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, as evidenced by their decent zero-shot transfer performance. While analyses have attributed this success to having cross-lingually shared representations, its contribution to transfer performance remains unquantified. Towards a better understanding, 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 adaptation (IWDA), that performs feature 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 pretrain-fine-tune paradigm, where the pretrained LM is treated as a contextual feature extractor on which new output layers are added, and then the entire model is fine-tuned mildly on downstream tasks (Devlin et al., 2019; Conneau & Lample, 2019). On multilingual versions of these LMs that are pretrained on unannotated and unparallel texts from more than one language, such as multilingual BERT (mBERT) and XLM-R (Conneau et al., 2020a), it is discovered that the extracted features are shared across languages. In particular, the cross-lingual knowledge acquired by these models has enabled *zero-shot* cross-lingual transfer, where after being fine-tuned on data in a source language, not only do they work well when evaluated on the source, but also quite decently on almost all languages they have seen during pretraining (Wu & Dredze, 2019).

Their success on zero-shot learning has prompted numerous studies on the emergence of their cross-lingual abilities. Among the many explanations, a recent one suggests that the deep representation combined with parameter sharing induced intermediate features that are shared across languages (K et al., 2019; Conneau et al., 2020b; Muller et al., 2021). Following this hypothesis, we begin with the study of 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 features 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 unsupervised cross-lingual learning, where unlabeled data in the target language are given in addition to labeled ones in the source language (this setting is referred to by UCL hereafter), we propose the use of importance-weighted domain adaptation (IWDA) that performs feature 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 analysis and proposed method are heavily inspired by Tachet des Combes et al. (2020), which studied the issues with domain adaptation (DA) under the presence of class prior shifts and proposed the inclusion of prior shifts estimation and correction in DA. Our proposed method also shares many components with the DA-based UCL method on mBERT by Keung et al. (2019), while addressing its limitations discovered in the present work (Section 2.2).

Feature Alignment, Old and New. The idea of learning invariant features have powered a long history of approaches for cross-lingual transfer on neural models. Under zero-shot setting, this is performed by aligning word embeddings on dictionary pairs (Mikolov et al., 2013; Smith et al., 2017; Artetxe & Schwenk, 2019; Cao et al., 2019), where the dictionary could even be induced via adversarial training (Zhang et al., 2017; Artetxe et al., 2018; Lample et al., 2018). On Transformer-based LMs, feature sharing can be alternatively achieved via a second-stage pretraining on frozen Transformer body (Artetxe et al., 2020b; Pfeiffer et al., 2020). Under UCL and supervised settings, techniques based on feature alignment have been proposed on models 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).

Analyzing Multilingual Neural LMs. The majority of works that studied mBERT's cross-lingual ability performed probing experiments, which have identified key factors including 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). Another line of works analyzed intermediate representations and observed that mBERT learns cross-lingually shared representations (Wu & Dredze, 2019; Conneau et al., 2020b; Muller et al., 2021).

UCL and Adapting to Domain Shift. By treating languages as domains, the problem of UCL closely resembles that of adapting to domain shift. Indeed, methods proposed for UCL, including ours, can be and have been applied on domain shift. Besides DA-based methods (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 formed in recent analyses that attributed the cross-lingual transfer success of multilingual neural LMs to having shared representations, and to understand its implications on the design of unsupervised transfer algorithms, in this section, we attempt to quantify the relation between representation invariance and transfer performance.

Let g denote the feature mapping provided by the pretrained multilingual LM, and $z = g(x)$ denote its feature representation for input texts/tokens x, usually being the last-layer token embeddings for token classification tasks, and the embedding of a special start-of-sentence marker for sequence classification (symbolized as `[CLS]` on BERT). Let h denote the classifier that often consists of a new linear layer added on top of g and an appropriate activation function (e.g. Softmax for k-class classification). By fine-tuning both g, h on labeled task data from the source language domain $(x, y) \sim p_S$, a zero-shot transfer model for the downstream task at hand is obtained, where predictions are made through $\hat{y} = h(g(x))$,

$$X \xrightarrow{g_{\text{zero-shot}}} Z \xrightarrow{h_{\text{zero-shot}}} \widehat{Y}.$$

It is called zero-shot because it has not seen any task data from the target language $(x', y') \sim p_T$. Yet, when evaluated on them, decent performance has been observed on most languages seen during pretraining, and for a variety of NLP tasks (Wu & Dredze, 2019; Conneau et al., 2020a).

For our empirical analysis, we studied the representations learned by mBERT after being fine-tuned on two multilingual downstream tasks and datasets: sentiment analysis on Multilingual Amazon Reviews Corpus (MARC), and named-entity recognition (NER) on WikiANN dataset (descriptions in Section 4). To obtain more data points for our analysis, we generated and evaluated on 500 smaller datasets from MARC and around 700 from WikiANN via undersampling. Specifically, from the following decomposition of the joint feature-label distribution into class-conditional feature and marginal prior distributions,

$$p(z, y) = p(z|y)p(y),$$

we examine how 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, and (2) the distributional shift in the class priors, with no shift being $p_S(y = j) = p_T(y = j)$ for all j. For legibility in later sections, we denote the class-conditional feature distribution and the marginal prior distribution by $p^{Z|Y}$ and p^Y.

2.1 Invariance of Feature Representations

To quantify representation invariance, we measure the discrepancy between the class-conditional feature distributions between the source and target domains[1], $p_S^{Z|Y}$ and $p_T^{Z|Y}$, by averaging the discrepancies computed on each class (hereafter referred to as *conditional feature shift*),

$$\frac{1}{k}\sum_{j=1}^{k} D(p_S^{Z|Y=j}, p_T^{Z|Y=j}),$$

and the discrepancy measure we use takes the ℓ_1-distance of the feature means on the two domains[2],

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

Measuring distances between feature means of languages has also appeared in prior work on analyzing mBERT. For instance this quantity is referred to as *language centroids* by Libovický et al. (2020). The important difference here is that the effects of class prior shift are eliminated explicitly by conditioning on the class labels, which we argue is an artifact of the dataset, as arbitrary amounts of *unconditioned* feature shift can be produced from altering the class priors.

