DA-Net: A Disentangled and Adaptive Network for Multi-Source Cross-Lingual Transfer Learning

Ling Ge[1], Chunming Hu[1,2,3,*], Guanghui Ma[1], Jihong Liu[4,*], Hong Zhang[5],
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] College of Software, Beihang University, Beijing, China
[3] Zhongguancun Laboratory, Beijing, China
[4]School of Mechanical Engineering and Automation, Beihang University, Beijing, China
[5]National Computer Network Emergency Response Technical Team / Coordination Center of China, Beijing, China
{geling, hucm, maguanghui, ryukeiko}@buaa.edu.cn, zhangh@isc.org.cn

Abstract

Multi-Source cross-lingual transfer learning deals with the transfer of task knowledge from multiple labelled source languages to an unlabeled target language under the language shift. Existing methods typically focus on weighting the predictions produced by language-specific classifiers of different sources that follow a shared encoder. However, all source languages share the same encoder, which is updated by all these languages. The extracted representations inevitably contain different source languages' information, which may disturb the learning of the language-specific classifiers. Additionally, due to the language gap, language-specific classifiers trained with source labels are unable to make accurate predictions for the target language. Both facts impair the model's performance. To address these challenges, we propose a Disentangled and Adaptive Network (DA-Net). Firstly, we devise a feedback-guided collaborative disentanglement method that seeks to purify input representations of classifiers, thereby mitigating mutual interference from multiple sources. Secondly, we propose a class-aware parallel adaptation method that aligns class-level distributions for each source-target language pair, thereby alleviating the language pairs' language gap. Experimental results on three different tasks involving 38 languages validate the effectiveness of our approach.

Introduction

Recent advances in multilingual models (Devlin et al. 2019; Conneau and Lample 2019) have enabled significant improvements in many cross-lingual tasks, owing partly to the availability of large-scale annotated data. However, some tasks in no-annotated languages may suffer from the so-called "data hungriness" and fail to enjoy this technological advancement. This problem can be alleviated by conducting cross-lingual transfer learning (Zeng et al. 2022; Moreo et al. 2023; Li et al. 2023), which seeks to transfer language-independent knowledge from the labelled language (source) to the unlabeled language (target) (Sherborne and Lapata 2022; Ding et al. 2022; Ge et al. 2023), so that unlabeled languages can benefit from existing techniques.

Despite impressive results, most existing works focus only on the single-source setting and fail to consider a more realistic scenario where multiple sources with different languages are available. Since the target language may be similar to different source languages over various aspects, i.e., word order, capitalization, and script style (Hu et al. 2021; de Vries et al. 2022), multi-source cross-lingual transfer learning (Moon et al. 2019; Wu et al. 2020a) generally produces superior performance (Hu et al. 2021) compared to single-source algorithms and has garnered much interest.

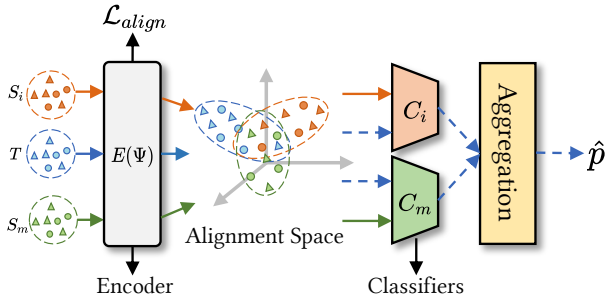

Figure 1: Existing works align all languages through alignment loss $\mathcal{L}_{align}$, causing the model only retaining information invariant across all languages.

The competitive approaches (Chen et al. 2019; Jin et al. 2022) typically learn a shared encoder, along with language-specific classifiers, which yield an ensemble prediction for target language. However, these approaches have two limitations. (1) Since multiple sources share one same encoder, the representations extracted by the encoder inevitably contain information from different source languages, which may confuse the optimisation of language-specific classifiers. (2) Due to the language gap, language-specific classifiers trained with source labels are unable to make accurate predictions for target samples. Although these methods attempt to alleviate this problem via performing all languages alignment (shown in Fig 1), unfortunately, this constraint is so strict that the shared features in each source-target language pair would be eliminated, thus deteriorating the adaptability of all language pairs and exacerbating this problem. Obviously, both limitations impair the model's performance.

In this paper, we propose a **D**isentangled and **A**daptive **Net**work (**DA-Net**) to tackle the two problems simultaneously. The proposed architecture has several separate lan-

*Corresponding authors.

guage branch networks following a shared encoder to extract specific structures. Each language branch network consists of a language-specific disentangler, adaptor and classifier.

Firstly, we propose a Feedback-guided Collaborative Disentanglement (FCD) method, which utilizes supervised signals from different classifiers to guide disentanglers to purify the shared representations, thereby mitigating mutual interference from multiple sources. We argue that a language-specific classifier can function as a detector to identify whether input representations contain the corresponding language information. The more information it contains about the related language, the better the language-specific classifier performs; otherwise, the classifier performs poorly. Considering this, we feed the representations generated by disentanglers into classifiers of their corresponding branch networks, and maximize the classification performance to preserve their respective language-specific information. Simultaneously, we feed these representations into other branches and minimize the classification performance to eliminate information irrelevant to the current source.

Secondly, we propose a Class-aware Parallel Adaptation (CPA) method, which seeks to align class-level distributions for source-target language pairs at each adaptor, thus bridging the language gap across languages. Class-level alignment can force the model better to acquire the shared semantics of class and avoid class-level feature mismatch issues across languages (Nguyen et al. 2021; Ge et al. 2023), thereby enhancing the performance of language-specific classifiers for target samples. To achieve this, we design a distribution-based contrastive learning to minimise the intra-class distribution distance and maximise the inter-class distribution distance across languages, thereby learning class discriminative shared features and ultimately improving language adaptation of language pairs.

Finally, we conduct experiments on three different tasks, including Named Entity Recognition (NER), Review Rating Classification (RRC) and Textual Entailment Prediction (TEP) involving 38 languages, and the results demonstrate the effectiveness of our DA-Net.

