UniPSDA: Unsupervised Pseudo Semantic Data Augmentation for Zero-Shot Cross-Lingual Natural Language Understanding

Dongyang Li[1,2], **Taolin Zhang**[2], **Jiali Deng**[1], **Longtao Huang**[2], **Chengyu Wang**[2], **Xiaofeng He**[1], **Hui Xue**[2]
[1]School of Computer Science and Technology, East China Normal University
[2]Alibaba Group
{dongyangli0612, jialideng1127}@gmail.com, hexf@cs.ecnu.edu.cn
{zhangtaolin.ztl, kaiyang.hlt, chengyu.wcy, hui.xueh}@alibaba-inc.com

Abstract

Cross-lingual representation learning transfers knowledge from resource-rich data to resource-scarce ones to improve the semantic understanding abilities of different languages. However, previous works rely on shallow unsupervised data generated by token surface matching, regardless of the global context-aware semantics of the surrounding text tokens. In this paper, we propose an **Un**superv**i**sed **P**seudo **S**emantic **D**ata **A**ugmentation (UniPSDA) mechanism for cross-lingual natural language understanding to enrich the training data without human interventions. Specifically, to retrieve the tokens with similar meanings for the semantic data augmentation across different languages, we propose a sequential clustering process in 3 stages: within a single language, across multiple languages of a language family, and across languages from multiple language families. Meanwhile, considering the multi-lingual knowledge infusion with context-aware semantics while alleviating computation burden, we directly replace the key constituents of the sentences with the above-learned multi-lingual family knowledge, viewed as pseudo-semantic. The infusion process is further optimized via three de-biasing techniques without introducing any neural parameters. Extensive experiments demonstrate that our model consistently improves the performance on general zero-shot cross-lingual natural language understanding tasks, including sequence classification, information extraction, and question answering.

Keywords: Cross-Lingual Representation, Data Augmentation, Zero-Shot Learning

1. Introduction

Cross-lingual representation learning facilitates resource-rich information to boost the performance of under-resourced languages in various downstream natural language understanding (NLU) tasks, such as text classification (Huang, 2022; Rathnayake et al., 2022; Li et al., 2021), sentiment analysis (Szolomicka and Kocon, 2022; Sazzed, 2020), information extraction (Huang et al., 2022a; Ahmad et al., 2021; Wang et al., 2019; Fan et al., 2019), and question answering (Limkonchotiwat et al., 2022; Perevalov et al., 2022). Although existing cross-lingual works (Li et al., 2023; Clouâtre et al., 2022) share explicit language semantics across different languages, they generally rely on supervised parallel corpora and simple, shallow unsupervised mechanisms such as back translation (Lam et al., 2022; Nishikawa et al., 2021) and random deletion (Sun et al., 2022).

The previous data augmentation (DA) approaches in cross-lingual representation learning can be roughly divided into two categories: supervised parallel data augmenters and unsupervised shallow data augmenters.

1. *Supervised Parallel Data Augmenter:* These works (Fernando et al., 2023; Lai et al., 2022; Riabi et al., 2021) utilize annotated parallel corpora (e.g., bilingual dictionaries and translation tools) to augment the training data by aligning the same meanings across different languages for low-resource tasks. However, the collection process for these parallel corpora is time-consuming and relies on human annotation.

2. *Unsupervised Shallow Data Augmenter:* Unlike the supervised approaches mentioned above, these methods employ unsupervised easy data augmentation (EDA) techniques (e.g., back translation, random deletion, and random replacement) to generate additional training samples for model training (Nishikawa et al., 2021; Bari et al., 2021; Chen et al., 2021). These methods focus solely on the surface semantics of the input samples to match cross-lingual data without considering the deeper linguistic connections.

As shown in Figure 1, techniques like “random deletion” and “random replacement” may alter the sentence’s intended meaning. Hence, we aim to expand the multilingual training samples based on a

Work done when Dongyang Li was doing an internship at Alibaba Group. Dongyang Li and Taolin Zhang contributed equally to this work. Correspondence to Chengyu Wang and Xiaofeng He.

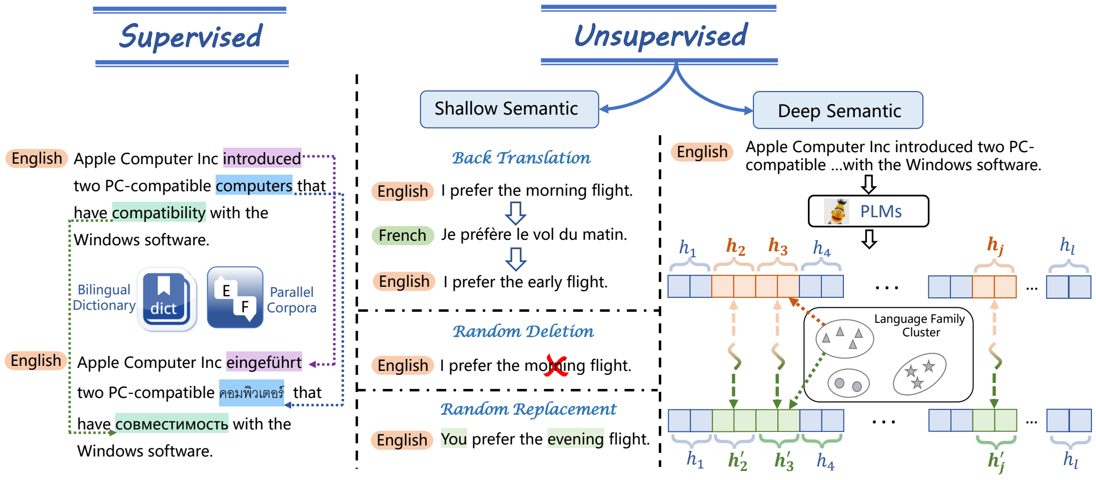
Supervised
Unsupervised
Shallow Semantic
Deep Semantic
English
Apple Computer Inc introduced two PC-compatible computers that have compatibility with the Windows software.
Bilingual Dictionary
dict
E
F
Parallel Corpora
English
Apple Computer Inc eingeführt two PC-compatible คอมพิวเตอร์ that have совместимость with the Windows software.
Back Translation
English
I prefer the morning flight.
French
Je préfère le vol du matin.
English
I prefer the early flight.
Random Deletion
English
I prefer the morning flight.
Random Replacement
English
You prefer the evening flight.
English
Apple Computer Inc introduced two PC-compatible ...with the Windows software.
PLMs
h_1 h_2 h_3 h_4 ... h_j ... h_l
Language Family Cluster
h_1 h'_2 h'_3 h_4 ... h'_j ... h_l

Figure 1: Examples of previous data augmentation techniques, including supervised methods that rely on parallel data, and unsupervised methods which carry the risk of losing sentential semantic coherence.

deep semantic understanding that the model can automatically derive, such as from the hidden layers of a pre-trained language model (PLM).