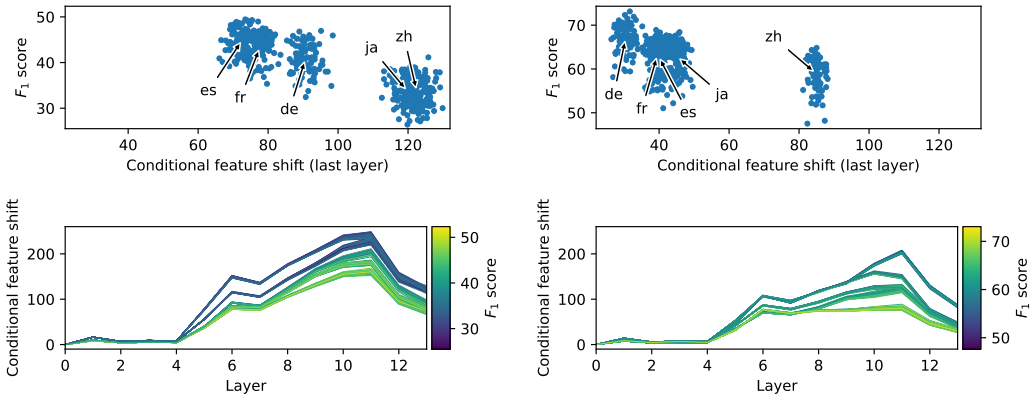
First, we study the contribution of representation invariance to zero-shot cross-lingual transfer. On a zero-shot mBERT model fine-tuned on the English portion of the MARC for four epochs, we plot on the top panel of Fig. 1a the F_1 scores[3] attained by the zero-shot model against conditional feature shifts between last-layer English features and those of the target language. It is observed that as representation invariance weakens, the model performance also decreases. This suggests that the model's zero-shot success is closely tied to having shared features and its implicit bias for selecting mostly invariant features during fine-tuning, and that better performance could be achieved by enhancing feature sharing and representation invariance.

Indeed, most successful zero-shot learning approaches proposed for mBERT explicitly aligned word embeddings on dictionary pairs (Cao et al., 2019), or compelled the model to use source features for target languages through a second-stage pretraining that adapts new language-specific parameters to a frozen source-pratrained Transformer body (Artetxe et al., 2020b; Pfeiffer et al., 2020). In addition, this finding is compatible with prior work on the cross-lingual ability of mBERT. For instance, linguistic similarity, a factor repeatedly argued to be important for transfer performance,

[1]The terms *domain* and *language* are used interchangeably in this paper.

[2]To be sure that first-moment statistics are informative, comparisons were made to RBF kernel maximum mean discrepancy (MMD) and CKA. We found that they correlate tightly on the measurement of class prior shift (Gretton et al., 2012; Kornblith et al., 2019).

[3]The use of macro-averaged F_1 score as the metric in this section is for studying the effects of class prior shift, where accuracy could be inflated on datasets with skewed priors (Azizzadenesheli et al., 2018).



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

(b) Bilingually supervised (English and target)

Figure 1: Class-conditional alignment of target language features with English features on MARC vs. mBERT transfer performance. **Top panels:** Final-layer features, with average F_1 scores and feature shifts marked by language. **Bottom panels:** Alignment of features as they pass through intermediate layers (lower means more aligned), with performance indicated by line colors (lighter means better). Performance correlates with the alignment of features on all Transformer layers. Lower-layer features see near-zero discrepancy due to using `[CLS]` for sentence embedding that are not informative until later layers; cf. token embeddings in Fig. 5a.

is reflected in the invariance of feature: Spanish, French and German are indeed more similar to English than Japanese and Chinese, whose features are less aligned with those of English. These results also corroborate domain adaptation generation bounds that related transfer performance to representation invariance (Ben-David et al., 2007).

While feature alignment is favorable for zero-shot transfer, is it also suitable for UCL? For our second study, we first recognize that the objective function of most unsupervised transfer algorithms consists of two parts: loss on labeled data, and terms that involve unlabeled data that are in theory linked to the true loss on unlabeled data. Based on this, we studied whether feature alignment is also favorable for UCL by comparing conditional feature shift to model performance after fine-tuned on the ideal objective, the combined true loss on source and target training data. The plot is provided on the top panel of Fig. 1b. Instead of observing an increase in conditional feature shifts, which would imply that language-specific features are preferred by mBERT over shared ones for good performance in each language, we observed better feature alignments between English and each of the target languages. This means that achieving representation invariance is not only an appropriate objective for zero-shot learning, but also for UCL, which we leverage in our proposed method.

Lastly, to demonstrate that the correlation between last-layer condition feature shift and model performance is not an artifact of having a linear classifier on these features, we show (1) on the bottom panels of Fig. 1 that conditional feature shift in intermediate layers also correlates with performance, as well as (2) *class-balanced* feature shift, plotted in Fig. 8 in the appendix, where the removal of prior shift effects is achieved by reweighting the class priors according to the uniform distribution (see Section 3 for formal definition).

2.2 Distributional Shift in Class Priors

Recent work on cross-lingual transfer has deliberately used class-balanced datasets for evaluation. Under zero-shot learning, this is for good reasons, otherwise the focus is shifted from "a high quality cross-lingual transfer to 'tricks' for how to best handle the class imbalance" (Schwenk & Li, 2018). However, for UCL, we argue that making the best use of unlabeled target data includes the detection and correction of prior shift, as we show in this section that class prior shift negatively affects cross-lingual transfer performance.

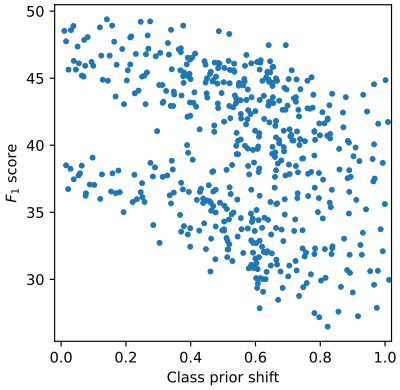


Figure 2: Distributional shifts in class priors between English and target datasets vs. zero-shot performance on MARC. Performance negatively correlates with prior shift.