Summarily, we make the following contributions: (1) We propose the DA-Net for multi-source cross-lingual tasks, which utilizes multiple source languages to enhance the model's generalisation for the unlabeled target language. (2) We devise the FCD method to mitigate mutual interference from multiple sources. (3) We propose the CPA method to capture the shared class-level semantics of language pairs.

Related Works

To alleviate the data-scarcity issue for some languages, many cross-lingual transfer learning methods have been developed to learn well adaptation models. Most recent studies focus on the bilingual transfer case and can be grouped into three categories: (1) Data-based approaches utilize machine translation and label projection(Jain et al. 2019; Yang et al. 2022) to create pseudo-training data for the target language. (2) Feature-based approaches rely on features alignment to diminish the language shift (Chen et al. 2021; Ge et al. 2023). (3) The distillation-based methods (Wu et al. 2020b; Liang et al. 2021; Ma et al. 2022) enable the student network to gain task knowledge from soft labels predicted by the teacher network on the target language. However, these studies are designed on the single-source assumption and fail to deal with multiple source languages.

The Multi-source cross-lingual transfer is both feasible and valuable in practice and has received increasing attention in application fields. Recent methods, driven by the weighted combining rule, learn language-specific classifiers and obtain a weighted ensemble prediction for target samples. For instance, MulTS (Chen et al. 2021) trains one specific network for each source language separately to derive the combined final prediction, which causes a substantial computational burden. MAN (Chen et al. 2019) and G-MOE (Jin et al. 2022) force multiple source languages to share one same encoder, avoiding the huge parameter computation problem. They all perform alignment among multiple sources and target to extract language-invariant knowledge. However, this rigorous constraint results in too little information the encoder learns. In addition, none of these approaches considers the language interference issue associated with the shared encoder.

Methodology

In multi-source scenario, we consider M labeled source languages datasets $\{S_i\}_{i=1}^{M}$, where $S_i = \{(x_{s_i}^j, y_{s_i}^j)\}_{j=1}^{|S_i|}$, and one unlabeled target language dateset $T = \{(x_t^j)\}_{j=1}^{|T|}$. Specifically, $x_{s_i}^j = \{w_n\}_{n=0}^{L}$ denotes the j-th sentence, with w_n indicating the n-th token and L representing the sentence length. For RRC and TEP tasks, $y_{s_i}^j$ stands for the sentence-level label. For NER, $y_{s_i}^j = \{y_n^j\}_{n=0}^{L}$ is a label sequence, where y_n^j specifies the entity class corresponding to token w_n. In this paper, we aim to train a model with $\{S_i\}_{i=1}^{M}$ and T, and expect it to generalise well in the target language T.

Framework and Basic Pipeline

Model Framework As shown in Figure 2, DA-Net consists of a shared encoder and multiple source branch networks. We train each source branch with one source-target language pair to encode language-specific knowledge separately, and ultimately, the predictions of target samples can be derived by weighting the outputs of multiple language-specific classifiers. To alleviate the semantic interference from multiple sources, we propose a feedback-guided collaborative disentanglement (FCD) method to purify input representations. In addition, to mitigate the language gap of language pairs, we propose a class-aware parallel adaptation (CPA) method to align class-level distributions across languages. In detail, our DA-Net has the following components. Note that we use the RRC task as an example to present model details throughout the paper.

Shared Multilingual Encoder We adopt mBERT (Pires et al. 2019) as the feature extractor $\boldsymbol{E}$ to obtain semantic representations for different languages. Formally, given one source language sequence, for instance, $x_{s_i}^j$, $\boldsymbol{E}$ maps it to a shared latent space and produces representations $h_{s_i}^j = \boldsymbol{E}(x_{s_i}^j; \Psi)$, where Ψ is the encoder parameters.

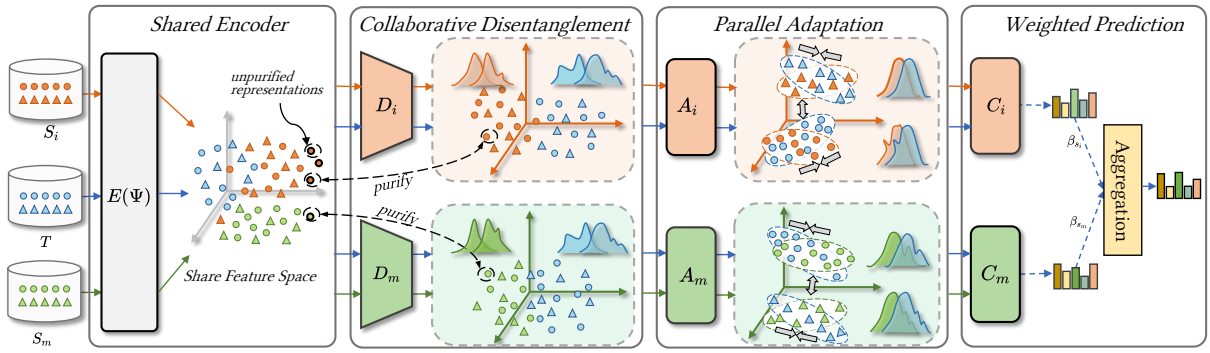

Figure 2: An illustration of our proposed model.

Since all source languages share the same parameters Ψ, which are in turn updated by all these languages, semantic representations of each source language extracted by $\boldsymbol{E}$ inevitably contain information from other source languages.

Disentangler To purify semantic representations from the shared $\boldsymbol{E}$, we equip each source language S_i with a disentangler D_i, one layer MLP and an activation function. Taking source representations $h^j_{s_i}$ as input, the D_i can produce new representations $z^j_{s_i} = D_i(h^j_{s_i}; \phi_i)$, where ϕ_i is the parameters of D_i. All disentanglers' parameters can be expressed as $\Phi = \{\phi_i\}^M_{i=1}$. Then, all these disentanglers, guided by our FCD method, can optimise new representations and disentangle multiple sources.