To overcome the issues mentioned above, we propose an **Un**supervised **P**seudo **S**emantic **D**ata **A**ugmentation (UniPSDA) mechanism, which mainly consists of two modules:

- **Domino Unsupervised Cluster:** To provide high-quality multilingual representations for performing the subsequent deep unsupervised data augmentation, we group languages into a hierarchical structure organized by language families[1] to learn multilingual relations. We perform the clustering process via the domino chain process[2] to collect semantically similar words across different languages by comparing the embeddings themselves, a method we name Domino Unsupervised Cluster. Specifically, the domino cluster is a chain-rule process comprised of three different sequential stages: the single language stage, the language family stage, and the multi-language stage.

- **Pseudo Semantic Data Augmentation:** Considering that previous data augmentation methods focus on the surface of naive training samples, we employ the learned multilingual internal representations to address the semantic deficiencies of the training samples. Specifically, the domino clustering-enhanced ultimate multilingual representations directly replace the important positions' hidden states in training samples, as recognized by the `<subject,verb,object>` (SVO) structure. The potential incompatibility phenomena of inserting clustering multilingual representations may result in biased parameter learning. To further alleviate the misalignment between the replaced embeddings space and the context output space of PLMs, we introduce three debiasing optimal transport affinity regularization techniques to make the learning process faster and more stable.

2. Methodology

2.1. Model Notations

The architecture of UniPSDA is shown in Figure 2. The goal of cross-lingual natural language understanding is to utilize a source language dataset $\mathcal{D}_{\text{lang}} = (\mathcal{X}_{\text{lang}}, \mathcal{Y}_{\text{lang}})$ to train a model $\mathcal{M}$. Then we apply the trained model $\mathcal{M}$ to tasks in other target languages $\mathcal{D}_{\text{lang}'} = (\mathcal{X}_{\text{lang}'}, \mathcal{Y}_{\text{lang}'})$, where $\mathcal{X}$ denotes the input samples and $\mathcal{Y}$ is the label set. In our work, each sentence of the training data is denoted as $S_i = (w_{i1}, w_{i2}, \cdots, w_{ij}, \cdots, w_{il_i})$, where w_{ij} denotes the j-th word in sentence S_i and l_i is the maximum word count of the sentence. The hidden state of word w_{ij} is $h_{w_{ij}} \in \mathbb{R}^{|u| \times d}$, where $|u|$ is the maximum number of tokens contained in the word and d is the dimension of the hidden state. The hidden state of sentence S_i is $h_{s_i} \in \mathbb{R}^{|L_s| \times d}$, where $|L_s|$ is the sentence's maximum token length. The specific notations for the three clustering stages in the `Domino Unsupervised Cluster` are as follows:

- In the single language stage, the words in the m-th single language G_{Sin_m} are clustered into $|G_{\text{Sin}_m}|$ clusters. The t-th cluster is denoted as Clu^{sin}_{mt}.

- In the language family stage, the words in the n-th language family G_{Fam_n} are clustered into

[1] https://www.ethnologue.com/browse/families
[2] https://en.wikipedia.org/wiki/Domino_effect

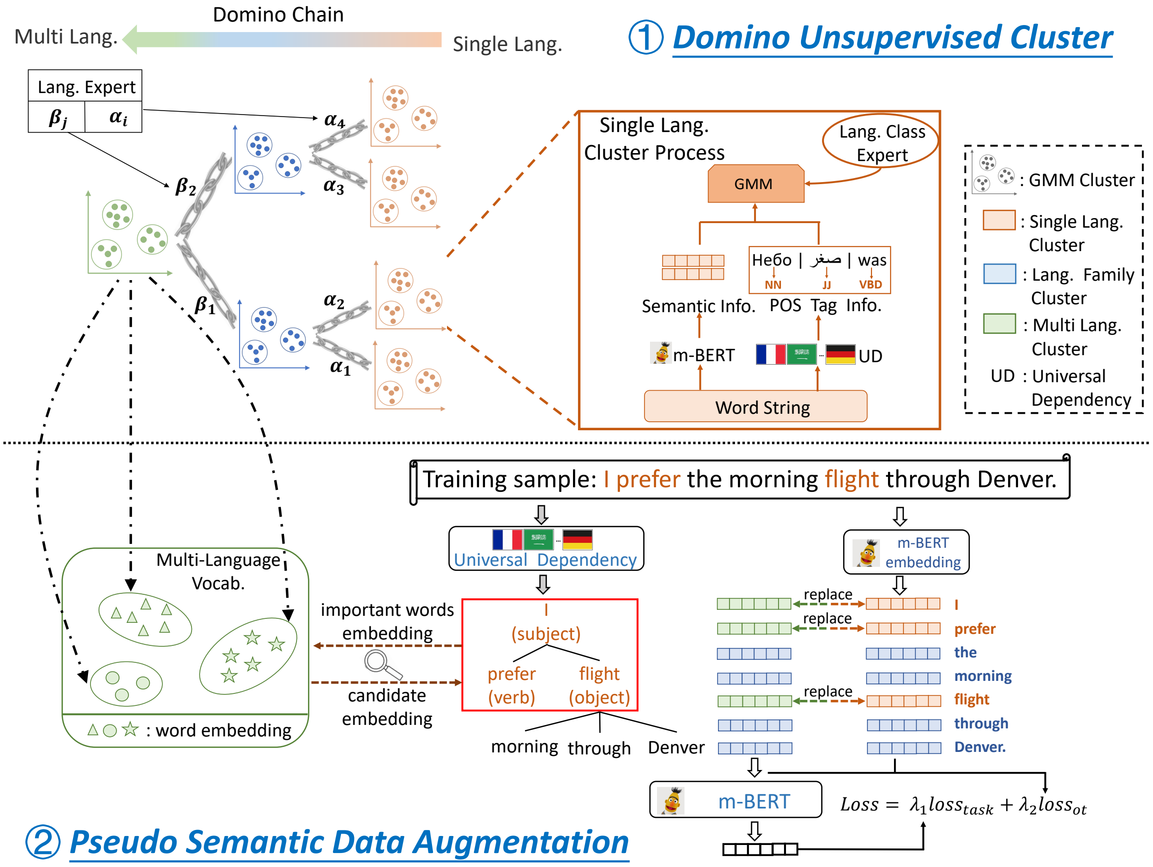

Figure 2: Model Overview of UniPSDA. (Best viewed in color)

$|G_{\text{Fam}_n}|$ clusters. The g-th cluster is denoted as Clu^{fam}_{ng}.