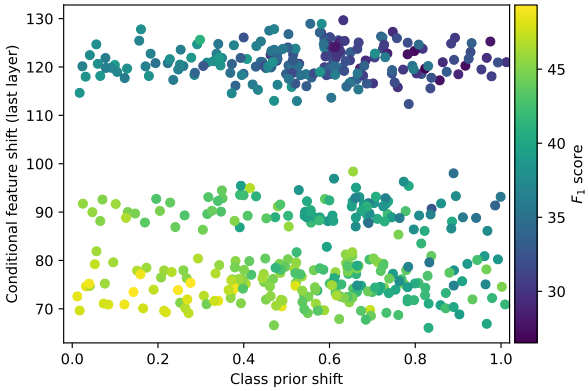


Figure 3: Conditional feature shift and class prior shift vs. zero-shot performance on MARC (indicated by scatter colors where lighter means better). Stronger correlations of each factor to performance than those in figs. 1 and 2 are revealed.

To examine its effects on cross-lingual transfer, we compare zero-shot transfer performance to the amount of prior shift, $D(p_S^Y, p_T^Y)$, which now reduces to total variation (TV), on the same 500 datasets undersampled from MARC. In fact, for this study, these datasets are undersampled according to class priors with varying degrees of shifts from that of English source data. The results are presented in Fig. 2, and show that zero-shot transfer performance generally suffers as prior shifts increases regardless of how well the features are aligned. Furthermore, its effects are more pronounced when the source prior distribution is skewed (Appendix A).

This finding calls for the estimation and correction of class prior shift when data from the target domain are available under UCL. Unfortunately, a lack of attention to this issue is discovered during our survey of the UCL literature, and found from our experiments that existing approaches based on semi-supervised learning do not correct for prior shift (Section 4.2), and those based on domain adaptation (DA) without explicit correction fails under even mild shifts (see Zhao et al. (2019) and Tachet des Combes et al. (2020) for a theoretical explanation). Interestingly, the DA-based method for mBERT proposed by Keung et al. (2019) circumvented this issue for NER on class-imbalanced CoNLL dataset (description in Section 4) by performing feature alignment on the sentence embeddings (from `[CLS]` special tokens) instead of the token embeddings of input words that are actual features to the linear classifier. This resulted in weaker alignments, and missed out opportunities for better results (Section 4.1). If the method were applied on token embeddings without prior shift correction, deteriorating performance would be observed (Fig. 7).

We close this section by putting both representation invariance and class prior shift together and illustrate their combined influence on zero-shot transfer (and cross-lingual transfer in general) in Fig. 3. Stronger correlations are revealed when one of the factors is fixed: (1) on target domains of the same level of feature alignment with the source, performance decreases with increased prior shift, and (2) on target domains of the same degree of prior shift, performance decreases when features are less strongly aligned. Lastly, figs. 1a, 3 and 8a for zero-shot transfer are also produced on 700 undersampled WikiANN NER datasets that covers a diverse set of 39 target languages, and included in figs. 5 and 9 in the appendix.

3 Importance-Weighted Domain Adaptation

The observations in Section 2 suggest that improvements to cross-lingual transfer can be made by (1) aligning the feature representations, (2) correcting for class prior shift, and (3) balancing the source dataset. For UCL, however, the main complication when attempting to achieve the above is that to perform principled feature alignment, class prior shift must be corrected using prior knowl-

edge of the target label distribution (Tachet des Combes et al., 2020). But not only is this quantity unknown, its estimation is generally unreliable when features do not align (Lipton et al., 2018).

Fortunately, because zero-shot fine-tuned multilingual LMs produce features that are largely aligned, the above complication can be overcome if we start from the zero-shot model and alternate between making small improvements to feature alignment and updating prior shift estimates on better aligned features. To this end, we propose class-importance-weighted domain adaptation (IWDA) for UCL. Below, a description of the two key components of IWDA is given, with implementation details relegated to Appendix B. We follow closely the work by Tachet des Combes et al. (2020).

Feature Alignment. While the ideal goal is conditional feature alignment, $p_S^{Z|Y=j} = p_T^{Z|Y=j}$ for all j, which eliminates erroneous alignments, it is hard to achieve under UCL as class labels are unknown. Thus, we aim for a weaker alignment that is necessary for conditional alignment, called *importance-weighted* alignment, $p_S^{Z,w} = p_T^Z$, where the importance weights (IWs) $w \in \mathbb{R}^k$ adjust the class priors via

$$p^w(x, y = j) = p(x|y = j)p(y = j)w_j.$$

Assume for now that the true class prior shift is known, then so are the true IWs, $w_j^\star = p_T(y = j)/p_S(y = j)$, through which $p_S^{Z,w^\star}$ and p_T^Z share the same class prior distribution, p_T^Y. While classic domain adaptation that aligned the *unweighted* source and target feature distributions, $p_S^Z = p_T^Z$, will suffer from erroneous alignments caused by prior shifts, the correction of prior shift in *importance-weighted* feature alignment protects against its effects (Tachet des Combes et al., 2020).

This leads to the following joint objective for the feature alignment part of IWDA, consisted of a discrepancy term between p_T^Z and $p_S^{Z,w}$ that is importance-weighted by a current IW estimate w, and a loss on the source domain that is added to prevent g from collapsing to uninformative and trivial mappings like constant features $g(x) = c$:

$$\min_{g,h} \left(\mathbb{E}_{(x,y)\sim p_S}[\ell(h(g(x)), y)] + \lambda D(p_S^{Z,w}, p_T^Z) \right).$$

Because sophisticated probability metrics D are difficult to compute while simpler ones lead to weak alignments, the discrepancy is usually approximated by an adversary following recent advances in adversarial learning, turning the above into a minimax objective (Ganin et al., 2016),

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

ℓ_{ad} is a specific choice of adversarial loss that is accompanied by an adversary function class $\mathcal{F}$, and f is referred to as the discriminator or critic. When f attains its maximum, it exactly computes the discrepancy induced by ℓ_{ad} (Goodfellow et al., 2014; Arjovsky & Bottou, 2017).