Adaptor To bridge the source-target language gap, we construct one adaptor A_{s_i} for each language S_i . The adapter is a one-layer MLP, followed by an activation function. We pass language representations $z^j_{s_i}$ through the corresponding adaptor and obtain $e^j_{s_i} = A_i(z^j_{s_i}; \theta_i)$ with θ_i denoting the parameters of A_i. We express parameters of M adaptors as $\Theta = \{\theta_i\}^M_{i=1}$. Then, we perform the CPA method on M adaptors to boost the adaptation of language pairs.

Classifier Similarly, we construct multiple classifiers $\{C_i\}^M_{i=0}$, each is a one-layer MLP, to obtain language-specific predictions. We input representations $e^j_{s_i}$ to classifier C_i, and finally obtain the prediction $p^j_{s_i} = C_i(A_i(z^j_{s_i}; \omega_i))$, where ω_i stands for parameters of C_i. All M classifiers' parameters can be expressed as $\Omega = \{\omega_i\}^M_{i=1}$.

With the multiple labelled source languages, our whole model can be optimized by minimizing cross-entropy loss:

$$\mathcal{L}_{CE}(\Psi, \Phi, \Theta, \Omega) = -\frac{1}{\sum^M_{i=1}|S_i|}\sum^M_{i=1}\sum^{|S_i|}_{j=1} y^j_{s_i} log(p^j_{s_i}) \quad (1)$$

Feedback-guided Collaborative Disentanglement

In this part, we detail our FCD method, which relies on supervised signals from multiple classifiers to guide disentanglers to perform a max-min game. In this way, the disentanglers can map each source language into one exclusive representation space, respectively, thus eliminating mutual interference among multiple sources. To illustrate this process, we take the source language S_i as an example (Fig 3).

Maximise Prediction Accuracy We sequentially feed the source representation $z^j_{s_i}$ produced by E_i to adaptor A_i and classifier C_i (located in source S_i branch) and obtain prediction $p^j_{s_i}$. To preserve the knowledge of source language S_i, in this step, we optimize the disentangler D_i to find representations that make the classifier C_i perform well. To achieve this, we minimize the following cross-entropy loss:

$$\mathcal{L}_{max}(\phi_i) = -\frac{1}{|S_i|}\sum^{S_i}_{j=1} y^j_{s_i} log(p^j_{s_i}) \quad (2)$$

Minimise Prediction Accuracy In this step, we expect disentanglers to learn representations on which other source branches' classifiers perform poorly, to remove these sources' information. We pass representations $z^j_{s_i}$ into adaptors $\{A_m\}^M_{m=1,m\neq i}$ and classifiers $\{C_m\}^M_{m=1,m\neq i}$, and obtain different predictions $\{p^j_{<s_i,s_m>}\}^M_{m=1,m\neq i}$. We optimize disentangler D_i to force each $p^j_{<s_i,s_m>}$ to approximate a uniform distribution q_{uni}. Since the uniform distribution presents the highest entropy and the most randomness (Gong, Liu, and Jain 2020), the above operation can maximize the prediction chaos of classifiers $\{C_m\}^M_{m=1,m\neq i}$ for representations $z^j_{s_i}$. Formally, we achieve the above by minimizing the MSE loss between $p^j_{<s_i,s_m>}$ and q_{uni}. Additionally, as different sources have different similarities to each other, undifferentiated disentanglement may impair the shared information between them. Thus, we employ the distance metric (i.e. $\alpha_{i,m}$) as the control factor to regulate the disentangling strength among source languages, denoted as:

$$\mathcal{L}_{min}(\phi_i) = -\frac{1}{M|S_i|}\sum^{M}_{\substack{m=1,\\ m\neq i}} \alpha_{i,m}\sum^{|S_i|}_{j=1}(p^j_{<s_i,s_m>} - q_{uni})^2 \quad (3)$$

where q_{uni} is a K-dimensional vector with all elements being $\frac{1}{K}$ and K is the number of classes. $\alpha_{i,m}$ is the MMD distance between S_i and S_m (see Eq 7).

To summarize, the loss $\mathcal{L}_{max}$ (Eq. 2) forces disentangler D_i to encode language information from S_i, while the loss $\mathcal{L}_{min}$ (Eq. 3) discourages D_i from encoding information of $\{S_m\}^M_{m=1,m\neq i}$. Consequently, the information from other

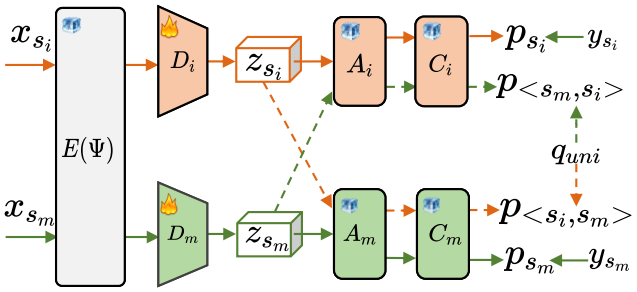

Figure 3: Collaborative disentanglement process.

sources can be removed from representations $z^j_{s_i}$. Eventually, we can use these two losses to optimise all disentanglers, and the total loss of this part can be expressed as:

$$\mathcal{L}_{FCD}(\Phi) = \sum_{i=0}^{M}(\mathcal{L}_{max}(\phi_i) + \mathcal{L}_{min}(\phi_i)) \tag{4}$$

Class-aware Parallel Adaptation

In this part, we present CPA method, which relies on aligning class-level distributions between language pairs on each adaptor to alleviate the language gap. Here, we utilize the language pair S_i and T as examples to illustrate the class-level adaptation process.