- In the multi-language stage, all language families are collected into G_{Mul}. All the words in G_{Mul} are clustered into $|G_{\text{Mul}}|$ clusters. The q-th cluster is denoted as Clu^{mul}_{q}.

2.2. Text Encoder

In this paper, m-BERT (Devlin et al., 2019) is utilized as our encoder[3] to obtain the hidden states, which are averaged from the embeddings of the first and last layers. The final hidden state of the j-th word in sentence S_i is formulated as:

$$h_{w_{ij}} = \frac{1}{2}\left(\mathcal{F}_{\text{first}}\left(w_{ij}\right) + \mathcal{F}_{\text{last}}\left(w_{ij}\right)\right) \quad (1)$$

where $\mathcal{F}_{\text{first}}$ and $\mathcal{F}_{\text{last}}$ denote the representations from the first and last layers, respectively. We average these representations to obtain the sentence's hidden state h_{s_i}.

2.3. Domino Unsupervised Cluster

To enable the model to learn relevant word information corresponding to different languages, we perform three hierarchical, chain-rule-based clustering steps, sequentially applied to representations of varying language granularities.

2.3.1. Single Language Cluster

In the single language cluster stage, we aim to group similar words within a specific language. To refine the clustering process, we clarify that "similar words" refers not only to semantic similarity but also to the concordance of part-of-speech (POS) tags. For instance, verbs with the meaning of "hope" are grouped together, distinct from nouns with similar meanings, thereby incorporating lexical POS knowledge into the clustering. Initially, we employ the Universal Dependencies[4]-based PyTorch tool Stanza[5], to obtain the POS tag for each word. The training data contains 17 types of POS tags, and we represent each word with a 17-dimensional one-hot

[3] Other multilingual pre-trained language models can also be considered as the backbone.

[4] https://universaldependencies.org/

[5] Stanza is an off-the-shelf cross-lingual linguistic analysis package. URL: https://stanfordnlp.github.io/stanza/

vector v_{POS} to signify the initial tag representations. This one-hot vector is then mapped to a context-aware space by a linear function $(W_{\text{POS}} v_{\text{POS}} + b_{\text{POS}})$. The final embeddings $h^{\text{final}}_{w_{ij}}$ are obtained by concatenating the original word representations with the projected POS tags:

$$h^{\text{final}}_{w_{ij}} = [h_{w_{ij}} \mid\mid (W_{_{\text{POS}}} v_{_{\text{POS}}} + b_{_{\text{POS}}})] \tag{2}$$

where "||" denotes the concatenation operation. Words with similar final embeddings are clustered together using an expectation-maximization algorithm based on Gaussian Mixture Models (GMM). For the m-th language, we obtain $|G_{\text{Sin}_m}|$ clusters at the end of the clustering process:

$$\begin{aligned} Clu^{\text{sin}}_{m1},\ Clu^{\text{sin}}_{m2},\ \ldots,\ Clu^{\text{sin}}_{m|G_{\text{Sin}_m}|} = \text{GMM}(\\ h_{w_1}, h_{w_2}, \ldots, h_{w_{|G_{\text{Sin}_m}|}}) \end{aligned} \tag{3}$$

where $|G_{\text{Sin}_m}|$ is the total number of words in the m-th language.

2.3.2. Language Family Cluster

In the Ethnologue[6] linguistic categorical tree, each language is considered a leaf node. Language families serve as ancestor nodes within this tree structure, and all descendant nodes of a particular ancestor node are grouped into the same language family. We aggregate the results of single language clusters within a specific language family and calculate the expert weight α_{mt} for each cluster using a Gate mechanism (Li et al., 2018). This weight is determined by the proportion of each cluster's word count in relation to the total sample size. In essence, we incorporate the size information of each cluster into the clustering process. The single language stage involves a total of $N_{\text{sin}} = |Clu^{\text{sin}}_{m1}| + |Clu^{\text{sin}}_{m2}| + \cdots + |Clu^{\text{sin}}_{m|G_{\text{Sin}_m}|}|$ elements, where $|Clu^{\text{sin}}_{m1}|$ denotes the number of elements in cluster Clu^{sin}_{m1}. Thus, the expert weight for the t-th cluster α_{mt} of language G_{Sin_m} is defined as $\alpha_{mt} = \frac{|Clu^{\text{sin}}_{mt}|}{N_{\text{sin}}}$. We represent all expert weights across r languages of the n-th language family G_{Fam_n} as the matrix A_{sin}. The cluster-center embeddings of single language clusters, denoted as Cen^{sin}_{mt}, are used in the language family cluster stage. The expert-weighted cluster-center embeddings $h_{Cen^{\text{sin}}_{mt}} \in \mathbb{R}^{|u| \times d}$ are then input into the GMM clustering process. The matrix H_{sin} comprises all cluster-center embeddings across r languages of the n-th language family G_{Fam_n}. GMM clustering groups semantically similar words from r languages into a specific cluster Clu^{fam}_{ng}.

$$\begin{aligned} Clu^{\text{fam}}_{n1}, Clu^{\text{fam}}_{n2}, \ldots, Clu^{\text{fam}}_{n|G_{\text{Fam}_n}|} = \text{GMM}(\\ \text{ele}(A_{\text{sin}} \odot H_{\text{sin}})) \end{aligned} \tag{4}$$

where the A_{sin} and H_{sin} matrices are defined as follows:

$$\begin{aligned} A_{\text{sin}} &= \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1|G_{\text{Sin}_1}|} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2|G_{\text{Sin}_2}|} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{r|G_{\text{Sin}_r}|} \end{bmatrix}, \\ H_{\text{sin}} &= \begin{bmatrix} h_{Cen^{\text{sin}}_{11}} & h_{Cen^{\text{sin}}_{12}} & \cdots & h_{Cen^{\text{sin}}_{1|G_{\text{Sin}_1}|}} \\ h_{Cen^{\text{sin}}_{21}} & h_{Cen^{\text{sin}}_{22}} & \cdots & h_{Cen^{\text{sin}}_{2|G_{\text{Sin}_2}|}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{Cen^{\text{sin}}_{r1}} & h_{Cen^{\text{sin}}_{r2}} & \cdots & h_{Cen^{\text{sin}}_{r|G_{\text{Sin}_r}|}} \end{bmatrix} \end{aligned} \tag{5}$$

where ele() denotes the operation of enumerating every element of the matrix, and $\odot$ represents the element-wise product.