Hence, the training procedure interleaves updates to g, h w.r.t. minimizing the source loss and discrepancy approximated by f, between updates to f w.r.t. maximizing the adversarial loss for maintaining its optimality and correctness in its approximation of the true discrepancy. If the number of passes through the dataset is limited, simultaneous updates to all g, h, f can be performed with a gradient reversal layer (GRL) that trades maximum attainable performance for faster convergence.

Prior Shift Estimation. Since the true IWs $w^\star$ are unknown, they are estimated and updated throughout the alignment process using a moment-matching method referred to as the confusion matrix approach (Saerens et al., 2002; Lipton et al., 2018). Assume that the features are conditionally aligned, $p_S^{Z|Y=j} = p_T^{Z|Y=j}$ for all j (better known as the *label shift* assumption), then for each output label class $\hat{y} = h(g(x)) \in [k]$, we have

$$p_T(\hat{y} = i) = \sum_{j=1}^{k} p_T(\hat{y} = i|y = j)p_T(y = j) = \sum_{j=1}^{k} p_S(\hat{y} = i|y = j)p_T(y = j) = \sum_{j=1}^{k} p_S(\hat{y} = i, y = j)w_j^\star,$$

Combining this equation with the rest for $i \in [k]$, $w^\star$ could be recovered by optimizing for w w.r.t. minimizing $(p_T(\hat{y}) - \sum_y p_S(\hat{y}, y)w_y)^2$ for all $\hat{y}$, a quadratic optimization problem with constraints $w \geq 0$ and $w^\top p_S^Y = 1$ (Tachet des Combes et al., 2020).

Summarizing our UCL pipeline, from the pretrained multilingual LM, we (1) fine-tune it on source labeled data to obtain a zero-shot model, and then (2) continue fine-tuning with the IWDA objective for feature alignment using source and target domain unlabeled data, while (3) updating IW estimates as well as the required statistics of confusion matrix and target output distribution. We expect IW estimates to improve with better feature alignment, and vice versa.

4 EXPERIMENTS

We evaluate the performance of importance-weighted domain adaptation (IWDA) for UCL on multilingual BERT (mBERT) on two NLP classification tasks: named-entity recognition (NER) and sentiment analysis, covering both token and sequence-level classification. As in most work on cross-lingual transfer, English is used as the source language for transfer in our evaluations.

Hyperparameter settings for IWDA are relegated to Appendix B. Due to unsupervised learning, performance varies between different initializations, hence all results are averaged over at least three runs. IWDA training time is roughly the same as that of performing supervised training on same amounts of source and target labeled data.

Datasets[4]**.** For 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 contains examples in English, German, Spanish and Dutch, and while the latter covers more than 200 languages, we choose a subset of 40 as in the XTREME benchmark (Hu et al., 2020). There are four and three entity types respectively, and labels are in IOB2 format. Although these datasets are not class-balanced, the effects of prior shift are mostly mild, likely because it is a structured prediction problem (Appendix A).

For sentiment analysis, the Multilingual Amazon Reviews Corpus (MARC) is used, a balanced dataset with product reviews in English, German, Spanish, French, Japanese and Chinese. The task is to predict the ratings associated with each review on a 5-star scale (using review body only). This task is considered hard, due to ambiguities and noisy labels on reviews with 2–4 stars. Domain shift is observed on the Chinese portion, where it contains significantly more book reviews.

Baselines[5]**.** Two semi-supervised learning (SSL) techniques are compared to in our experiments. For NER, a technique based on knowledge distillation (KD) by Wu et al. (2020) is considered, where the task fine-tuned mBERT is distilled on pretrained mBERT. The bottom three layers of mBERT are frozen throughout training, a trick found to be helpful for transfer in Wu & Dredze (2019). We also implemented and considered a KD method (our KD) that follows more closely the original work by Hinton et al. (2015). On sentiment analysis, a method based on self-training (ST) is implemented and considered that follows closely the work by Dong & de Melo (2019). More details on our implementations are included in Appendix C.

4.1 RESULTS ON STANDARD BENCHMARKS

The results on the standard CoNLL NER and MARC sentiment analysis datasets are presented in tables 1 and 2, respectively. Results on WikiANN NER are in Table 4 in the appendix.

On CoNLL NER, the performance of SSL methods are better than IWDA on average. We suspect it is due to the noise introduced by DA, from optimizing for importance-weighted alignment instead of the ideal goal of class-conditional alignment, and also from insufficiently optimized critics. This is because for fair comparisons with baselines, we fine-tuned on only four passes over the training set and opted for simultaneous updates to all g, h, f, as opposed to alternating updates common in GAN literature that would otherwise require more passes and longer training time. Indeed, with alternating updates and longer training, IWDA delivers improved performance (Table 3 in the appendix).

[4]Zero-shot transfer benchmarks such as XNLI (Conneau et al., 2018) were not used, whose training data are machine-translated from the English set. This is because in UCL, we do not expect unlabeled data in the target language to be parallel to those in the source language, nor translation artifacts (Artetxe et al., 2020a).

[5]As we noticed in our experiments that SSL techniques typically requires separate hyperparameter settings optimized for each dataset, only ones with reported mBERT results are considered (in contrast, we used the same IWDA hyperparameters on all experiments).