Class Pseudo Labels Since there is no annotation in the target language, the class-wise alignment between source and target is not trivial. For unannotated target samples, an intuitive way is to employ the training model to predict their pseudo labels, which may result in unstable predictions. Therefore, we create a momentum model $\mathcal{M}_{cur}$ derived from the training model instead. Since $\mathcal{M}_{cur}$ utilizes a moving average technique for step-level updates, it can provide consistent model representation and stable predictions at each step (He et al. 2020).

$$\mathcal{M}_{cur} = \gamma * \mathcal{M}_{cur} + (1-\gamma) * \mathcal{M}_{cur-1} \tag{5}$$

where $\gamma \in (0, 1)$ is the moving coefficient, cur and $cur-1$ denote the current and previous moment, respectively.

Then, we feed target samples to $\mathcal{M}_{cur}$ and obtain M set of probabilities. The pseudo labels of target samples can be derived via weighted summing M different probabilities (Eq. 9). The class labels for source are their golden labels.

Class-aware Distribution Alignment After obtaining class labels for source and target samples, we can intuitively utilize the lately prevalent supervised contrastive learning (Khosla et al. 2020) to achieve class-level alignment between language pair. However, this method tends to pull all samples of the same class together, making the learned representations ignore intra-class diversity, which is unfavourable to generalize to the target language (Mou et al. 2022). Motivated by the fact that larger intra-class diversity facilitates transfer learning (Zhao et al. 2021; Feng et al. 2021), we design a distribution-based contrastive learning technique to bridge the gap between languages, and maintain the diversity of intra-class features (Wang et al. 2023).

In detail, we assume that source and target samples belonging to class k obey the distributions $P^k_{s_i}$ and Q^k_t, respectively. We regard one of them as the anchor distribution, such as $P^k_{s_i}$, the other as the positive distribution, such as Q^k_t, and the remaining class-wise distributions $\{P^o_{s_i}\}^K_{o=1,o\neq k}$ (from source) and $\{Q^o_t\}^K_{o=1,o\neq k}$ (from target) as negative distributions, denoted as $P^k_{s_i,neg}$. Our goal is to minimise the distance of same-class distributions while maximising the distance of different-class distributions, denoted as:

$$\mathcal{L}_{CA}(\Psi, \theta_i, \omega_i) = -\sum_{k=1}^{K} \log \frac{2e(d(P^k_{s_i}, Q^k_t)/\tau)}{\sum_{neg} e(d(P^k_{s_i}, P^k_{s_i,neg})/\tau) + \sum_{neg} e(d(Q^k_t, Q^k_{t,neg})/\tau)} \tag{6}$$

where $e(\cdot)$ denotes exponential function $exp(\cdot)$. $Q^k_{t,neg}$ denotes negative distributions of Q^k_t. τ is a temperature factor. $d(\cdot)$ is a distance function between language distributions.

As shown, the intra-class distribution discrepancy (numerator) is minimized to align language distributions within a class. In contrast, the inter-class distribution discrepancy (denominator) is maximized to push the distributions of different classes away from decision boundaries. These two are jointly optimized to improve the adaptation performance.

Distribution Metric To optimise distances of different class-wise distributions, we suggest using maximum mean discrepancy (MMD) (Borgwardt et al. 2006) metric to measure these distances ($d(\cdot)$ in Eq.6). MMD is an effective non-parametric metric for comparing two distributions based on two sets of observed samples belonging to them (Gretton et al. 2012). Suppose we have two class-wise distributions $P^u_{s_i}$, Q^v_t and their respective observations samples $\{x^j_u\}^{|S^u_i|}_{j=1}$, $\{x^j_v\}^{|T^v|}_{j=1}$, with u, v denoting different classes. Through introducing kernel tricks and empirical kernel mean embeddings (Gretton et al. 2012), the empirical estimation of MMD between $P^u_{s_i}$ and Q^v_t is expressed as:

$$\begin{aligned}\mathcal{D}^2_{\text{MMD}}(P^u_{s_i}, Q^v_t) = &\frac{1}{|\boldsymbol{S}^u_i|\,|\boldsymbol{S}^u_i|}\sum_{j=1}^{|\mathcal{S}^u_i|}\sum_{l=1}^{|\mathcal{S}^u_i|} G\left(e^j_u, e^l_u\right) \\ &+ \frac{1}{|\boldsymbol{T}^v|\,|\boldsymbol{T}^v|}\sum_{j=1}^{|\mathcal{T}^v|}\sum_{l=1}^{|\mathcal{T}^v|} G\left(e^j_v, e^l_v\right) \\ &- \frac{2}{|\boldsymbol{S}^u_i|\,|\boldsymbol{T}^v|}\sum_{j=1}^{|\mathcal{S}^u_i|}\sum_{l=1}^{|\mathcal{T}^v|} G\left(e^j_u, e^l_v\right)\end{aligned} \tag{7}$$

where e^j_u and e^j_v are representations corresponding to samples x^j_u and x^j_v, respectively, generated by adaptor A_i. Note that, $x^j_u \in S^u_i$ (class u of source S_i), $x^j_v \in T^v$ (class v of target language T). G refers to the Gaussian kernel.