2.3.3. Multi Languages Cluster

Finally, we perform clustering on all language family cluster-center embeddings obtained from the language family cluster stage. For example, the first cluster-center of cluster Clu^{fam}_{n1} in the n-th language family is denoted as Cen^{fam}_{n1}. Each cluster-center embedding is associated with a multi-language expert weight β_{ng}, computed using the same Gate mechanism as in the language family cluster stage. We represent all expert-weight elements across z language families of the multi-language pool G_{Mul} as the matrix B_{fam}. These expert-weighted cluster-center embeddings $h_{Cen^{\text{fam}}_{ng}} \in \mathbb{R}^{|u| \times d}$ are then used in the GMM clustering process. The matrix H_{fam} contains all cluster-center embeddings across the z language families of G_{Mul}. GMM clustering groups semantically similar words from the z language families into specific clusters, denoted as Clu^{mul}_{q}.

$$\begin{aligned} Clu^{\text{mul}}_{1}, Clu^{\text{mul}}_{2}, \ldots, Clu^{\text{mul}}_{|G_{\text{Mul}}|} = \text{GMM}(\\ \text{ele}(B_{\text{fam}} \odot H_{\text{fam}})) \end{aligned} \tag{6}$$

where the matrices B_{fam} and H_{fam} are defined as follows:

$$\begin{aligned} B_{\text{fam}} &= \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1|G_{\text{Sin}_1}|} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2|G_{\text{Sin}_2}|} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{z1} & \beta_{z2} & \cdots & \beta_{z|G_{\text{Sin}_z}|} \end{bmatrix}, \\ H_{\text{fam}} &= \begin{bmatrix} h_{Cen^{\text{fam}}_{11}} & h_{Cen^{\text{fam}}_{12}} & \cdots & h_{Cen^{\text{fam}}_{1|G_{\text{Fam}_1}|}} \\ h_{Cen^{\text{fam}}_{21}} & h_{Cen^{\text{fam}}_{22}} & \cdots & h_{Cen^{\text{fam}}_{2|G_{\text{Fam}_2}|}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{Cen^{\text{fam}}_{z1}} & h_{Cen^{\text{fam}}_{z2}} & \cdots & h_{Cen^{\text{fam}}_{z|G_{\text{Fam}_z}|}} \end{bmatrix} \end{aligned} \tag{7}$$

where ele() denotes the operation of enumerating each element of the matrix, and $\odot$ represents the element-wise product.

[6] https://www.ethnologue.com/

2.4. Pseudo Semantic Data Augmentation

To enrich the training data with diverse linguistic information, we augment the model with global multilingual semantics obtained from the last domino unsupervised cluster module.

2.4.1. Pseudo Semantic Replacement

We propose two approaches for handling sentence semantics. First, we pass sentences through m-BERT to generate embeddings for each sentence. Second, we use Universal Dependencies to extract syntactic parsing trees for the sentences. To guide the model toward learning more accurate representations of crucial sentence words, we focus on key elements identified through syntactic parsing. Given that the subject, verb, and object (SVO) components are essential for comprehension in many tasks (Dai et al., 2017; Zhang et al., 2017), we treat the SVO as the crucial words of each sentence. Subsequently, we mark the position of each SVO component within the sentence:

$$\begin{aligned} h_{s_i} = [&e_{t_{i1}}, e_{t_{i2}}, \dots, e_{w_{iS}}, \dots, \\ &e_{w_{iV}}, \dots, e_{w_{iO}}, \dots, e_{t_{i|L_s|}}] \end{aligned} \tag{8}$$

where $e_{w_{iS}}$, $e_{w_{iV}}$, and $e_{w_{iO}}$ denote the embeddings of the sentence's subject, verb, and object components, respectively, and $e_{t_{i1}}$ represents the embedding of the first token in sentence S_i. We replace the original sentence's SVO word embeddings with randomly selected candidate embeddings from the same cluster. These candidate word embeddings share similar semantics with the SVO words but come from different languages. Through the replacement of crucial words with cross-lingual knowledge, we can guide the model to learn more about the critical linguistic elements of sentences and achieve better semantic representations. The updated sentence representation is expressed as:

$$\begin{aligned} h^{'}_{s_i} = [&e_{t_{i1}}, e_{t_{i2}}, \dots, e_{can'_{iS}}, \dots, \\ &e_{can'_{iV}}, \dots, e_{can'_{iO}}, \dots, e_{t_{i|L_s|}}] \end{aligned} \tag{9}$$

where $e_{can'_{iS}}$, $e_{can'_{iV}}$, and $e_{can'_{iO}}$ denote the candidate embeddings for the subject, verb, and object components, respectively. After the embedding replacement, cross-lingual pseudo semantic information is introduced to the training data. We then feed these enhanced representations into transformer models with base-level parameter sizes to refine the embeddings.

2.4.2. De-biasing Optimal Transport Affinity Regularization

Spatial misalignment exists between the original sentence and the enhanced sentence, as noted by (Huang et al., 2022b). To diminish the discrepancy between the replaced sentence embedding $h^{'}_{s_i}$ and the original sentence embedding h_{s_i}, we introduce an integrated regularization term based on the optimal transport mechanism, named Optimal Transport Affinity Regularization.

(1) Wasserstein Distance Abbreviation: To align the space of original sentence representations with that of cross-lingual knowledge-enhanced sentence representations (Wang and Henao, 2022; Alqahtani et al., 2021), we employ optimal transport ($\mathcal{OT}$) to facilitate the adjustment process. We calculate a transport plan P that maps the original sentence to the augmented sentence with optimal cost $C \in \mathbb{R}^{|L_s| \times |L_s|}$, using the Euclidean distance (Danielsson, 1980) between the two sentence representations as a measure of cost:

$$C(h_{s_i}, h^{'}_{s_i}) = \left(\sum_{j=1}^{d} \left| h_{s_ij} - h^{'}_{s_ij} \right|^2 \right)^{\frac{1}{2}} \tag{10}$$