Table 1: CoNLL NER UCL transfer results (from English; macro-averaged F_1).

	en	de	es	nl
Supervised and zero-shot learning				
Supervised (Wu & Dredze, 2019)	91.97	82.82	87.38	90.94
Zero-shot	90.57	69.77	74.14	78.28
Unsupervised learning				
Keung et al. (2019)	-	71.9	74.3	77.6
Wu et al. (2020)	-	73.22	76.94	**80.89**
Our KD	90.84	73.25	76.35	80.52
IWDA	90.77	72.56	76.11	78.63
+ Our KD	90.89	**73.71**	**77.14**	79.72
IWDA (oracle)	90.75	72.58	76.48	79.17

Table 2: MARC sentiment analysis UCL transfer results (from English; accuracy).

	en	de	es	fr	ja	zh
Supervised and zero-shot learning						
Supervised	58.50	61.19	57.76	57.05	57.97	53.74
Zero-shot	58.50	44.80	46.49	46.02	37.37	38.48
Unsupervised learning						
ST	58.25	50.74	48.80	47.96	42.10	41.40
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
IWDA (oracle)	55.96	51.95	50.83	**50.01**	**44.91**	**45.96**

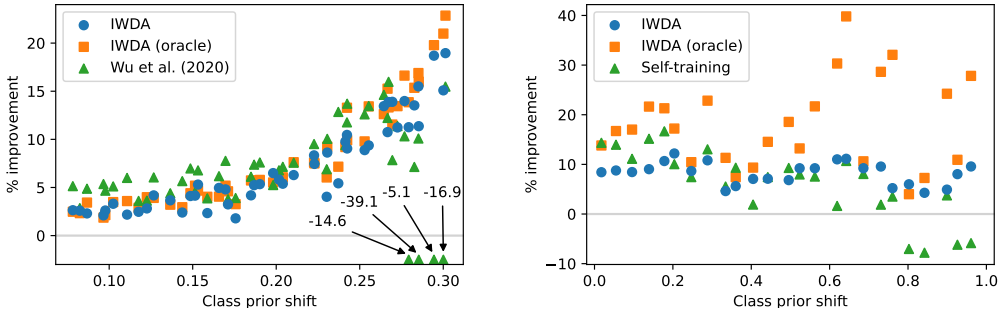
The remarks above and recent works that suggested the label smoothing effects of KD motivated the experiment of complementing IWDA with a final-stage KD (Yuan et al., 2020; Zhang & Sabuncu, 2020). With this procedure (IWDA + our KD), further performance gains were observed and in most cases surpassing that of KD alone. This is in fact expected, because SSL algorithms are designed on the assumption that labeled and unlabeled data are sampled from the same domain, so IWDA feature alignment can be viewed as a data preprocessing step to satisfy this assumption.

On MARC sentiment analysis, IWDA consistently outperforms the ST baseline, and further performance gains are also observed when IWDA is followed by ST. Note the decreased source performance on models fine-tuned with IWDA, an expected effect especially on harder problems since the IWDA joint objective does not solely minimize source error.

Finally, we include IWDA results where the ground-truth target domain priors (true IWs) are provided (oracle). On the datasets presented here, the small gaps between IWDA and oracle results shows the efficiency of our prior shift estimation. Yet, as observed on WikiANN results (Table 4), while the estimation procedure generally succeeds at producing accurate IW estimates, it is not always guaranteed. The average decrease in conditional feature shift is 21.60% and 22.33% with IWDA and IWDA (oracle) respectively on CoNLL (with 100% indicating perfect alignment), 51.53% and 64.77% on MARC, and 32.91% and 36.34% on WikiANN. We again expect that performing alternating IWDA updates could improve the stability and accuracy of IW estimates, but leave further investigations to future work.

4.2 Results Under Class Prior Shifts

We have demonstrated the performance of IWDA on datasets with mild to no prior shifts. Now, we study and compare the performance of IWDA and SSL under large prior shifts, a likely property of real-world data that has been insufficiently investigated for UCL. Similar to Section 2, macro-averaged F_1 score is used as the performance metric, and UCL methods are evaluated on the transfer from the English portion of the standard CoNLL dataset to 50 undersampled datasets of German CoNLL, and English portion of the standard MARC dataset to 25 undersampled datasets of Japanese



(a) CoNLL NER English to German

(b) MARC English to Japanese

Figure 4: Percent improvement over zero-shot performance of IWDA and SSL-based UCL baselines under varying class prior shifts, evaluated on datasets undersampled from benchmarks. SSL without prior shift correction worsens as shift increases.

MARC, all with varying amounts of prior shifts (in TV). Percent improvements over zero-shot performance are plotted in Fig. 4.

On CoNLL NER, while the KD-based SSL baseline performs generally better than IWDA under mild to no prior shifts as observed earlier, its performance quickly deteriorates as prior shift increases and is sometimes worse than doing nothing (zero-shot). On the other hand, IWDA is able to provide consistent improvement over zero-shot under all ranges of prior shifts. The same are observed on MARC sentiment analysis, where we evaluated IWDA and ST under even larger shifts.

These results show that not only are vanilla DA methods not protected from the effects of prior shift (Section 2.2), but SSL methods as well, which could amplify pseudolabel noise and result in worse performance than simply performing zero-shot transfer under the presence of large prior shifts. The relative success of IWDA means that it is possible to effectively estimate and correct for prior shift using unlabeled data, and is necessary if the best use of available data were to be made for UCL.

While small gaps between IWDA and oracle on CoNLL NER reflect the effectiveness of the prior shift estimation procedure of IWDA, on harder problems such as MARC, larger gaps are expected as the accuracy of the IW estimates is affected by the suboptimality of the classifier and the lack of class-conditional alignment that is now harder to achieve (Lipton et al., 2018). The average percent correction of prior shift on these experiments, defined as the decrease in TV between the true target domain priors and end-of-training estimation compared to the TV between source and target priors, is 87.33% on CoNLL (with 100% achieved with perfect IW estimation), and 28.08% on MARC with shifts larger than 0.8. On MARC with shifts smaller than 0.8 that are harder problems with noisy labels, the estimates are worse than assuming a uniform prior distribution (-335.80%), due to the reasons mentioned above. Yet, consistent improvements upon zero-shot baseline are still observed, indicating that the benefits of correct alignments outweight the harms from incorrect ones.