Methods	af	hr	it	fr	fa	ur	zh-yue	he	id	az	ja	et	eo	no	sh	Avg
Source Language: en, es and ru																
mBERT	81.33	83.99	82.83	85.57	56.93	38.69	49.45	60.93	66.47	72.82	36.35	78.81	68.84	83.13	81.55	68.51
M-MOE	80.41	82.26	83.99	83.49	55.00	46.54	51.35	**63.34**	64.04	70.75	37.25	77.44	68.75	82.35	82.05	68.60
MAN	80.22	84.01	83.14	84.95	57.41	49.34	48.65	60.38	64.69	72.47	36.73	80.00	65.82	84.30	79.03	68.74
MulTS	81.28	84.26	83.44	84.22	53.20	42.32	48.82	61.43	66.68	73.73	35.01	82.87	67.25	84.15	79.21	68.52
G-MOE	79.12	85.39	84.38	85.12	58.16	49.71	49.55	62.58	70.36	73.31	33.59	80.97	66.39	86.00	82.82	69.83
DA-Net	**83.57**	**86.24**	**85.02**	**87.16**	**66.08**	**56.79**	**52.44**	61.84	**72.51**	**74.52**	**38.38**	**83.65**	**70.19**	**86.58**	**87.50**	**72.83**
Source Language: en, ar and zh																
mBERT	77.31	79.81	78.28	81.37	65.74	37.26	68.06	61.75	67.33	69.30	46.08	78.62	60.59	77.92	56.24	67.04
M-MOE	77.58	78.88	82.33	79.75	65.87	48.24	69.86	63.32	64.43	69.83	43.86	79.49	62.49	80.78	58.57	68.28
MAN	78.17	80.19	81.07	80.67	65.34	46.33	70.14	61.89	64.04	70.76	42.64	79.84	61.50	78.01	56.91	67.81
MulTS	80.33	81.92	81.78	82.85	53.87	52.44	65.86	63.01	58.96	71.85	43.22	**82.90**	61.49	82.03	60.50	68.20
G-MOE	78.64	82.29	82.88	83.13	62.17	52.07	70.26	63.61	74.33	73.47	47.64	81.15	60.80	80.69	72.50	71.04
DA-Net	**81.46**	**83.46**	**83.20**	**83.91**	**80.81**	**69.63**	**72.42**	**63.73**	**81.27**	**74.39**	**49.57**	81.83	**62.53**	**83.53**	**83.34**	**75.67**

Table 1: The NER task results on the Wikiann dataset.

We can employ the MMD metric to calculate distances between each pair of class-wise distributions in Eq 6, regardless of whether these distributions originate from source or target, and belong to the same class or not. In addition, the MMD can also be utilized to derive the value of $\alpha_{i,m}$ (Eq 3). In this case, we adopt samples from S_i and S_m to estimate the distance between language distributions P_{s_i} and P_{s_m}.

Subsequently, we will implement class-level distribution alignment on each branch via $\mathcal{L}_{CA}$. Eventually, our class-aware parallel adaptation loss can be defined as follows:

$$\mathcal{L}_{CPA}(\Psi, \Theta, \Omega) = \sum_{i=0}^{M} \mathcal{L}_{CA}(\Psi, \theta_i, \omega_i) \tag{8}$$

Training and Prediction

We first describe the training procedure, followed by the target prediction of our model.

Trainning Procedure To equip our model with initial classification performance to serve as language detectors in disentangling stage and to acquire pseudo labels in adaptation stage, firstly, we warm up our model via $\mathcal{L}_{CE}$ (Eq 1) with multiple labelled source languages in the first epoch.

Then, we iteratively perform disentangling and adaptation via $\mathcal{L}_{FCD}$ (Eq 4) and $\mathcal{L}_{CPA}$(Eq 8) at the batch level. In the disentangling stage, we fix all the parts except disentanglers, and minimize the loss $\mathcal{L}_{FCD}$ to update these disentanglers. In the adaptation step, we fix all the disentanglers and update the other components of the model by minimizing the losses $\mathcal{L}_{CE} + \eta\mathcal{L}_{CPA}$, with η denoting the trade-off factor.

Target Prediction We use x_t^j as an example to describe the prediction process. Firstly, we input this sample into our model and obtain multiple predictions $\{p^j_{<t,s_i>}\}_{i=1}^{M}$ from language-specific classifiers. Then, all M predictions are weighted via a point-to-set Mahalanobis distance-based metric (McLachlan 1999), which measures the similarity between target sample x_t^j and each source.

$$y_t^j = \sum_{i=1}^{M} \beta(x_t^j, S_i) p^j_{<t,s_i>} \tag{9}$$

where $\beta(x_t^j, S_i) = -((x_t^j - \mu_{s_i})^T \Sigma_{s_i}^{-1} (x_t^j - \mu_{s_i}))^{\frac{1}{2}}$. The μ_{s_i} and $\Sigma_{s_i}^{-1}$ are the mean encoding and the inverse covariance matrix of S_i, respectively.

Experiments and Analysis

Experiment Setting

Datasets For NER task, we employ the adopt the benchmark dataset Wikiann [1] (Rahimi, Li, and Cohn 2019), where each word is marked by the BIO scheme, and annotated with LOC (Location), PER (Person), or ORG (Organisation).

For RRC task, we adopt the benchmark Amazon Reviews Corpus [2] (Keung et al. 2020) dataset, where each customer review is classified into five sentiment star ratings (classes).

For TEP task, we adopt the benchmark dataset XNLI [3] (Keung et al. 2020), where each sentence pair is labeled with entailment, neutral or contradiction. Due to the computational resources limitation, we take $20,000$ samples on each class for each language separately as training set.

Our experiments cover 38 languages and contain 4 low-resource languages, namely Esperanto (eo), Norwegian (no), Serbo-Croatian (sh), and Cantonese (zh-yue). These languages are not involved at all during the pre-trained model mBERT pretraining. For all datasets, we adopt the original training, development, and evaluation sets. Note that, the source and target data of the Amazon Reviews Corpus and XNLI share the same label distribution, while those of Wikiann have different label distributions, as detailed in the given dataset addresses. To simulate the zero-shot scenario, we train the model with both the labelled source and unlabeled target training sets, validate the model on the source validation set, and evaluate the model on the target test set.