We aim to find the optimal transport plan $P \in \mathbb{R}^{|L_s| \times |L_s|}$ that minimizes the cost C. This problem is formulated as minimizing the p-Wasserstein distance d_{p-Wass}. Due to the high computational complexity of calculating P, we approximate it using the Sinkhorn algorithm (Altschuler et al., 2017):

$$\mathbf{K} = \exp\left(-\frac{C(h_{s_i}, h^{'}_{s_i})}{\varepsilon} \right) \tag{11}$$

$$P(h_{s_i}, h^{'}_{s_i}) = diag(u)\mathbf{K}diag(v) \tag{12}$$

We compute u and v iteratively, starting with $v^{(0)} = \mathbf{1}_{|L_s|}$, using the following update formulas:

$$u^{(l+1)} = \frac{a(h_{s_i})}{\mathbf{K}v^{(l)}}, \quad v^{(l+1)} = \frac{b(h^{'}_{s_i})}{\mathbf{K}^T u^{(l+1)}} \tag{13}$$

where a and b are distribution mapping functions. The $\mathcal{OT}$ loss can be defined as:

$$loss_{\mathcal{OT}} = \langle P(h_{s_i}, h^{'}_{s_i}), C(h_{s_i}, h^{'}_{s_i}) \rangle \tag{14}$$

To mitigate the $\mathcal{OT}$ learning biases between the two sentence representations, we introduce two auxiliary de-biasing terms to calibrate the loss.

(2) De-biasing Eigenvectors Shrinkage: We utilize a linear orthogonal mapping parameter $\mathbf{W} \in \mathbb{R}^{|L_s| \times |L_s|}$ to approximate the replaced embeddings to the original ones, $h_{s_i} \approx \mathbf{W}h^{'}_{s_i}$. Singular value decomposition (SVD) is directly applied to compute $\mathbf{W}$ (Xing et al., 2015):

$$\mathbf{U\Sigma V^T} = \text{SVD}(h^{'T}_{s_i} h_{s_i}) \tag{15}$$

$$\mathbf{W} = \mathbf{VU^T} \tag{16}$$

We initialize the linear mapping function with weight $\mathbf{W}$ to simplify the learning process. However, eigenvectors with small singular values can lead to poor transformations if not suppressed (Chen et al., 2019). Thus, we penalize the smallest k singular values of Σ, which is ordered by magnitude. The eigenvectors shrinkage loss is defined as:

$$loss_{eig} = -\eta \sum_{r=1}^{k} \sigma_r^2 \tag{17}$$

where η is a hyper-parameter to control the degree of penalty, and σ_r is the r-th smallest singular value.

(3) De-biasing Distance Shrinkage: To guide the framework's learning direction towards minimizing the discrepancy, we add a term based on the distance between the two embeddings to the loss function. The distance shrinkage loss is defined as:

$$loss_{dis} = 1 - \text{sim}(h_{s_i}, h^{'}_{s_i}) \tag{18}$$

where sim$()$ represents the similarity measure.

Finally, the auxiliary $\mathcal{OT}$ affinity regularization is given by:

$$loss_{Reg} = \rho_1 loss_{\mathcal{OT}} + \rho_2 loss_{eig} + \rho_3 loss_{dis} \tag{19}$$

where ρ_i denotes the controlled weight of each regularization component, with the constraint that the sum of ρ_i equals 1.

2.5. Training Objective

Our training objective combines the task-specific loss with the $\mathcal{OT}$ affinity regularization. The overall objective function is formulated as:

$$loss_{total} = \lambda_1 loss_{task} + \lambda_2 loss_{Reg} \tag{20}$$

where λ_i controls the relative contribution of each component, and the sum of λ_i is constrained to be 1.

3. Experiments

3.1. Tasks and Datasets

Sequence Classification tasks include text classification and sentiment analysis. We selected the following datasets for these tasks: MLDoc (Schwenk and Li, 2018) for text classification, and the MultiBooked Catalan and Basque (Barnes et al., 2018)[7] for sentiment analysis. The evaluation metrics for these tasks are accuracy and macro F1.

For **Information Extraction**, we focus on Relation Extraction as a representative task. Here, the goal is to predict the correct relation type present in the data. We use the ACE2005 dataset (Walker et al., 2006), which spans three languages: English, Chinese, and Arabic. The performance is measured using micro F1.

Question Answering involves retrieving answers for specific questions from a given passage. We conduct experiments on the cross-lingual question answering dataset BiPaR (Jing et al., 2019), which is commonly used for evaluating such systems. The evaluation metrics for this task are Exact Match (EM) and micro F1.

3.2. Experiment Settings

Given computational resource constraints, we employ the base-level version of multilingual BERT (m-BERT) to obtain hidden states for words and sentences. The encoder consists of 12 Transformer layers with 12 self-attention heads, and the hidden state dimension is set to 768. During training, we experiment with learning rates in $\{1\text{e}-5, 2\text{e}-5, 3\text{e}-5, 1\text{e}-6, 2\text{e}-6, 3\text{e}-6\}$. AdamW is chosen as the optimizer, with a learning rate of $1\text{e}-3$ and weight decay of $1\text{e}-5$. For the Wasserstein distance, we set $p = 1$, while the Sinkhorn algorithm's control parameter ε is 0.1. The last $k = 300$ singular values are used in the De-biasing Eigenvectors Shrinkage section, with η in the $loss_{\text{eig}}$ formula being 0.001. The weight ρ of $loss_{\text{Reg}}$ is set to $\{0.4, 0.2, 0.4\}$, and the λ of the total loss is set to $\{0.5, 0.5\}$ independently. Statistical results are based on 5 runs, and t-tests confirm that improvements are statistically significant, with $p < 0.05$ for all results.[8]

3.3. Baselines

We compare our approach against a variety of baselines:

MLDoc (Schwenk and Li, 2018) introduces a cross-lingual text classification dataset, with baseline results from basic neural network models.

BLSE (Barnes and Klinger, 2018) presents a model for the sentiment analysis task, relying on supervised parallel bilingual data.

LASER (Artetxe and Schwenk, 2019) proposes a system utilizing a BiLSTM to learn sentence representations across 93 languages, assessed on natural language understanding (NLU) tasks.

m-BERT (Devlin et al., 2019) offers a language model pre-trained on over 100 languages, generating representations for different languages.

XLM-R (Conneau et al., 2020) is a transformer-based masked language model known for its strong cross-lingual performance.

mUSE (Yang et al., 2020b) is pre-trained in 16 languages to project multilingual corpora into a single semantic space.

CoSDA-ML (Qin et al., 2020) proposes a model using shallow string surface data augmentation

[7] We refer to these datasets collectively by the term "OpeNER".