5 Conclusion

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 adaptation (IWDA) for unsupervised cross-lingual learning (UCL). Experiment results show that IWDA is effective at the estimation and correction of class prior shift from unlabeled data including large shifts where existing UCL approaches based on SSL fall short. As prior shift is a prevalent issue under unsupervised settings and IWDA has demonstrated that its estimation and correction are possible, more care to this issue is essential for the development of UCL algorithms that will work well on not just class-balanced NLP benchmarks but also real-world data. While the confusion matrix approach for importance weight (IW) estimation in our IWDA implementation is largely effective, future work on the search and study of improved IW estimation methods would enable IWDA to achieve better and more consistent transfer performance.

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.

Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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.

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.

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.

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.

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.

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.

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.

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 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.

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. Do Explicit Alignments Robustly Improve Multilingual Encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4471–4482, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.362.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3902–3910, June 2020. doi: 10.1109/CVPR42600.2020.00396.

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.

Zhilu Zhang and Mert Sabuncu. Self-distillation as instance-specific label smoothing. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2184–2195. Curran Associates, Inc., 2020.

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.

A Effects of Class Prior Shift When Source Prior Is Skewed

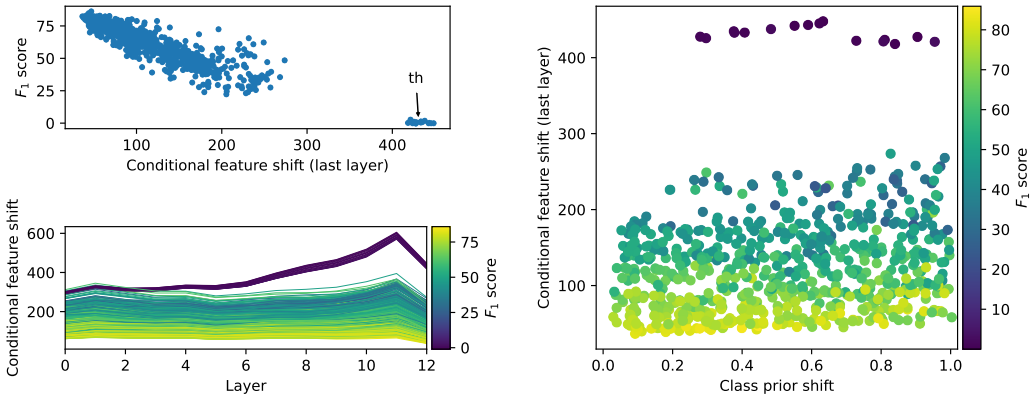
In Section 2, the relation between feature invariance, class prior shift and zero-shot transfer performance is studied when transferring from a relatively balanced source dataset. Here, we show that when the model is fine-tuned on source datasets with skewed prior distributions, the effects of class prior shift are more pronounced.

For this study, on MARC, we undersample from the English portion a source dataset according to the prior distribution of 0.1486, 0.0643, 0.0803, 0.1928 and 0.5141, reflective of the distribution of product ratings on Amazon.com. We produce the same plot of Fig. 3 for a mBERT model fine-tuned on this skewed dataset in Fig. 6a. Compared to when the source is class-balanced, here, the correlation between performance and conditional feature alignment has weakened, and that with class prior shift has strengthened.

On WikiANN, potentially due to NER being a structured prediction problem where there exist dependencies between tokens in the same sentence, although the original English portion of the dataset is considerably skewed (with 50.75% of the tokens belonging to class O, i.e. not part of an entity mention), there appears to be little to no correlation between class prior shift and model performance (Fig. 5b); not until we evaluated an mBERT that was zero-shot fine-tuned on undersampled English WikiANN where only sequences containing tokens with label O are kept, so that 80.92% of the tokens belong to class O (Fig. 6b).

B IWDA Details and Hyperparameters

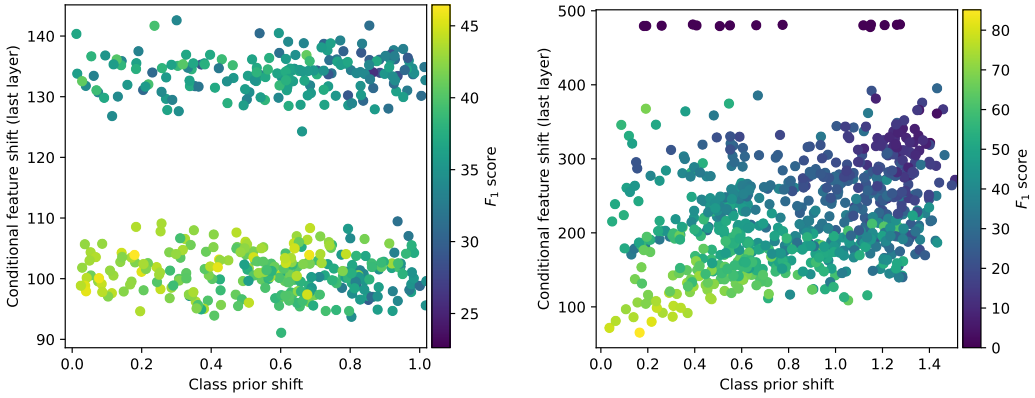
Feature Alignment. In our implementations, we used Wasserstein-1 loss for the adversarial loss ℓ_{ad} (Arjovsky et al., 2017), which we observed to be more responsive to importance-weighted class prior correction than cross-entropy loss, which corresponds to Jensen-Shannon divergence under optimality (Goodfellow et al., 2014; Ganin et al., 2016). The Lipschitz requirement on the adversarial function class $\mathcal{F} \ni f$ is satisfied with the soft constraints provided by zero-centered gradient penalty, which penalizes the gradient of f evaluated on random points on the feature space between source and target inputs (Thanh-Tung et al., 2019).



(a) Last and intermediate feature alignment

(b) Feature alignment and class prior shift

Figure 5: Feature invariance, class prior shift, and mBERT zero-shot transfer performance trained in English and evaluated on undersapmled WikiANN NER datasets, covering 39 target languages (cf. the same plots on MARC, figs. 1a and 3). Note that Thai is much less aligned with English.