[1] https://huggingface.co/datasets/wikiann

[2] https://huggingface.co/datasets/amazon_reviews_multi

[3] https://huggingface.co/datasets/xnli

Methods	de	fr	bg	th	sw	vi	hi	de	fr	bg	th	sw	vi	hi	Avg
	Source Language: en, es and ru							Source Language: en, ar and zh							
mBERT	68.22	70.49	67.29	52.07	50.17	67.11	60.77	66.81	68.54	66.37	53.19	50.68	67.45	60.12	62.09
M-MOE	68.90	69.92	67.96	52.91	49.76	67.13	57.28	68.42	68.80	67.18	54.55	50.71	66.26	60.79	62.18
MAN	69.04	69.15	67.94	53.19	49.82	66.94	60.65	67.48	68.90	67.94	53.29	49.82	67.32	60.65	62.30
MulTS	68.32	69.50	67.20	53.37	48.54	67.66	61.45	67.14	68.68	67.10	52.55	49.40	67.96	60.69	62.11
G-MOE	68.89	69.71	68.67	53.59	50.73	66.33	61.46	67.63	69.77	67.59	52.96	49.16	68.35	59.80	62.62
DA-Net	**69.96**	**71.54**	**69.18**	**55.29**	**51.16**	**68.49**	**61.47**	**69.94**	**70.98**	**68.88**	**55.52**	**51.71**	**68.96**	**61.52**	**63.90**

Table 2: The TEP task results on the XNLI dataset.

Source	en,es, fr			en, ja, zh			Avg
Methods	de	ja	zh	de	fr	es	
mBERT	50.08	40.46	40.77	49.46	50.74	50.42	46.99
M-MOE	50.94	40.86	40.92	49.69	51.94	49.28	47.27
MAN	50.19	40.94	41.42	48.84	50.16	51.33	47.15
MulTS	49.88	39.41	40.25	**49.98**	49.34	50.02	46.48
G-MOE	49.90	41.36	41.19	49.75	51.18	50.34	47.29
DA-Net	**50.99**	**42.48**	**42.54**	49.84	**53.58**	**52.24**	**48.61**

Table 3: The RRC task results on the Amazon dataset.

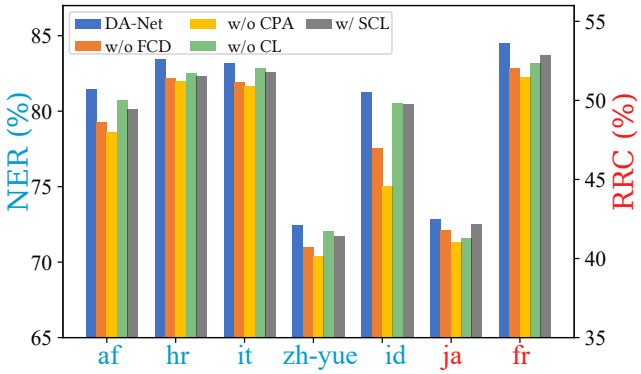

Figure 4: Ablation Study.

Implementation Details Following previous works (Keung et al. 2020; Ma et al. 2022), we adopt token-level F1 metric for NER task and accuracy metric for RRC and TEP tasks. We train our model using Adam (Kingma and Ba 2015) optimizer. We set the batch size to 128 for NER and 64 for other tasks. We use maximum sequence length = 128, the moving coefficient $\gamma = 0.0001$, and the dropout = 0.5 empirically. We utilize the grid search technology to obtain the optimal super-parameters, including the temperature factor τ selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, the trade-off factor η selected from $\{0.01, 0.05, 0.1, 0.5\}$, the learning rate for encoder selected from $\{1e-5, 3e-5, 5e-5\}$ and for other components selected from $\{0.0001, 0.0005\}$. We repeat all experiments 5 times and report their mean results.All experiments are implemented using PyTorch and are conducted on NVIDIA Tesla V100 GPU.

Experimental Results and Aanlysis

Baseline and SOTAs We compare our DA-Net with several previous multi-source approaches. **(1) mBERT** (Pires et al. 2019), **(2) M-MOE** (Guo et al. 2018), **(3) MAN** (Chen et al. 2019), **(4) MulTS** (Wu et al. 2020a) and **(5) G-MOE** (Jin et al. 2022). Note that the original M-MOE and MAN use LSTM as the encoder. For a fair comparison, we replace the encoder with mBERT.

Performance Comparison To evaluate the model performance, we conduct experiments using two distinct source combinations for each task: the same language family (Setting 1) and different language families (Setting 2). For example, in NER, sources en, es, and ru belong to the same language family (Indo-European), while en, ar, and zh belong to the Indo-European, Afro-Asian, and Sino-Tibetan, respectively. As shown in Table 1, 2 and 3, our method outperforms the baseline and the SOTAs in almost all target languages for three tasks under different settings. In detail, for NER, we exceed the SOTA G-MOE by 3% points in Setting 1,and 4.63% in Setting 2 on average. Since our method employs FCD and CPA strategies in training, we can simultaneously reduce interference among sources and enhance source-target languages' adaptation.

More specifically, in Setting 2, our model is 18.64% and 17.56% higher than SOTA for fa and ur languages. While both belong to the Indo-European language family, they are written using the Arabic alphabet. Benefiting from our CPA method, we can preserve the shared knowledge of each language pair. Thus, these target languages can transfer grammatical and semantic knowledge from en and script knowledge from ar. In contrast, the SOTA model forces all languages to be aligned, compromising the exclusive knowledge of the language pairs. In addition, our method also performs well when the target language is outside the range of the sources' language family, e.g., id (Austronesian language family), ja (Ryukyu language family). On low-resource languages such as zh-yue, eo, no, sh, our model still surpasses SOTA, demonstrating prominent generalisation.