[8] The source code and data are available at `https://github.com/MatNLP/UniPSDA`

Model	en	de	zh	es	fr	it	ja	ru	Average
MLDoc	87.2	71.7	73.5	65.3	70.2	65.1	69.8	56.9	69.9$_{(\pm 0.7)}$
LASER	86.5	86.0	70.4	71.3	73.9	65.6	58.5	63.4	72.0$_{(\pm 0.5)}$
m-BERT	92.1	74.3	72.5	67.0	70.5	61.8	69.7	61.5	71.2$_{(\pm 0.3)}$
XLM-R	90.7	78.5	70.3	66.4	67.8	63.9	64	64.0	70.7$_{(\pm 0.6)}$
ZSIW	91.3	82.8	79.6	71.7	78.1	67.0	68.5	64.3	75.4$_{(\pm 0.5)}$
DAP	94.1	86.7	81.7	76.2	84.3	67.6	73.9	66.7	78.9$_{(\pm 0.2)}$
SOGO$_{cos}$	93.2	87.0	81.8	76.2	82.5	68.7	73.7	63.9	78.4$_{(\pm 0.1)}$
X-STA	93.8	86.4	81.7	77.2	84.3	68.4	73.4	64.8	78.8$_{(\pm 0.2)}$
CoSDA-ML	92.4	79.1	72.7	69.9	74.5	64.3	70.6	**66.9**	73.8$_{(\pm 0.6)}$
UniPSDA	**94.5**	**87.1**	**82.3**	**77.4**	**84.4**	**69.4**	**74.0**	65.5	**79.3**$_{(\pm 0.2)}$

Table 1: General results of text classification in terms of accuracy (%) on the MLDoc dataset.

to include various language strings in the training data.

CCCAR (Nguyen et al., 2021) designs a model for information extraction tasks, leveraging datasets in three target languages.

ZSIW (Li et al., 2021) introduces a zero-shot instance-weighting model for cross-lingual text classification.

HERBERTa (Seganti et al., 2021) uses an unconventional two-BERT-model pipeline for information extraction.

X-METRA-ADA (M'hamdi et al., 2021) employs meta-learning for cross-lingual transfer capability enhancement.

SSDM (Wu et al., 2022) proposes a siamese semantic disentanglement model to separate syntax knowledge across languages.

LaBSE (Feng et al., 2022) is a BERT-based sentence embedding model supporting 109 languages.

DAP (Li et al., 2023) integrates sentence-level and token-level dual-alignment for cross-lingual pretraining.

SOGO$_{\mathbf{cos}}$ (Zhu et al., 2023) employs saliency-based substitution and a novel token-level alignment strategy for cross-lingual spoken language understanding.

X-STA (Cao et al., 2023) leverages an attentive teacher, gradient disentangled knowledge sharing, and multi-granularity semantic alignment for cross-lingual machine reading comprehension.

3.4. General Experimental Results

Sequence Classification: The results for sequence classification are presented in Table 1 for text classification and Table 2 for sentiment analysis. We observe that: (1) Our approach outperforms strong baselines and nearly reaches state-of-the-art performance for each task. (2) The performance of the text classification task is significantly improved by leveraging the Domino Cluster to select appropriate candidates and injecting pseudo semantic knowledge into critical components of the sentences. We achieve an average accuracy of 79.3%, with a particularly notable improvement for French, where accuracy increases by approximately 10% (from 74.5% to 84.4%) compared to the method proposed by (Qin et al., 2020). (3) In the sentiment analysis task, our UniPSDA model boosts the average macro F1 score to 86.0, as shown in Table 2. This represents the best reported result for cross-lingual sentiment analysis on the OpeNER dataset.

Model	en	eu	ca	Average
BLSE	86.1	88.5	73.9	82.8$_{(\pm 0.4)}$
m-BERT	89.9	87.9	75.2	84.3$_{(\pm 0.2)}$
mUSE	88.7	90.0	75.1	84.6$_{(\pm 0.3)}$
XLM-R	87.9	87.6	71.7	82.4$_{(\pm 0.2)}$
DAP	90.5	91.7	74.9	85.7$_{(\pm 0.1)}$
SOGO$_{cos}$	90.1	91.2	75.0	85.4$_{(\pm 0.1)}$
X-STA	90.4	90.1	74.9	85.1$_{(\pm 0.1)}$
CoSDA-ML	90.4	91.6	74.6	85.3$_{(\pm 0.5)}$
UniPSDA	**90.7**	**91.9**	**75.3**	**86.0**$_{(\pm 0.1)}$

Table 2: General experimental results of sentiment analysis in terms of macro F1 (%) and baselines on the OpeNER dataset.

Information Extraction: The results for information extraction are shown in Table 3. The findings demonstrate that: (1) Our methodology is effective for the relation extraction task, achieving more accurate cross-lingual representations as evidenced by higher macro F1 scores compared to prior work. (2) The enhanced focus on acquiring pertinent cross-lingual knowledge regarding crucial sentence components has led to a solid average F1 score of approximately 44.1 in our experiments, marking an improvement of 2.7.

Question Answering: Table 4 presents the performance of our question answering framework. The results suggest that: (1) Despite the zero-shot experimental setup, the pseudo data augmentation

Model	en	zh	ar	Average
m-BERT	57.1	44.8	21.5	41.1$_{(\pm0.5)}$
XLM-R	54.9	45.2	21.7	40.6$_{(\pm0.4)}$
LaBSE	55.1	45.8	26.6	42.5$_{(\pm0.2)}$
HERBERTa	54.5	45.7	26.4	42.2$_{(\pm0.3)}$
CoSDA-ML	58.3	45.5	20.4	41.4$_{(\pm0.5)}$
DAP	59.0	46.2	26.2	43.8$_{(\pm0.2)}$
SOGO$_{cos}$	58.7	46.2	26.5	43.8$_{(\pm0.3)}$
X-STA	57.9	45.7	26.2	43.3$_{(\pm0.1)}$
CCCAR	56.6	43.9	18.3	39.6$_{(\pm0.3)}$
UniPSDA	**59.1**	**46.3**	**26.8**	**44.1**$_{(\pm0.1)}$

Table 3: General experimental results of information extraction in terms of micro F1 (%) and baselines on the ACE2005 dataset.