(a) MARC

(b) WikiANN

Figure 6: Conditional feature shift and class prior shift vs. zero-shot performance on MARC and WikiANN. **Model was trained on skewed source training data.** Cf. figs. 3 and 5b (axis limits have changed), the impact on performance from class prior shift is more pronounced.

(Pseudo-)Class conditioned feature alignment, proposed by Long et al. (2018), is performed for added stability, under which the outer product between features $z = g(x)$ and the output probability produced by the current classifier $h(z)$, $z \otimes h(z)$, is taken and fed to the adversary f.

Combining everything, the joint loss computed on a minibatch of source examples $(x_i, y_i)_{i=1}^n$ and target examples $(x'_j)_{j=1}^m$ is given by the following (cf. the minimax objective in Section 3)

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

where $\tilde{n} = \min(n, m)$ and $\tilde{z}_i$ is a random point on the line connecting $g(x_i)$ and $g(x'_i)$ ($\lambda_{\text{da}} = 0.005, \lambda_{\text{gp}} = 10$ in our experiments). The adversary f is optimized to maximize this loss, Transformers feature mapping g is optimized to minimize the first two terms, and the linear classifier h is optimized to minimize the first term only.

Prior Shift Estimation. On the estimation of importance weights (IWs), for added stability, we use the regularized estimator inspired by (Azizzadenesheli et al., 2018) that penalizes the amount

of estimated shift. Here, for simplicity, we use ℓ_2-squared penalty implemented as weight decay, $\sum_{i=1}^{k}(w_i - w_0)^2$. By default, $w_0 = \mathbf{1}$, which corresponds to no prior shift.

Given first moment statistics of source confusion, $C_{ij} := p_S(\hat{y} = i, y = j)$, and target output distribution, $\nu_i := p_T(\hat{y} = i)$, the IWs are estimated by minimizing the following objective w.r.t. w,

$$\sum_{i=1}^{k}\left(\nu_i - \sum_{j=1}^{k} C_{ij} w_j\right) + \lambda_{\text{iw}} \sum_{i=1}^{k}(w_i - w_0)^2,$$

($\lambda_{\text{iw}} = \eta_{\text{iw}}$ in our experiments, where η_{iw} is the learning rate for w). Because obtaining the first moment statistics from the entire training set is expensive, they are updated with exponential decay on (unbiased) estimates $\hat{C}, \hat{\nu}$ obtained from minibatches in each forward-pass, $C_t = (1 - \eta_{\text{iw-stats}})C_{t-1} + \eta_{\text{iw-stats}}\hat{C}$, and similarly for $\hat{\nu}$ ($\eta_{\text{iw-stats}} = 0.005$ in our experiments).

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).

For initialization, before the start of IWDA fine-tuning, C_0, ν_0 are warmed up with data from the training sets, and w is computed using C_0, ν_0 with the quadratic program described in Tachet des Combes et al. (2020). Since w will be likely be incorrect as features are not yet aligned, we set $w_0 = (1 - \lambda_{\text{iw-init}})w + \lambda_{\text{iw-init}}\mathbf{1}$ ($\lambda_{\text{iw-init}} = 0.25$ in our experiments, and linearly decays to 0 over the first 10% of optimization steps).

Optimization. AdamW optimizer is applied to all parameters, with learning rate of 5×10^{-4} for the adversary and IW, and 1×10^{-5} for mBERT. Linear learning rate schedule is applied to mBERT and IW, and 10% warm up is applied to mBERT and λ_{da}. The batch size is set to 8. No truncation on the input texts is performed (although sequence length is limited to 512 due to modeling constraints of mBERT).

For each minibatch of examples from the source and target domains, one gradient step is applied on updating h, w (followed by a projection to satisfy $w \geq 0$ and $w^\top p_S^Y = 1$), and the first moment statistics C, ν are updated. When simultaneous adversary updates are used, one gradient step is also applied on g and f. Otherwise, with alternating updates, it is applied on either g or f.

This set of hyperparameters was found to work well for IWDA on most datasets and languages evaluated in the present work. For discussions on hyperparameter tuning for unsupervised cross-lingual learning, we refer interested readers to Artetxe et al. (2020c).

B.1 Number of Fine-Tuning Steps

All zero-shot models are obtained from fine-tuning mBERT on the source dataset for 4 epochs with a batch size of 8. Our IWDA experiments performed fine-tuning on source and target data for the number of optimization steps equal to 4 epochs over the target dataset, where batches of size 8 are sampled from each of the source and target domains at every step.

On NER experiments, we noticed that the average length of the sequences differs between languages. Because classification loss is averaged over the losses on each token in the minibatch, this means that the magnitude of back-propagated gradients received by each token will be on average smaller on domains with longer sequences compared to those with shorter sequences, which could in turn result in undertrained models on domains with longer sequences when the number of fine-tuning epochs is fixed.

Based on this observation, in our NER experiments, by using English data as reference, we scaled the number of steps on target domains so that each token from the target domain would be seen by the model the same number of times on average. More concretely, the scaling is given by (average # of tokens per sequence in target)/(average # of tokens per sequence in English). For CoNLL NER, the scaling is 1.17 for German, 2.18 for Spanish, and 0.88 for Dutch, and fine-tuning for the equivalent of 4 epochs over the English portion with batch size of 8 means 7135 steps on German, 9086 on Spanish, and 6954 on Dutch (with 7021 steps on English).

C SSL BASELINES DETAILS AND HYPERPARAMETERS

Knowledge Distillation (Our KD). In our implementation, we follow the original paper by Hinton et al. (2015) more closely. Where the KD method by Wu et al. (2020) sets Softmax temperature to 1 and does not use examples from the source domain for distillation, we set the temperature to 3 and use the source labeled data for distillation for both soft targets and hard targets with a relative weight of 0.1. Same as Wu et al. (2020), the losses on the soft targets are computed with MSE for numerical stability. The learning rate is 1×10^{-4}, and the batch size is 32.