Ablation Study To explore the contribution of each component, we design five variant models. (1) DA-Net$_{w/o\ \mathrm{FCD}}$ removes the FCD method. (2) DA-Net$_{w/o\ \mathrm{CPA}}$ wipes out the CPA method. (3) DA-Net$_{w/o\ \mathrm{CL}}$ employs MMD loss between language distributions instead of class-aware MMD loss in the CPA method. (4) DA-Net$_{w/\ \mathrm{SCL}}$ replaces the distributional alignment to supervised contrast learning in

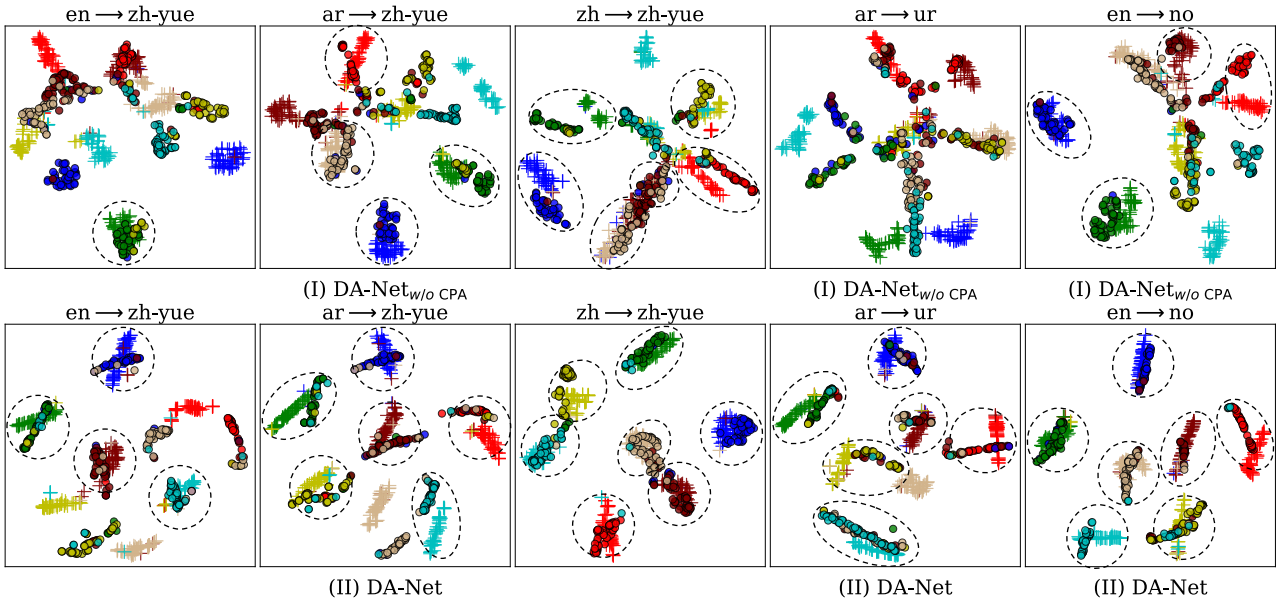

Figure 5: Our method achieves more class-level alignment and learns better class-discriminative representations. The plus (+) and circles (•) indicate representations of the source and target languages. Different colours represent different classes.

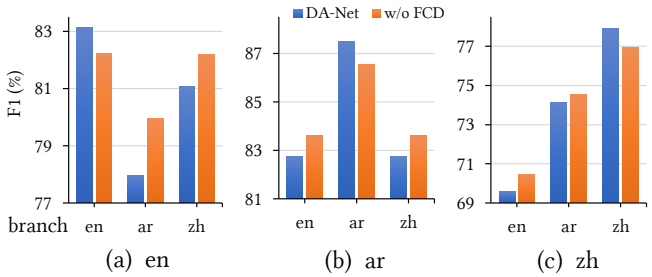

Figure 6: Collaborative Disentanglement Analysis.

the CPA method. As shown in Fig 4, for NER, compared to DA-Net, the af of DA-Net$_{w/o\ \mathrm{FCD}}$ decreases by 2.16%. This suggests that the FCD method facilitates performance improvement by alleviating the language interference issue. DA-Net$_{w/o\ \mathrm{CPA}}$ decreased by 6.25% on id, indicating that CPA method can effectively improve model generalisation since it bridges the gap of language pairs. The slight drop of DA-Net$_{w/\ \mathrm{SCL}}$ and DA-Net$_{w/o\ \mathrm{CL}}$ indicates that the distribution alignment and class information are helpful for final performance. Similar results can be observed on RRC task.

Collaborative Disentanglement Analysis To demonstrate that the FCD approach can mitigate interference between multiple sources, we compare the performance of DA-Net and DA-Net$_{w/o\ \mathrm{FCD}}$ in the NER task in Figure 6. We input three source languages (en, ar and zh) into the three branch networks of the two models, and get prediction results. For example, Figure 6 (a) is the results of en in three branches. As shown, DA-Net improves the F1 of each source language on the corresponding branch and decreases its F1 on the other branches. This indicates that DA-Net can purify representations from the shared encoder, keeping language-specific classifiers from interfering with each other during training, and improving the performance of language-specific classifiers for their corresponding source.

Visualize Representations To present that the CPA method can achieve class-level alignment in each branch, we compare the representations produced by adaptors of DA-Net and DA-Net$_{w/o\ \mathrm{CPA}}$. Fig 5 (a) visualises representations of target zh-yue and three different sources (en, ar, zh), while Fig 5 (b) visualises representations of other two target languages and different sources. Due to the language gap, the feature distributions of source and target language produced by DA-Net$_{w/o\ \mathrm{CPA}}$ are significantly different. Many target language examples of one class are incorrectly aligned to source examples of a different class, resulting in poor adaptation. In contrast, DA-Net correctly aligns more class-level distributions and captures more shared knowledge across languages. More importantly, DA-Net successfully separates and pulls apart clusters of different classes, learning better class-differentiated representations.

Conclusion

This paper presents a model DA-Net for multi-source cross-lingual tasks. DA-Net proposes a FCD approach to purify input representations of the classifier, thus mitigating inter-multisource interference and optimizing the learning of language-specific classifiers. In addition, DA-Net introduces a CPA method to improve language pairs' adaption, thus enhancing the performance of language-specific classifiers for target languages. Experimental results demonstrate DA-Net is effective and outperforms previous SOTAs.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work is funded by the National Key Research and Development Program of the Ministry of Science and Technology of China (No. 2021YFB1716201). Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by Kernel Maximum Mean Discrepancy. In *Proceedings 14th International Conference on Intelligent Systems for Molecular Biology 2006, Fortaleza, Brazil, August 6-10, 2006*, 49–57.

Chen, W.; Jiang, H.; Wu, Q.; Karlsson, B.; and Guan, Y. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. 743–753. Association for Computational Linguistics.