Model	en		zh		Average
	EM	F1	EM	F1	
m-BERT	31.9	44.3	23.5	26.1	31.5$_{(\pm0.6)}$
XLM-R	32.3	45.0	23.3	26.2	31.7$_{(\pm0.3)}$
LaBSE	33.7	43.4	24.2	25.7	31.8$_{(\pm0.5)}$
CoSDA-ML	34.2	44.5	24.7	25.9	32.3$_{(\pm0.2)}$
X-METRA-ADA	33.9	44.9	23.8	26.8	32.4$_{(\pm0.1)}$
SSDM	34.1	45.8	24.1	26.2	32.6$_{(\pm0.2)}$
DAP	34.3	**45.9**	24.6	27.1	33.0$_{(\pm0.3)}$
SOGO$_{cos}$	34.2	45.5	24.7	27.1	32.9$_{(\pm0.2)}$
X-STA	33.8	44.9	24.2	26.7	32.4$_{(\pm0.2)}$
UniPSDA	**34.4**	45.7	**25.0**	**27.3**	**33.1**$_{(\pm0.2)}$

Table 4: General experimental results of question answering and baselines on the BiPaR dataset.

mechanism employed by our framework demonstrates a robust transfer capability. This translates to effective performance on the BiPaR dataset, with our work producing more accurate representations than most of the baselines. (2) The scores obtained in the two languages evaluated affirm the efficacy of UniPSDA. However, our F1 scores for English are lower than those achieved by SSDM (Wu et al., 2022) and DAP (Li et al., 2023). This discrepancy can be attributed to the fact that SSDM and DAP utilize specific parallel data for training, which was not the case in our approach.

4. Detailed Analysis

4.1. Ablation Study

In our ablation study, we independently remove key components—namely, the Domino Unsupervised Cluster module and the Pseudo Semantic Data Augmentation module—to evaluate their individual contributions to the framework's performance. The results of the ablation experiments are presented in Table 5. We draw two main conclusions: (1) The Domino Unsupervised Cluster module is crucial for generating precise representations, while the Pseudo Semantic Data Augmentation module significantly enhances the model's performance by providing additional cross-lingual information. (2) The absence of the cluster module leads to a noticeable decline in performance across all downstream tasks. Specifically, in text classification, accuracy falls by 4.4% (from 79.3% to 74.9%). This indicates that clustering based on semantic embeddings is more beneficial to the model than clustering based on shallow string representations. The removal of the Pseudo Semantic Data Augmentation module also results in a marked decrease in performance due to the lack of cross-lingual knowledge.

Model	OpeNER	MLDoc	ACE05	BiPaR
UniPSDA	86.0	79.3	44.1	45.7
-Dom. Unsup.	85.8	74.9	41.9	45.1
-Pse. Seman.	85.1	73.7	41.1	44.8
-Aff. Regul.	84.2	71.1	40.9	44.0

Table 5: Ablation study of our work on four datasets. "-" means returning to the original setting.

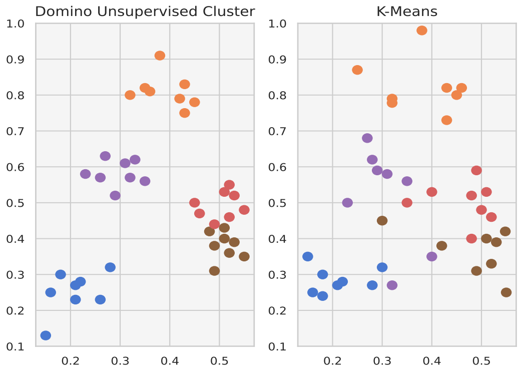

Figure 3: The words representations of five different semantics after t-SNE dimensional reduction.

4.2. Influence of Domino Unsupervised Cluster

We employ t-SNE (van der Maaten and Hinton, 2008) to project the high-dimensional word representations into a two-dimensional space, facilitating the visualization of the embeddings. The resulting plots compare word embeddings clustered by the Domino Unsupervised Cluster and the naive K-Means algorithm. As depicted in Figure 3, the domino unsupervised cluster gathers similar word embeddings more compactly, whereas the naive K-Means approach results in a more diffuse distribution of similar word embeddings.

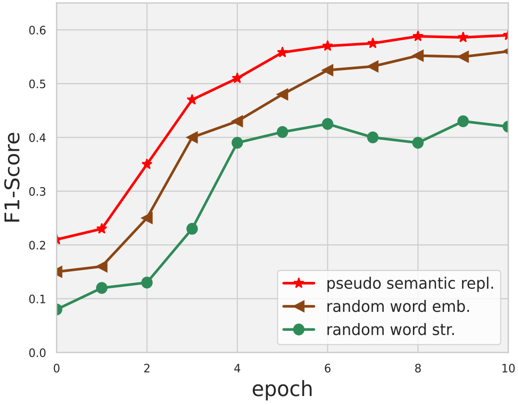

Figure 4: Results comparison of different data augmentation skills in Pseudo Semantic Data Augmentation module.

4.3. The Influence of Pseudo Semantic Data Augmentation

To examine the effect of Pseudo Semantic Data Augmentation on the information extraction task, we experiment with three distinct replacement strategies on the English test set. The comparative results are illustrated in Figure 4.

Observations indicate that replacements using random word strings or random word embeddings are less effective than those leveraging pseudo semantic methods. The pseudo semantic data augmentation approach demonstrates a superior ability to preserve the semantic integrity of sentences, leading to more meaningful augmentations and potentially better model performance.

5. Related Work

5.1. Cross-Lingual Pre-trained Models

Recent cross-lingual pre-trained language models (PLMs) can be categorized into two groups:

1. **Monolingual Training Data Models:** Multilingual BERT (m-BERT) (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) utilize monolingual corpora for training with a masked language modeling task.

2. **Multilingual Training Data Models:** Extensions of XLM-R by Jiang et al. (2022); Hämmerl et al. (2022); Chi et al. (2022); Barbieri et al. (2022) demonstrate improvements with high-quality static embedding alignments. Tools facilitating bilingual language alignment (Tran et al., 2020; Chi et al., 2021; Yang et al., 2020a; Schuster et al., 2019) enable the learning of additional languages. These models often depend on parallel data and alignment tools to enrich the corpus diversity.

5.2. Cross-Lingual Data Augmentations

Cross-lingual data augmentation approaches are typically divided into:

1. **Supervised Data Augmentation:** Methods such as CoSDA-ML (Qin et al., 2020) and MulDA (Liu et al., 2021) utilize parallel corpora to integrate knowledge from other languages. Dong et al. (2021) employ parallel language alignments for shared representational spaces.

2. **Unsupervised Data Augmentation:** Techniques like adversarial training and cross-lingual sample generation are employed by Riabi et al. (2021); Bari et al. (2021); Dong et al. (2021); Guo et al. (2021) to improve multilingual model performance. Nishikawa et al. (2021) use back translation for enhancing word embeddings, while Cheng et al. (2022) replace words based on a probabilistic distribution. Chen et al. (2021) focus on sentence selection from low-resource languages. These models tend to prioritize surface string variations, often overlooking the rich, context-aware semantics.