In addition, Wu et al. (2020) observed that when distilling from a zero-shot fine-tuned teacher mBERT, better performance is achieved when the teacher is also trained with bottom three layers frozen. Therefore, for fair comparison, the zero-shot models on which our KD implementation is applied are also fine-tuned with frozen bottom layers (Table 1). We did not freeze any layers during the training of the models involved in IWDA experiments.

Self-Training (ST). Our implementation follows Dong & de Melo (2019) closely, in which we first obtain a zero-shot fine-tuned model, and then continue to fine-tune it on source and target data with high-confidence pseudolabels. Self-training is performed for 4 epochs. In each epoch, on MARC, we fine-tune on 50000 examples randomly selected on the source domain, and 4000 examples from each class on the target domain by confidence with their predicted labels. For experiments in Section 4.2, the sample sizes are scaled down to 5000 and 400, as the size of undersampled MARC datasets is 20000. The learning rate is 1×10^{-5}, and the batch size is 128.

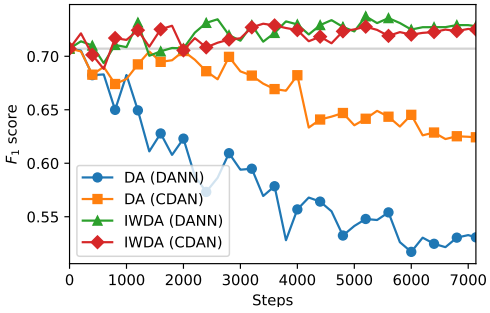
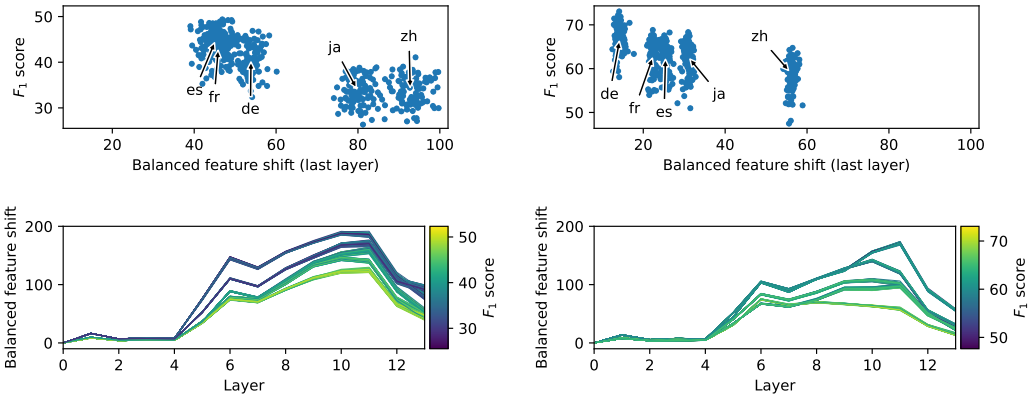


Figure 7: mBERT UCL transfer performance from English CoNLL NER to German, using IWDA feature alignment applied on last-layer token embeddings, and domain adaptation (DA) without class prior correction (using Wasserstein-1 adversarial loss). F_1 is plotted against the number of optimizer steps. Because of increasingly erroneous alignment as result of prior shift, DA performance deteriorates as training progresses. Performing (pseudo)class-conditional feature alignment (CDAN) protects DA from the effects of prior shift to some degree (Long et al., 2018), compared to DANN (Ganin et al., 2016).



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

(b) Bilingually supervised (English and target)

Figure 8: Class-balanced alignment of target language features with English features on MARC vs. mBERT transfer performance (cf. class-conditional alignment in Fig. 1). These are intended to illustrate that the correlation between performance and feature alignment is not an artifact of having a linear predictor on last-layer features.

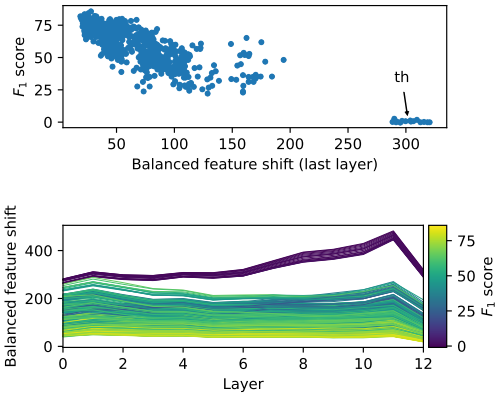


Figure 9: Class-balanced alignment vs. mBERT zero-shot transfer performance on WikiANN (trained on English only; cf. class-conditional alignment in Fig. 5a), see Fig. 8 for remarks.

Table 3: CoNLL NER UCL transfer results (from English; macro-averaged F_1), with 40 epochs of IWDA fine-tuning, $\lambda_{\text{iw}} = 2\eta_{\text{iw}}$, and alternating updates to the adversary with 4 updates to f per update to g (see Appendix B for definitions; cf. Table 1).

	en	de	es	nl
Supervised (Wu & Dredze, 2019)	91.97	82.82	87.38	90.94
Zero-shot	90.57	69.77	74.14	78.28
IWDA	91.19	72.91	76.93	78.62
+ Our KD	90.68	73.83	77.87	79.96
IWDA (oracle)	91.32	72.47	77.24	79.31

Table 4: WikiANN NER UCL transfer results (from English; macro-averaged F_1). Partial zero-shot results from explicitly word-aligned mBERT (Wu & Dredze, 2020) are included for comparison.

	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
Wu & Dredze (2020)	84.0	-	40.3	-	-	78.7	-	70.7	-	-
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
Wu & Dredze (2020)	-	-	79.0	-	67.2	-	-	-	-	-
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
Wu & Dredze (2020)	-	-	-	-	-	-	-	-	-	64.9
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
Wu & Dredze (2020)	-	-	-	-	-	-	-	69.1	-	52.0
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