Chen, X.; Awadallah, A. H.; Hassan, H.; Wang, W.; and Cardie, C. 2019. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 3098–3112. Association for Computational Linguistics.

Conneau, A.; and Lample, G. 2019. Cross-lingual Language Model Pretraining. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 7057–7067.

de Vries, W.; et al. 2022. Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 7676–7685. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Ding, K.; Liu, W.; Fang, Y.; Mao, W.; Zhao, Z.; Zhu, T.; Liu, H.; Tian, R.; and Chen, Y. 2022. A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning. In Calzolari, N.; Huang, C.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.; Ryu, P.; Chen, H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S., eds., *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 4372–4380. International Committee on Computational Linguistics.

Feng, Y.; Jiang, J.; Tang, M.; Jin, R.; and Gao, Y. 2021. Rethinking supervised pre-training for better downstream transferring. *arXiv preprint arXiv:2110.06014*.

Ge, L.; Hu, C.; Ma, G.; Zhang, H.; and Liu, J. 2023. ProKD: An Unsupervised Prototypical Knowledge Distillation Network for Zero-Resource Cross-Lingual Named Entity Recognition. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 12818–12826. AAAI Press.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly De-Biasing Face Recognition and Demographic Attribute Estimation. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, 330–347. Springer.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. J. 2012. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 13: 723–773.

Guo, J.; et al. 2018. Multi-Source Domain Adaptation with Mixture of Experts. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 4694–4703. Association for Computational Linguistics.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. of CVPR*.

Hu, Z.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; and Tu, K. 2021. Multi-View Cross-Lingual Structured Prediction with Minimum Supervision. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2661–2674. Association for Computational Linguistics.

Jain, A.; et al. 2019. Entity Projection via Machine Translation for Cross-Lingual NER. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 1083–1092. Association for Computational Linguistics.

Jin, H.; Dong, T.; Hou, L.; Li, J.; Chen, H.; Dai, Z.; and Qu, Y. 2022. How Can Cross-lingual Knowledge Contribute Better to Fine-Grained Entity Typing? In Mure-

san, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 3071–3081. Association for Computational Linguistics.

Keung, P.; Lu, Y.; Szarvas, G.; and Smith, N. A. 2020. The Multilingual Amazon Reviews Corpus. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 4563–4568. Association for Computational Linguistics.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, Z.; Huang, S.; Zhang, Z.; Deng, Z.; Lou, Q.; Huang, H.; Jiao, J.; Wei, F.; Deng, W.; and Zhang, Q. 2023. Dual-Alignment Pre-training for Cross-lingual Sentence Embedding. *CoRR*, abs/2305.09148.

Liang, S.; Gong, M.; Pei, J.; Shou, L.; Zuo, W.; Zuo, X.; and Jiang, D. 2021. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. In *Proc. of KDD*.

Ma, J.; Chen, B.; Gu, J.; Ling, Z.; Guo, W.; Liu, Q.; Chen, Z.; and Liu, C. 2022. Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition. In *Proc. of EMNLP*.

McLachlan, G. J. 1999. Mahalanobis distance. *Resonance*, 4(6): 20–26.

Moon, T.; Awasthy, P.; Ni, J.; and Florian, R. 2019. Towards Lingua Franca Named Entity Recognition with BERT. *CoRR*, abs/1912.01389.

Moreo, A.; et al. 2023. Generalized Funnelling: Ensemble Learning and Heterogeneous Document Embeddings for Cross-Lingual Text Classification. *ACM Trans. Inf. Syst.*, 41(2): 36:1–36:37.

Mou, Y.; He, K.; Wang, P.; Wu, Y.; Wang, J.; Wu, W.; and Xu, W. 2022. Watch the Neighbors: A Unified K-Nearest Neighbor Contrastive Learning Framework for OOD Intent Discovery. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 1517–1529. Association for Computational Linguistics.

Nguyen, M. V.; Nguyen, T. N.; Min, B.; and Nguyen, T. H. 2021. Crosslingual Transfer Learning for Relation and Event Extraction via Word Category and Class Alignments. In *Proc. of EMNLP*.

Pires, T.; et al. 2019. How Multilingual is Multilingual BERT? In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4996–5001. Association for Computational Linguistics.

Rahimi, A.; Li, Y.; and Cohn, T. 2019. Massively Multilingual Transfer for NER. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 151–164. Association for Computational Linguistics.

Sherborne, T.; and Lapata, M. 2022. Zero-Shot Cross-lingual Semantic Parsing. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 4134–4153. Association for Computational Linguistics.

Wang, W.; Li, H.; Ding, Z.; Nie, F.; Chen, J.; Dong, X.; and Wang, Z. 2023. Rethinking Maximum Mean Discrepancy for Visual Domain Adaptation. *IEEE Trans. Neural Networks Learn. Syst.*

Wu, Q.; Lin, Z.; Karlsson, B.; Lou, J.; and Huang, B. 2020a. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 6505–6514. Association for Computational Linguistics.

Wu, Q.; Lin, Z.; Karlsson, B. F.; Huang, B.; and Lou, J. 2020b. UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. In *Proc. of IJCAI*.

Yang, J.; Huang, S.; Ma, S.; Yin, Y.; Dong, L.; Zhang, D.; Guo, H.; Li, Z.; and Wei, F. 2022. CROP: Zero-shot Cross-lingual Named Entity Recognition with Multilingual Labeled Sequence Translation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 486–496. Association for Computational Linguistics.

Zeng, J.; Jiang, Y.; Yin, Y.; Wang, X.; Lin, B.; and Cao, Y. 2022. DualNER: A Dual-Teaching framework for Zero-shot Cross-lingual Named Entity Recognition. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 1837–1843. Association for Computational Linguistics.

Zhao, N.; Wu, Z.; Lau, R. W. H.; and Lin, S. 2021. What Makes Instance Discrimination Good for Transfer Learning? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.