the prior distribution on the target domain, which is approximated by the IW estimates from IWDA.