We address this limitation by incorporating global cross-lingual semantics into monolingual training data, thereby enriching the diversity of language knowledge.

6. Conclusion

In this work, we introduce UniPSDA, an unsupervised data augmentation mechanism that leverages semantic embeddings to enrich cross-lingual natural language understanding (NLU) tasks with diverse linguistic information. The Domino Unsupervised Cluster module identifies semantically similar cross-lingual content, while the Pseudo Semantic Data Augmentation module injects context-aware semantics into the training corpus. Furthermore, affinity regularization serves to minimize the representational gap between original and augmented sentences. Through extensive experimentation, our methods demonstrate superior performance relative to other strong baselines, underscoring their effectiveness in enhancing cross-lingual NLU.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments. This work was supported in part by National Key R&D Program of China (No. 2022ZD0120302) and Alibaba Group through Alibaba Research Intern Program.

Bibliographical References

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. GATE: graph attention transformer encoder for cross-lingual relation and event extraction. In *AAAI*, pages 12462–12470.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of EMNLP*, pages 3904–3919.

Jason M. Altschuler, Jonathan Weed, and Philippe Rigollet. 2017. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NeurIPS*, pages 1964–1974.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2022. XLM-T: multilingual language models in twitter for sentiment analysis and beyond. In *LREC*, pages 258–266.

M. Saiful Bari, Tasnim Mohiuddin, and Shafiq R. Joty. 2021. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. In *ACL*, pages 1978–1992.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. In *LREC*.

Jeremy Barnes and Roman Klinger. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL*, pages 2483–2493.

Tingfeng Cao, Chengyu Wang, Chuanqi Tan, Jun Huang, and Jinhui Zhu. 2023. Sharing, teaching and aligning: Knowledgeable transfer learning for cross-lingual machine reading comprehension. In *Findings of EMNLP*, pages 455–467.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, pages 1906–1916.

Yanda Chen, Chris Kedzie, Suraj Nair, Petra Galuscáková, Rui Zhang, Douglas W. Oard, and Kathleen R. McKeown. 2021. Cross-language sentence selection via data augmentation and rationale training. In *ACL*, pages 3881–3895.

Qiao Cheng, Jin Huang, and Yitao Duan. 2022. Semantically consistent data augmentation for neural machine translation via conditional masked language model. In *COLING*, pages 5148–5157.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *ACL*, pages 3418–3430.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: cross-lingual language model pre-training via ELECTRA. In *ACL*, pages 6170–6182.

Louis Clouâtre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022. Detecting languages unintelligible to multilingual models through local structure probes. In *Findings of EMNLP*, pages 5375–5396.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3298–3308.

Per-Erik Danielsson. 1980. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Xin Dong, Yaxin Zhu, Zuohui Fu, Dongkuan Xu, and Gerard de Melo. 2021. Data augmentation with adversarial training for cross-lingual NLI. In *ACL*, pages 5158–5167.

Yan Fan, Chengyu Wang, Boxing Chen, Zhongkai Hu, and Xiaofeng He. 2019. SPMM: A soft piecewise mapping model for bilingual lexicon induction. In *SDM*, pages 244–252. SIAM.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *ACL*, pages 878–891.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2023. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowl. Inf. Syst.*, 65(2):571–612.

Yingmei Guo, Linjun Shou, Jian Pei, Ming Gong, Mingxing Xu, Zhiyong Wu, and Daxin Jiang. 2021. Learning from multiple noisy augmented data sets for better cross-lingual spoken language understanding. In *EMNLP*, pages 3226–3237.

Katharina Hämmerl, Jindrich Libovický, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. In *Findings of ACL*, pages 2316–2329.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022a. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *ACL*, pages 4633–4646.

Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. 2022b. Learning where to learn in cross-view self-supervised learning. In *CVPR*, pages 14431–14440.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *NAACL*, pages 717–723.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. XLM-K: improving cross-lingual language model pre-training with multilingual knowledge. In *AAAI*, pages 10840–10848.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. Bipar: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *EMNLP*, pages 2452–2462.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *ACL*, pages 262–271.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *ACL*, pages 245–254.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *EMNLP*, pages 3824–3833.

Irene Li, Prithviraj Sen, Huaiyu Zhu, Yunyao Li, and Dragomir R. Radev. 2021. Improving cross-lingual text classification with zero-shot instance-weighting. In *ACL*, pages 1–7.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. In *ACL*, pages 3466–3478.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Cl-relkt: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *NAACL*, pages 2141–2155.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual NER. In *ACL/IJCNLP*, pages 5834–5846.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-METRA-ADA: cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *NAACL*, pages 3617–3632.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *EMNLP*, pages 5414–5426.

Sosuke Nishikawa, Ryokan Ri, and Yoshimasa Tsuruoka. 2021. Data augmentation with unsupervised machine translation improves the structural similarity of cross-lingual word embeddings. In *ACL-IJCNLP*, pages 163–173.

Aleksandr Perevalov, Andreas Both, Dennis Diefenbach, and Axel-Cyrille Ngonga Ngomo. 2022. Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In *WWW*, pages 977–986.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *IJCAI*, pages 3853–3860.

Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2022. Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowl. Inf. Syst.*, 64(7):1937–1966.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *EMNLP*, pages 7016–7030.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource bengali language. In *EMNLP*, pages 50–60.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *NAACL*, pages 1599–1613.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *LREC*.

Alessandro Seganti, Klaudia Firlag, Helena Skowronska, Michal Satlawa, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *EACL*, pages 1946–1955.

Xin Sun, Tao Ge, Shuming Ma, Jingjing Li, Furu Wei, and Houfeng Wang. 2022. A unified strategy for multilingual grammatical error correction with pre-trained cross-lingual language model. In *IJCAI*, pages 4367–4374.

Joanna Szolomicka and Jan Kocon. 2022. Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis. In *ICDM*, pages 443–450.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *NIPS*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. pages 2579–2605.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Chengyu Wang, Yan Fan, Xiaofeng He, and Aoying Zhou. 2019. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction. In *WWW*, pages 1965–1976. ACM.

Rui Wang and Ricardo Henao. 2022. Wasserstein cross-lingual alignment for named entity recognition. In *ICASSP*, pages 8342–8346.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. 2022. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *ACL*, pages 991–1000.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, pages 1006–1011.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020a. Alternating language modeling for cross-lingual pre-training. In *AAAI*, pages 9386–9393.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020b. Multilingual universal sentence encoder for semantic retrieval. In *ACL*, pages 87–94.

Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In *ICCV*, pages 4243–4251.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In *EMNLP*, pages 7849–7856.